

Data Stream Mining & Processing

PROCEEDINGS of the
2018 IEEE Second International Conference on
Data Stream Mining & Processing (DSMP)



IEEE Ukraine Section (Kharkiv)
SP/AP/C/EMC/COM
Societies Joint Chapter

IEEE Ukraine Section (West)
AP/ED/MTT/CPMT/SSC
Societies Joint Chapter

August 21–25, 2018

Lviv, Ukraine



Proceedings of the 2018 IEEE Second International Conference on Data Stream Mining & Processing (DSMP)

Organized by

IEEE Ukraine Section

IEEE Ukraine Section (Kharkiv) SP/AP/C/EMC/COM Societies Joint Chapter

IEEE Ukraine Section (West) AP/ED/MTT/CPMT/SSC Societies Joint Chapter

IT Step University

Ukrainian Catholic University

Lviv Polytechnic National University

Kharkiv National University of Radio Electronics

Lviv, Ukraine
August 21-25, 2018

Copyright and Reprint Permission: Abstracting is permitted with credit to the source. Libraries are permitted to photocopy beyond the limit of U.S. copyright law for private use of patrons those articles in this volume that carry a code at the bottom of the first page, provided the per-copy fee indicated in the code is paid through Copyright Clearance Center, 222 Rosewood Drive, Danvers, MA 01923. For reprint or republication permission, email to IEEE Copyrights Manager at pubs-permissions@ieee.org. All rights reserved. Copyright ©2018 by IEEE.

Additional copies may be ordered from:

IEEE Conference Operations

445 Hoes Lane, P.O. Box 1331, Piscataway, NJ
08855-1331 USA

DSMP'2018 Organizing Committee

IT Step University,
83a Zamarstyniv'ska st., 79019, Lviv, Ukraine

E-mail: dsmp.conference@gmail.com

IEEE Catalog Number: CFP18J13-CDR

ISBN: 978-1-5386-8175-6

DSMP'2018 Conference Committee

Honorary Chairpersons

Yuriy Rashkevych, Ukraine

Yevgeniy Bodyanskiy, Ukraine

General Chairs

Dmytro Peleshko, Ukraine

Olena Vynokurova O., Ukraine

Yaroslav Prytula, Ukraine

Technical Program Committee Chair

Dmytro Peleshko, Ukraine

Publication and Finance Chair

Olena Vynokurova O., Ukraine

Technical Program Committee

Aizenberg I., USA	Nakonechny A., Ukraine
Antoshchuk S., Ukraine	Petlenkov E., Estonia
Babichev S., Czech Republic	Qu S.C., China
Balasubramaniam J., India	Rekik A., Tunisia
Berezkiy O., Ukraine	Romanyshyn Yu., Ukraine
Bidyuk P., Ukraine	Rusyn B., Ukraine
Bogomolov S., Australian	Sachenko A., Ukraine
Boyun V., Ukraine	Setlak G., Poland
Churyumov G., Ukraine	Šipeky L., Slovakia
Didmanidze I., Georgia	Shelevytsky I., Ukraine
Du K., China	Slipchenko A., Netherlands
Dyvak M., Ukraine	Smolarz A., Poland
Gabsi M., France	Snytyuk V., Ukraine
Gabsi M., Tunisia	Sokolov O., Poland
Gozhiy O., Ukraine)	Sokolovsky Ya., Ukraine
Gryniv R., Ukraine	Souii M., Ph.D., Tunisia
Hnatushenko V., Ukraine	Stepashko V., Ukraine
Hu W.B., China	Štěpnička M., Czech Republic
Kareem Kamal A. Ghany, Egypt	Schlesinger M., Ukraine
Karlik B., Albania	Su J., China
Khaled G., Tunisia	Szymanski Z., Poland
Kharchenko V., Ukraine	Temani M., Tunisia
Klawonn F., Germany	Tkachenko R., Ukraine
Kokshenev I., Ph.D., Brazil	Tsmots I., Ukraine
Krylov V., Ukraine	Vassiljeva K., Estonia
Lu C.W., China	Voloshyn V., Ukraine
Lytvynenko V., Ukraine	Vorobyov S., Finland
Lyubchik L., Ukraine	Wójcik W., Poland
Lubinets Ya., Ukraine	Wu J.Q., China
Malyar M., Ukraine	Yanovsky F., Ukraine
Markov K., Bulgaria	Yatsymirskyy M., Poland
Mashkov V., Czech Republic	Ye Z.W., China
Mashtalir V., Ukraine	Yegorova E., United Kingdom
Mikhalyov O., Ukraine	Zhengbing Hu, China
Morklyanyk B., Ukraine	Zaychenko Yu., Ukraine

Local Organizing Committee Chair

Taras Rak, Ukraine

PR Manager

Maria Shepel, Ukraine

Event Manager

Yulia Vasylets, Ukraine

Members of Local Organizing Committee

Alekseyev V., Ukraine	Miyushkovych Yu., Ukraine
Andriychuk M., Ukraine	Molchanovskyi O., Ukraine
Batyuk A., Ukraine	Mulesa P., Ukraine
Berezko O., Ukraine	Panchenko T., Ukraine
Borzov Yu., Ukraine	Perova I., Ukraine
Doroshenko A., Ukraine	Pichkalov I., Ukraine
Didyk O., Ukraine	Povkhan I., Ukraine
Dumin O., Ukraine	Seniuk V., Ukraine
Ivanov Yu., Ukraine	Shateyev O., Ukraine
Figura R., Poland	Sviridova T., Ukraine
Kostyuk N., Ukraine	Sydorenko R., Ukraine
Klyuvak A., Ukraine	Tsiura N., Ukraine
Kyselova A., Ukraine	Tyshchenko O., Ukraine
Lotoshynska N., Ukraine	Veselovsky S., Ukraine
Malets I., Ukraine	Vysotska V., Ukraine
Menshikova O., Ukraine	

List of Reviewers

- Yu. Rashkevych, Ukraine
D. Peleshko, Ukraine
O. Vynokurova, Ukraine
V. Voloshyn, Ukraine
Ie. Gorovyi, Ukraine
Yu. Romanyshyn, Ukraine
S. Antoshchuk, Ukraine
Ya. Sokolovskyy, Ukraine
I. Perova, Ukraine
O. Didyk, Ukraine
Ye. Pavlov, Ukraine
N. Kulishova, Ukraine
I. Shelevytsky, Ukraine
L. Lyubchik, Ukraine
M. Yatsymirskyy, Poland
A. Dolotov, Ukraine
D. Puchala, Poland
P. Tarasiuk, Poland
Ie. Burov, Ukraine
O. Karabin, Ukraine
G. Kriukova, Ukraine
N. Lamonova, Ukraine
V. Lytvyn, Ukraine
Ya. Todorov, Finland
G. Ponomaryova, Ukraine
O. Gorokhovatskyi, Ukraine
A. Chernodub, Ukraine
- Ye. Bodyanskiy, Ukraine
O. Gozhyj, Ukraine
I. Aizenberg, USA
T. Panchenko, Ukraine
R. Ali, Tunisia
V. Hnatushenko, Ukraine
V. Mashtalir, Ukraine
V. Aliksieiev, Ukraine
O. Berezsky, Ukraine
E. Yegorova, United Kingdom
B. Tiwana, USA
O. Dumin, Ukraine
V. Lytvynenko, Ukraine
G. Churyumov, Ukraine
T. Rak, Ukraine
A. Slipchenko, Netherlands
K. Stokfiszewski, Poland
A. Berko, Ukraine
V. Volkova, Ukraine
N. Gandhi, USA
L. Kirichenko, Ukraine
A. Kuzyk, Ukraine
U. Ozkaya, Turkish
R. Upadhyay, Tatarstan
O. Menshikova, Ukraine
S. Babichev, Czech Republic

Partners

Exclusive partner

SoftServe

www.softserve.ua

Gold partner

GlobalLogic

www.globallogic.com

Perfectial

www.perfectial.com

Silver partner

ROMB

Partners

Lviv City Council

<http://city-adm.lviv.ua/>

Lviv Convention Bureau

<http://www.lvivconvention.com.ua/en/>

Kyivstar

<https://kyivstar.ua/>

Skhidnytska 118

<http://skhidnytska.ua/>

Welcome Letter

Dear Colleagues,

We would like to personally encourage each of you to join us at IEEE Second International Scientific Conference Data Stream Mining and Processing (DSMP'2018), which is held in Lviv – Kryve Ozero, UKRAINE, 21-25 August, 2018. Our main goal is not only to provide an opportunity for networking and learning recent scientific achievements but also a chance to be involved in real time panel discussions with IT representatives to review and discuss their practical outcomes on real projects.

The DSMP is organized by IEEE Ukraine Section, IEEE Ukraine Section (Kharkiv) SP/AP/C/EMC/COM Societies Joint Chapter, IEEE Ukraine Section (West) AP/ED/MTT/CPMT/SSC Societies Joint Chapter, IT Step University, Ukrainian Catholic University, Lviv Polytechnic National University, and Kharkiv National University of Radio Electronics.

Agenda of the DSMP'2018 is very rich. This year we have nominated a 120 number of accepted papers coming from about 27 countries which makes DSMP a truly international high impact conference. Major highlights of DSMP'2018 are its keynotes speakers. This conference proved to be extremely important given the fruitful dialog and a chance to exchange ideas and sharing valuable hands-on experience.

This year program is based on the following topics: Hybrid Systems of Computational Intelligence, Machine Vision and Pattern Recognition, Dynamic Data Mining & Data Stream Mining, Big Data & Data Science Using Intelligent Approaches and also panel with participation of IT Companies.

We are proud of the fact that DSMP proceedings have been included into the IEEE Xplore Digital Library as well as other Abstracting and Indexing (A&I) databases (Scopus, Web of Science and etc.). High quality of the DSMP program would not be possible without the contribution of authors, keynote speakers, organizers, students, 53 reviewers who devoted a lot of enthusiasm and hard work to prepare papers, presentations, organization infrastructure and carefully review all submissions. We are very grateful for their efforts.

We would like to thank each of your for attending our conference and bringing your expertise to our gathering.

We would like to express our gratitude to our partners and sponsors for being so generous and sponsoring our conference.

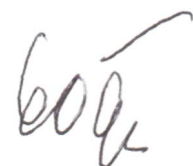
We wish all participants an excellent conference, fruitful discussions and pleasant stay in Lviv and Conference venue.

Sincerely

Yuriy Rashkevych



Yevgeniy Bodyanskiy



General List of Topics

Topic #1. Big Data & Data Science Using Intelligent Approaches	1
Topic #2. Dynamic Data Mining & Data Stream Mining	111
Topic #3. Hybrid Systems of Computational Intelligence	303
Topic #4. Machine Vision and Pattern Recognition	453
Panels	615

Table of Contents

Topic #1. Big Data & Data Science Using Intelligent Approaches	1
Iryna Perova, Olena Litovchenko, Yevgeniy Bodyanskiy, Yelizaveta Brazhnykova, Igor Zavgorodnii and Pavlo Mulesa. MEDICAL DATA-STREAM MINING IN THE AREA OF ELECTROMAGNETIC RADIATION AND LOW TEMPERATURE INFLUENCE ON BIOLOGICAL OBJECTS	3
Polina Zhernova, Anastasiia Deineko, Yevgeniy Bodyanskiy and Vladyslav Riepin. ADAPTIVE KERNEL DATA STREAMS CLUSTERING BASED ON NEURAL NETWORKS ENSEMBLES IN CONDITIONS OF UNCERTAINTY ABOUT AMOUNT AND SHAPES OF CLUSTERS	7
Ganna Ponomaryova, Igor Nevlydov, Oleksandr Filipenko and Mariya Volkova. MEMS-BASED INERTIAL SENSOR SIGNALS AND MACHINE LEARNING METHODS FOR CLASSIFYING ROBOT MOTION.	13
Dmytro Lande, Valentyna Andrushchenko and Iryna Balagura. DATA SCIENCE IN OPEN-ACCESS RESEARCH ON-LINE RESOURCES	17
Nina Khairova, Svitlana Petrasova and Włodzimierz Lewoniewski. BUILDING THE SEMANTIC SIMILARITY MODEL FOR SOCIAL NETWORK DATA STREAMS	21
Gautam Pal, Gangmin Li and Katie Atkinson. BIG DATA REAL TIME INGESTION AND MACHINE LEARNING	25
Andrii Berko and Vladyslav Aliksieiev. A METHOD TO SOLVE UNCERTAINTY PROBLEM FOR BIG DATA SOURCES.	32
Yuriy Kondratenko and Nina Kondratenko. COMPUTATIONAL LIBRARY OF THE DIRECT ANALYTIC MODELS FOR REAL-TIME FUZZY INFORMATION PROCESSING	38
Oleksandr Gerasin, Yuriy Zaporozhets and Yuriy Kondratenko. MODELS OF MAGNETIC DRIVER INTERACTION WITH FERROMAGNETIC SURFACE AND GEOMETRIC DATA COMPUTING FOR CLAMPING FORCE LOCALIZATION PATCHES	44
Volodymyr Ostakhov, Viktor Morozov and Nadiia Artykulna. MODELS OF IT PROJECTS KPIS AND METRICS	50
Yuliya Kozina, Natalya Volkova and Daniil Horpenko. MOBILE APPLICATION FOR DECISION SUPPORT IN MULTI-CRITERIA PROBLEMS	56
Olena Basalkevych and Olexandr Basalkevych. FUZZY RECONSTRUCTIONS IN LINGUISTICS	60
Mykola Malyar, Oleksy Voloshyn, Volodymyr Polishchuk and Marianna Sharkadi. FUZZY MATHEMATICAL MODELING FINANCIAL RISKS	65
Peter Bidyuk, Aleksandr Gozhyj, Iryna Kalinina, Zdislaw Szymanski and Volodymyr Beglytsia. THE METHODS BAYESIAN ANALYSIS OF THE THRESHOLD STOCHASTIC VOLATILITY MODEL	70
Max Garkavtsev, Natalia Lamonova and Alexander Gostev. CHOSING A PROGRAMMING LANGUAGE FOR A NEW PROJECT FROM A CODE QUALITY PERSPECTIVE	75

Viktor Putrenko, Nataliia Pashynska and Sergiy Nazarenko. DATA MINING OF NETWORK EVENTS WITH SPACE-TIME CUBE APPLICATION	79
Vasyl Palchykov and Yuriy Holovatch. BIPARTITE GRAPH ANALYSIS AS AN ALTERNATIVE TO REVEAL CLUSTERIZATION IN COMPLEX SYSTEMS	84
Dariusz Puchala, Kamil Stokfiszewski, Kamil Wieloch and Mykhaylo Yatsymirskyy. COMPARATIVE STUDY OF MASSIVELY PARALLEL GPU REALIZATIONS OF WAVELET TRANSFORM COMPUTATION WITH LATTICE STRUCTURE AND MATRIX-BASED APPROACH	88
Vladyslav Alieksieiev. ONE APPROACH OF APPROXIMATION FOR INCOMING DATA STREAM IN IOT BASED MONITORING SYSTEM.	94
Anatoliy Batyuk, Volodymyr Voityshyn and Volodymyr Verhun. SOFTWARE ARCHITECTURE DESIGN OF THE REAL-TIME PROCESSES MONITORING PLATFORM	98
Myroslav Komar, Vladimir Golovko, Anatoliy Sachenko, Vitaliy Dorosh and Pavlo Yakobchuk. DEEP NEURAL NETWORK FOR IMAGE RECOGNITION BASED ON THE CAFFE FRAMEWORK	102
Mansouri Sadek, Mbarek Charhad, Ali Rekik and Mounir Zrigui. A FRAMEWORK FOR SEMANTIC VIDEO CONTENT INDEXING USING TEXTUAL INFORMATION	107
Topic #2. Dynamic Data Mining & Data Stream Mining	111
Olena Vynokurova, Yevgeniy Bodyanskiy, Dmytro Peleshko and Yuriy Rashkevych. THE AUTOENCODER BASED ON GENERALIZED NEO-FUZZY NEURON AND ITS FAST LEARNING FOR DEEP NEURAL NETWORKS	113
Gennady Chuiko, Olga Dvornik and Yevhen Darnapuk. SHAPE EVOLUTIONS OF POINCARÉ PLOTS FOR ELECTROMYOGRAMS IN DATA ACQUISITION DYNAMICS	119
Petro Kravets. GAME MODEL FOR DATA STREAM CLUSTERING	123
Vasyl Lytvyn, Victoria Vysotska, Yevhen Burov and Andriy Demchuk. DEFINING AUTHOR'S STYLE FOR PLAGIARISM DETECTION IN ACADEMIC ENVIRONMENT	128
Volodymyr Yuzevych, Ruslan Skrynkovskyy and Bohdan Koman. INTELLIGENT ANALYSIS OF DATA SYSTEMS FOR DEFECTS IN UNDERGROUND GAS PIPELINE	134
Liliya Chyrun, Iaroslav Kis, Victoria Vysotska and Lyubomyr Chyrun. CONTENT ANALYSIS METHOD FOR CUT FORMATION OF HUMAN PSYCHOLOGICAL STATE	139
Vasyl Lytvyn, Victoria Vysotska, Olga Lozynska, Oksana Oborska and Dmytro Dosyn. METHODS OF BUILDING INTELLIGENT DECISION SUPPORT SYSTEMS BASED ON ADAPTIVE ONTOLOGY	145
Fedir Geche, Oksana Mulesa, Veronika Voloshchuk and Anatoliy Batyuk. ABOUT KERNEL STRUCTURE CONSTRUCTION OF THE GENERALIZED NEURAL FUNCTIONS	151

Olga Smotr, Nazarii Burak, Yuriy Borzov and Solomija Ljaskovska. IMPLEMENTATION OF INFORMATION TECHNOLOGIES IN THE ORGANIZATION OF FOREST FIRE SUPPRESSION PROCESS	157
Oleg Riznyk, Olexandr Povshuk, Yurii Kynash and Yurii Noga. TRANSFORMATION OF INFORMATION BASED ON NOISY CODES	162
Anna Vergeles, Dmytro Prokopenko, Alexander Khaya and Nataliia Manakova. UNSUPERVISED REAL-TIME STREAM-BASED NOVELTY DETECTION TECHNIQUE	166
Anastasiia Deineko, Polina Zhernova, Boris Gordon, Oleksandr Zayika, Iryna Pliss and Nelya Pabyrivska. DATA STREAM ONLINE CLUSTERING BASED ON FUZZY EXPECTATION-MAXIMIZATION APPROACHING FORMATION ON SUBMISSION	171
Solomija Ljaskovska, Igor Malets, Yevgen Martyn and Oleksandr Prydatko. INFORMATION TECHNOLOGY OF PROCESS MODELING IN THE MULTIPARAMETER SYSTEMS	177
Gennadiy Churyumov, Vladimir Tokarev, Vitalii Tkachov and Stanislav Partyka. SCENARIO OF INTERACTION OF THE MOBILE TECHNICAL OBJECTS IN THE PROCESS OF TRANSMISSION OF DATA STREAMS IN CONDITIONS OF IMPACTING THE POWERFUL ELECTROMAGNETIC FIELD	183
Oleksandr Prydatko, Ivan Solotvinskyy, Yuriy Borzov, Oleksii Didyk and Olga Smotr. INFORMATIONAL SYSTEM OF PROJECT MANAGEMENT IN THE AREAS OF REGIONAL SECURITY SYSTEMS' DEVELOPMENT	187
Leonid Lyubchik and Galyna Grinberg. ONLINE RANKING LEARNING ON CLUSTERS	193
Vitalii Bulakh, Lyudmyla Kirichenko and Tamara Radivilova. TIME SERIES CLASSIFICATION BASED ON FRACTAL PROPERTIES	198
Olga Zavgorodnia, Ivan Mikheev and Oleksandr Zyma. IDENTIFYING EUROPEAN E-LEARNER PROFILE BY MEANS OF DATA MINING	202
Galyna Kriukova and Mykola Glybovets. HIGH-PERFORMANCE DATA STREAM MINING BY MEANS OF EMBEDDING HIDDEN MARKOV MODEL INTO REPRODUCING KERNEL HILBERT SPACES	207
Daniel Ambach and Oleksandra Ambach. FORECASTING THE OIL PRICE WITH A PERIODIC REGRESSION ARFIMA-GARCH PROCESS	212
Valentyna Volkova, Ivan Deriuga, Vadym Osadchyi and Olga Radyvonenko. IMPROVEMENT OF CHARACTER SEGMENTATION USING RECURRENT NEURAL NETWORKS AND DYNAMIC PROGRAMMING	218
Sergiy Golub and Nataliia Khymytsia. THE METHOD OF CLIODINAMIK MONITORING	223
Sergii Khlamov, Vadym Savanevych, Olexander Briukhovetskyi, Artem Pohorelov, Vladimir Vlasenko and Eugen Dikov. COLITEC SOFTWARE FOR THE ASTRONOMICAL DATA SETS PROCESSING	227
Anastasiya Doroshenko. PIECEWISE-LINEAR APPROACH TO CLASSIFICATION BASED ON GEOMETRICAL TRANSFORMATION MODEL FOR IMBALANCED DATASET	231

Alexey Roenko, Feliks Sirenko, Yevhen Chervoniak and Ievgen Gorovyi. DATA PROCESSING METHODS FOR MOBILE INDOOR NAVIGATION	236
Yurij Holovatch, Ralph Kenna and Olesya Mryglod. DATA MINING IN SCIENTOMETRICS: USAGE ANALYSIS FOR ACADEMIC PUBLICATIONS	241
Hanna Rudakova, Oksana Polyvoda and Anton Omelchuk. USING RECURRENT PROCEDURES TO IDENTIFY THE PARAMETERS OF THE LARGE-SIZED OBJECT MOVING PROCESS MODEL IN REAL TIME	247
Andriy Lozynskyy, Igor Romanyshyn, Bohdan Rusyn and Volodymyr Minialo. ROBUST APPROACH TO ESTIMATION OF THE INTENSITY OF NOISY SIGNAL WITH ADDITIVE UNCORRELATED IMPULSE INTERFERENCE	251
Bohdan Pavlyshenko. USING STACKING APPROACHES FOR MACHINE LEARNING MODELS	255
Romanna Malets, Igor Malets, Heorgiy Shynkarenko and Petro Vahin. MODELING OF THERMOVISCOELASTICITY TIME HARMONIC VARIATIONAL PROBLEM FOR A THIN WALL BODY	259
Oleh Suprun, Olena Sipko and Vitaliy Snytyuk. EDUCATIONAL SCHEDULE DEVELOPMENT USING EVOLUTION TECHNOLOGIES	265
Volodymyr Lyubinets, Deon Nicholas and Taras Boiko. AUTOMATED LABELING OF BUGS AND TICKETS USING ATTENTION-BASED MECHANISMS IN RECURRENT NEURAL NETWORKS	271
Yehor Lyebyedyev and Mykola Makhortykh. #EUROMAIDAN: QUANTITATIVE ANALYSIS OF MULTI-LINGUAL FRAMING OF 2013-2014 UKRAINIAN PROTESTS ON TWITTER	276
Serhii Rybalchenko. BIG DATA AUTOMATIC SYSTEM OF ANALYSIS AND TRADING ON FINANCIAL MARKETS	281
Mesbaholdin Salami, Farzad Movahedi Sobhani and Mohammad Sadegh Ghazizadeh. DEVELOPMENT OF A NEW ALGORITHM BASED ON SIMULATION – OPTIMIZATION ALGORITHMS FOR BIG DATA MINING TO IMPROVE PREDICTION OF FUTURE ELECTRICITY PRICES IN THE IRANIAN ELECTRICITY MARKET	286
Topic #3. Hybrid Systems of Computational Intelligence	303
Olena Vynokurova, Dmytro Peleshko, Viktor Voloshyn, Semen Oskerko and Yuriy Borzov. HYBRID MULTIDIMENTIONAL WAVELET-NEURO-SYSTEM AND ITS LEARNING USING CROSS ENTROPY COST FUNCTION FOR PATTERNS RECOGNITION	305
Sergej Korjagin, Pavel Klachek and Irina Liberman. DEVELOPMENT OF HYBRID COMPUTATIONAL INTELLIGENCE BY KNOWLEDGE GENESIS METHOD	310
Igor Aizenberg and Kashifuddin Qazi. CLOUD DATACENTER WORKLOAD PREDICTION USING COMPLEX-VALUED NEURAL NETWORKS	315
Yegor Kovylin and Oleg Volkovsky. COMPUTER SYSTEM OF BUILDING OF THE SEMANTIC MODEL OF THE DOCUMENT INFORMATION ON SUBMISSION	322

Alina Shafronenko, Yevgeniy Bodyanskiy, Artem Dolotov and Galina Setlak. FUZZY CLUSTERING OF DISTORTED OBSERVATIONS BASED ON OPTIMAL EXPANSION USING PARTIAL DISTANCES	327
Nataliia Kashpruk, Anna Walaszek-Babiszewska and Marek Rydel. ON THE EQUIVALENCE BETWEEN AR FAMILY TIME SERIES MODELS AND FUZZY MODELS IN SIGNAL PROCESSING	331
Sergii Babichev, Volodymyr Lytvynenko, Maxim Korobchynskiy, Jiří Škvor and Maria Voronenko. INFORMATION TECHNOLOGY OF GENE EXPRESSION PROFILES PROCESSING FOR PURPOSE OF GENE REGULATORY NETWORKS RECONSTRUCTION	336
Ali Rekik and Nissen Masmoudi. A NEW APPROACH FOR FORMING A PROBABILISTIC RISK ASSESSMENT MODEL OF INNOVATIVE PROJECT IMPLEMENTATION UNDER RISK	342
Viktor Morozov, Olena Kalnichenko, Andrii Khрутba, Grigory Steshenko and Iuliia Liubyma. MANAGING OF CHANGE STREAMS IN PROJECTS OF DEVELOPMENT DISTRIBUTED INFORMATION SYSTEM	346
Alexander Vlasenko, Olena Vynokurova, Nataliia Vlasenko and Marta Peleshko. A HYBRID NEURO-FUZZY MODEL FOR STOCK MARKET TIME-SERIES PREDICTION	352
Vladyslav Kotsovsky, Fedir Geche and Anatoliy Batyuk. FINITE GENERALIZATION OF THE OFFLINE SPECTRAL LEARNING	356
Nelya Pabyrivska and Viktor Pabyrivskyy. INVERSE PROBLEM FOR TWO-DIMENSIONAL HEAT EQUATION WITH AN UNKNOWN SOURCE	361
Yuliia Tatarinova. AVIA: AUTOMATIC VULNERABILITY IMPACT ASSESSMENT ON THE TARGET SYSTEM	364
Olexiy Azarov, Leonid Krupelnitsky and Hanna Rakytyanska. A FUZZY MODEL OF TELEVISION RATING CONTROL WITH TREND RULES TUNING BASED ON MONITORING RESULTS	369
Yaroslav Sokolovskyy, Maryana Levkovich, Olha Mokrytska and Vitalij Atamanyuk. MATHEMATICAL MODELING OF TWO-DIMENSIONAL DEFORMATION-RELAXATION PROCESSES IN ENVIRONMENTS WITH FRACTAL STRUCTURE	375
Shashi Bhushan, Raju Pal and Svetlana Antoshchuk. ENERGY EFFICIENT CLUSTERING PROTOCOL FOR HETEROGENEOUS WIRELESS SENSOR NETWORK: A HYBRID APPROACH USING GA AND K-MEANS	381
Pavlo Vitynskyi, Roman Tkachenko, Ivan Izonin and Hakan Kutucu. HYBRIDIZATION OF THE SGTN NEURAL-LIKE STRUCTURE THROUGH INPUTS POLYNOMIAL EXTENSION	386
Igor Aizenberg and Zain Khaliq. ANALYSIS OF EEG USING MULTILAYER NEURAL NETWORK WITH MULTI-VALUED NEURONS	392
Galyna Chornous and Ihor Nikolskyi. BUSINESS-ORIENTED FEATURE SELECTION FOR HYBRID CLASSIFICATION MODEL OF CREDIT SCORING	397

Zhengbing Hu and Oleksii Tyshchenko. A HYBRID NEURO-FUZZY ELEMENT: A NEW STRUCTURAL NODE FOR EVOLVING NEURO-FUZZY SYSTEMS	402
Kostyantyn Kharchenko, Oleksandr Beznosyk and Valeriy Romanov. IMPLEMENTATION OF NEURAL NETWORKS WITH HELP OF A DATA FLOW VIRTUAL MACHINE	407
Viktor Mashkov, Jiří Fišer, Volodymyr Lytvynenko and Maria Voronenko. SELF-DIAGNOSIS OF THE SYSTEMS WITH INTERMITTENTLY FAULTY UNITS	411
Dmytro Chumachenko. ON INTELLIGENT MULTIAGENT APPROACH TO VIRAL HEPATITIS B EPIDEMIC PROCESSES SIMULATION	415
Sergii Kondratiuk and Iurii Krak. DACTYL ALPHABET MODELING AND RECOGNITION USING CROSS PLATFORM SOFTWARE	420
Lukasz Wieczorek and Przemyslaw Ignaciuk. INTELLIGENT SUPPORT FOR RESOURCE DISTRIBUTION IN LOGISTIC NETWORKS USING CONTINUOUS-DOMAIN GENETIC ALGORITHMS	424
Ihor Shelevytsky, Victorya Shelevytska, Vlad Golovko and Bogdan Semenov. SEGMENTATION AND PARAMETRIZATION OF THE PHONOCARDIOGRAM FOR THE HEART CONDITIONS CLASSIFICATION IN NEWBORNS	430
Oleksandr Dumin, Dmytro Shyrokorad, Gennadiy Pochanin, Vadym Plakhtii and Oleksandr Prishchenko. SUBSURFACE OBJECT IDENTIFICATION BY ARTIFICIAL NEURAL NETWORKS AND IMPULSE RADIOLOCATION	434
Ivan Tsmots, Oleksa Skorokhoda, Yurii Tsymbal, Taras Tesluyk and Viktor Khavalko. NEURAL-LIKE MEANS FOR DATA STREAMS ENCRYPTION AND DECRYPTION IN REAL TIME	438
Mykola Dyvak, Iryna Oliynyk, Andriy Pukas and Andriy Melnyk. SELECTION THE "SATURATED" BLOCK FROM INTERVAL SYSTEM OF LINEAR ALGEBRAIC EQUATIONS FOR RECURRENT LARYNGEAL NERVE IDENTIFICATION	444
Paweł Tarasiuk and Mykhaylo Yatsymirskyy. OPTIMIZED CONCISE IMPLEMENTATION OF BATCHER'S ODD-EVEN SORTING	448
Topic #4. Machine Vision and Pattern Recognition	453
Dmytro Peleshko, Oleksii Maksymiv, Taras Rak, Orysia Voloshyn and Bohdan Morklianyk. CORE GENERATOR OF HYPOTHESES FOR REAL-TIME FLAME DETECTING	455
Oleksii Gorokhovatskyi and Olena Peredrii. SHALLOW CONVOLUTIONAL NEURAL NETWORKS FOR PATTERN RECOGNITION PROBLEMS	459
Volodymyr Gorokhovatskyi, Yevgenyi Putyatin, Oleksii Gorokhovatskyi and Olena Peredrii. QUANTIZATION OF THE SPACE OF STRUCTURAL IMAGE FEATURES AS A WAY TO INCREASE RECOGNITION PERFORMANCE	464
Ali Al-Ammouri, Hasan Al-Ammori, Arsen Klochan and Anastasia Degtiarova. LOGIC-MATHEMATICAL MODEL FOR RECOGNITION THE DANGEROUS FLIGHT EVENTS	468

Yevgeniy Bodyanskiy, Nonna Kulishova and Daria Malysheva. THE MULTIDIMENSIONAL EXTENDED NEO-FUZZY SYSTEM AND ITS FAST LEARNING FOR EMOTIONS ONLINE RECOGNITION	473
Nataliya Boyko, Nataliya Shakhovska and Oleg Basystiuk. PERFORMANCE EVALUATION AND COMPARISON OF SOFTWARE FOR FACE RECOGNITION, BASED ON DLIB AND OPENCV LIBRARY	478
Andriy Klyuvak, Oksana Kliuvak and Ruslan Skrynkovskyy. PARTIAL MOTION BLUR REMOVAL	483
Sergei Yelmanov and Yuriy Romanyshyn. A GENERALIZED DESCRIPTION FOR THE PERCEIVED CONTRAST OF IMAGE ELEMENTS	488
Maksym Korobchynskiy, Alexander Mariliv, Mihail Slonov and Serhii Mieshkov. METHOD FOR DETERMINING THE RATIONAL TIME INTERVALS FOR DETECTING OBJECTS BY THERMAL IMAGER	494
Vitaliy Boyun. BIOINSPIRED APPROACHES TO THE SELECTION AND PROCESSING OF VIDEO INFORMATION	498
Vyacheslav Moskalenko, Alona Moskalenko, Artem Korobov, Olha Boiko, Serhii Martynenko and Oleksandr Borovenskiy. MODEL AND TRAINING METHODS OF AUTONOMOUS NAVIGATION SYSTEM FOR COMPACT DRONES	503
Kirill Smelyakov, Dmytro Yeremenko, Vitalii Polezhai, Anton Sakhon and Anastasiya Chupryna. BRAILLE CHARACTER RECOGNITION BASED ON NEURAL NETWORKS	509
Sergey Rassomakhin, Alexandr Kuznetsov, Vladimir Shlokin, Ivan Bilozetsev and Roman Serhienko. MATHEMATICAL MODEL FOR THE PROBABILISTIC MINUTIA DISTRIBUTION IN BIOMETRIC FINGERPRINT IMAGES	514
Yevgeniy Bodyanskiy, Iryna Pliss, Daria Kopaliani and Olena Boiko. DEEP 2D-NEURAL NETWORK AND ITS FAST LEARNING	519
Andriy Yerokhin, Valerii Semenets, Alina Nechyporenko, Oleksii Turuta and Andrii Babii. F-TRANSFORM 3D POINT CLOUD FILTERING ALGORITHM	524
Petr Hurtik, David Číž, Oto Kaláb, David Musiolek, Petr Kočárek and Martin Tomis. SOFTWARE FOR VISUAL INSECT TRACKING BASED ON F-TRANSFORM PATTERN MATCHING	528
Ievgen Gorovyi, Vitalii Vovk, Maksim Shevchenko, Valerii Zozulia and Dmytro Sharapov. EMBEDDED VISION MODULES FOR TEXT RECOGNITION AND FIDUCIAL MARKERS TRACKING	534
Roman Martysyshyn, Yulia Miyushkovych, Lubomyr Sikora, Natalya Lysa and Rostyslav Tkachuk. TECHNOLOGY OF REMOTE RECOGNITION THE DART-ARROW ON THE TARGET	538
Anatoliy Kovalchuk and Nataliia Lotoshynska. ELEMENTS OF RSA ALGORITHM AND EXTRA NOISING IN A BINARY LINEAR-QUADRATIC TRANSFORMATIONS DURING ENCRYPTION AND DECRYPTION OF IMAGES	542

Sergii Mashtalir, Volodymyr Mashtalir and Mykhailo Stolbovyi. REPRESENTATIVE BASED CLUSTERING OF LONG MULTIVARIATE SEQUENCES WITH DIFFERENT LENGTHS	545
Sergii Mashtalir, Olena Mikhnova and Mykhailo Stolbovyi. SEQUENCE MATCHING FOR CONTENT-BASED VIDEO RETRIEVAL	549
Oleh Berezsky, Oleh Pitsun, Natalia Batryn, Kateryna Berezska, Nadiya Savka and Taras Dolynyuk. IMAGE SEGMENTATION METRIC-BASED ADAPTIVE METHOD	554
Igor Malets, Oleksandr Prydatko, Vasyl Popovych and Andriy Dominik. INTERACTIVE COMPUTER SIMULATORS IN RESCUER TRAINING AND RESEARCH OF THEIR OPTIMAL USE INDICATOR	558
Roman Melnyk and Yurii Kalychak. ANALYSIS OF METAL DEFECTS BY CLUSTERING THE SAMPLE AND DISTRIBUTED CUMULATIVE HISTOGRAM	563
Sergei Yelmanov and Yuriy Romanyshyn. IMAGE CONTRAST ENHANCEMENT USING A MODIFIED HISTOGRAM EQUALIZATION	568
Yevhen Zadorozhnii, Yevhenii Tverdokhlib, Tetiana Fedoronchak and Natalia Myronova. DEVELOPMENT AND IMPLEMENTATION OF HUMAN FACE ALIGNMENT AND TRACKING IN VIDEO STREAMS	574
Mariya Nazarkevych, Ivanna Klyujnyk and Hanna Nazarkevych. INVESTIGATION THE ATEB-GABOR FILTER IN BIOMETRIC SECURITY SYSTEMS	580
Bohdan Durnyak, Oleksandr Tymchenko Jr., Oleksandr Tymchenko and Bohdana Havrysh. APPLYING THE NEURONETCHIC METHODOLOGY TO TEXT IMAGES FOR THEIR RECOGNITION	584
Volodymyr Sherstiuk, Marina Zharikova and Igor Sokol. FOREST FIRE MONITORING SYSTEM BASED ON UAV TEAM, REMOTE SENSING, AND IMAGE PROCESSING	590
Yuriy Furgala, Yuriy Mochulsky and Bohdan Rusyn. EVALUATION OF OBJECTS RECOGNITION EFFICIENCY ON MAPES BY VARIOUS METHODS	595
Tetiana Gladkykh, Taras Hnot and Roman Grubnyk. MUSIC CONTENT SELECTION AUTOMATION	599
Galyna Shcherbakova, Victor Krylov, Maksym Gerganov, Svitlana Antoshchuk, Marina Polyakova and Anatoly Sachenko. AREAL MULTISTART METHOD OF OPTIMIZATION FOR IMAGE RECOGNITION	605
Maksym Kovalchuk, Vasyl Koval, Anatoliy Sachenko and Diana Zahorodnia. DEVELOPMENT OF REAL-TIME FACE RECOGNITION SYSTEM USING LOCAL BINARY PATTERNS	609
Panels	615
Author's Index	xvii

Topic #1

Big Data & Data Science Using Intelligent Approaches

Medical Data-Stream Mining in the Area of Electromagnetic Radiation and Low Temperature Influence on Biological Objects

Iryna Perova
Biomedical Engineering Department
Kharkiv National University of Radio
Electronics
Kharkiv, Ukraine
rikywenok@gmail.com

Yelizaveta Brazhnykova
Biomedical Engineering Department
Kharkiv National University of Radio
Electronics
Kharkiv, Ukraine
yelyzaveta.brazhnykova@nure.ua

Olena Litovchenko
Molecular Biology and Biotechnology
Department
V. N. Karazin National University
Kharkiv, Ukraine
latyshkaelena@gmail.com

Igor Zavgorodnii
Hygiene and Ecology No.2 Department
Kharkiv National Medical University
Kharkiv, Ukraine
zavnikua@gmail.com

Yevgeniy Bodyanskiy
Artificial Intelligence Department
Kharkiv National University of Radio
Electronics
Kharkiv, Ukraine
yevgeniy.bodyanskiy@nure.ua

Pavlo Mulesa
Cybernetics and Applied Mathematics
Department
Uzhhorod National University
Uzhhorod, Ukraine
ppmulesa@gmail.com

Abstract—At this paper medical data stream mining in the area of influence by different fields (electromagnetic radiation and positive low temperature) on laboratory white rats is investigated. The most informative features in multidimensional time series using neural network based on Oja's neurons and the most informative physical field influence on biological objects (white rats) are detected.

Keywords—Principal Component Analysis, Medical Data Stream Mining, Electromagnetic Radiation Influence

I. INTRODUCTION

The current state of the environment is represented by environmental factors of a different nature (physical, chemical, biological, social factors) that influence various biological objects, including human beings, both in everyday life and at work. Such a variety of factors predetermines their complex, mixed or combined, simultaneous or successive influence [1]. In such combinations of factors, biological systems form a whole complex of responses in organs and systems that depend on the strength, concentration, and time of action of these factors [2,3]. Ecological and biological researches should be aimed at determining the nature of these influences, the features of the development of mechanisms of biological effects at different levels of the body functioning, as well as determining the portion of the contribution made by each of the factors belonging to the complex of acting factors [1,4].

A fact that dimensionality of medical data sets is excess for solving of diagnostics tasks is one of the biggest problems in the area of Medical Data Mining [5-8]. This problem means containing of excess information in medical dataset that lead to complexity of analysis and interpretation of results. Data Reduction stage became a needed one for solving a common Data Mining tasks.

This stage permits to represent dataset in the compact and easy visualized form. Principal Component Analysis (PC-analysis) is one of most usable methods for solving data reduction tasks when dataset presents in the form of table "object-property".

When dataset is fed to processing in the form of data stream PC-analysis cannot be used and solving of data reduction task can be performed using neural network technologies. A neural network based on Oja's neuron [9] with its learning algorithm is the system that permits to perform data reduction sequentially in online mode.

II. ELECTROMAGNETIC RADIATION AND POSITIVE LOW TEMPERATURE INFLUENCE ON BIOLOGICAL OBJECTS

Scientists describe the present ecological situation in the world as «electromagnetic pollution of the environment» [10] in connection with the current spread of electromagnetic radiation (EMR) generated by different equipment. At present, the levels of exogenous electric and magnetic fields exceed significantly the natural background of the Earth and are an unfavorable factor, which influence on biological objects grows from year to year, acquiring a global character [11,12]. Sources of radiation can be found in all spheres of human activity, which are used in industry, medical practice, in the educational process, life and entertainment [13].

Recently, more and more researches are devoted to the problems of electromagnetic ecology. The world community recognizes that EMR is a significant environmental factor and has a high biological activity. Many international organizations are engaged in the development of this problem. Their research is aimed at studying the biological effect of EMR with subsequent development of the principles of regulation (methodology) in order to protect the population from the negative influence of EMR, as well as the substantiation of unified world standards for electromagnetic safety [3,11].

Numerous literature data indicate that EMR affects all organs and systems of the body: the blood system, cardiovascular, endocrine, immune, nervous and sexual systems with different biological activity (depressing, trigger or phase) in all frequency ranges [14,15].

However, despite a large number of studies, there is still no generally accepted theory of the effect of EMR on the body, its nature and the mechanisms of action on

physiological systems. The reason for this may be attributed to the fact that EMR never acts as a monofactor. Other factors of a different nature that influence the body in combination with EMR are not excluded, which predetermined the urgency of the problem of medical and biological study of EMR influence on the human body in environmental conditions [1,14].

Another important environmental factor is the air temperature, which can provide comfortable or uncomfortable living conditions. The adverse effect of temperature on the body is possible under various circumstances, especially in winter, when the air temperature decreases [16].

Cold is a stressor for the body; in response, the activity of the most important regulatory systems is activated in order to maintain the temperature constant of the body. Hypothermia has a general effect on the body, causing biological reactions that manifest themselves as a complex of biochemical, pathophysiological, morphofunctional changes [17,18]. In response to irritation with cold, the body reveals a number of complex physiological reactions. It is proved that the corresponding reactions of the body to the effect of low temperature depend on the volume and duration of the effect of this factor [19].

Thus, the global spread of EMR, its combined simultaneous effect with a temperature factor can affect a person under different conditions. Proceeding from the above, it is necessary to study the influence of EMR in combination with a positive low temperature (PLT) on the body to determine the biological effects of the combined action: additivity, synergism, and antagonism. It is also necessary to determine the portion of the contribution made by each factor to the total effect, with subsequent development of criteria for evaluating the biological effect of EMR under conditions of cold stress followed by hygienic assessment and development of measures to prevent their impact.

With that end in view, we developed an original research model. An experimental study of the effect of physical factors on the animal organism is carried out: EMR and PLT, both in isolated action and in combined effect. The study was carried out under conditions of a subacute experiment during 30 days.

The Exposure Chamber equipment was created, which allowed to simulate both the influence of the required range of air temperature on laboratory animals and EMR parameters (Fig. 1). This model is protected by copyright [20].

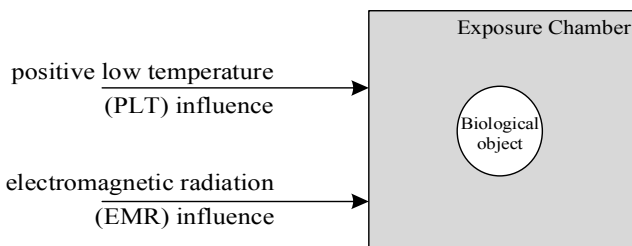


Fig. 1. Exposure chamber equipment

III. EXPERIMENT RESULTS PROCESSING

Laboratory white rats (males of the WAG line) were chosen as a biological object. The animals were distributed into 2 groups (N = 60). The research group consisted of 30 animals, which were divided into three groups: the first group of 10 animals was subjected to isolated EMR (operating frequency 70 kHz, electric voltage 600 V / m); the second group of 10 animals was influenced by the isolated effect of PLT in the range of + 4 °C ± 2 °C; the third group (10 animals) experienced combined effects of EMR and PLT at the same parameters as the groups of isolated action. The intact group of 30 animals was in comfortable conditions (+ 25 °C ± 2 °C) and served as a control in relation to groups 1, 2, and 3. Expositions were conducted during 4 hours, 5 times a week.

To reveal biological effects, the blood serum of animals was studied at the stage of 5, 15, 30 days in the dynamics of the experiment. The following biochemical parameters were determined in the blood serum: the content of diene conjugates (DC), malonic dialdehyde (MDA), SH-groups, ceruloplasmin, cholesterol, triglycerides, high density lipoprotein (HDL), low density lipoprotein (LDL), very low density lipoproteins (VLDL), urea, acid and alkaline phosphatase, chlorides, calcium, magnesium, phosphorus, total protein, glucose, catalase and superoxide dismutase activity; atherogenicity index (AI) was calculated. The indices were determined using commercial test systems with help of the biochemical analyzer "Labline-80" (Austria) in accordance with the instructions attached to them.

From mathematical point of view each object-white rat is described by multidimensional time series that contain information about the blood serum in 5, 15, 30 days:

$$X(k) = \{x_{il}(k)\},$$

where $k = 1, \dots, N, \dots$ – number of object-white rat in matrix (in our case $N = 60$), $i = 1, \dots, n$ – number of time series for each of white rat (in our case $n = 20$), $l = 1, \dots, q$ – number of time instants, that correspond to number of days ($q = 3$).

To present information in easy reception mode we have used neural network described in [9], [22]-[23] for each of time series. On Fig.2-Fig.4 position of each white rat in space of three principal components was presented. Intact group of white rats was marked by o-dots and research group – by *-dots. In each of groups different influence type was marked by different colors: black for the influence by the isolated effect of PLT, magenta for influence by isolated EMR, red for combined effects of EMR and PLT.

It is easy to see that proposed visualization presents that combined effects of EMR and PLT influence has significant correlation with the influence by the isolated effect of PLT on 5-th and 15-th days.

To determine this relation, it is easy to introduce a fuzzy clustering procedure between centers of groups with different influence type (one with PLT and one with EMR influence) and centers of group of combined effects [24].

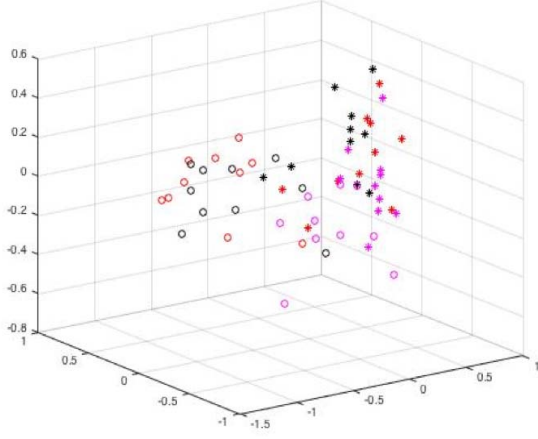


Fig. 2. PCA-visualization of each white rat on 5-th day of experiment

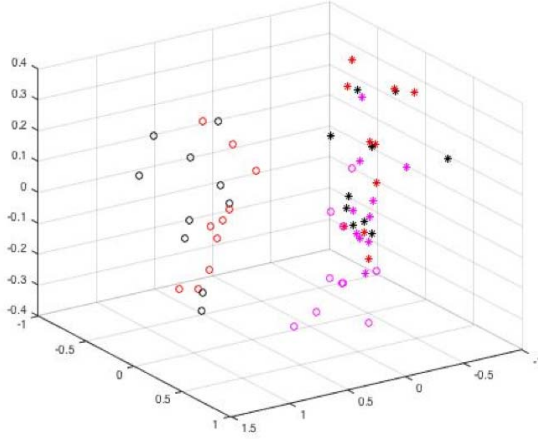


Fig. 3. PCA-visualization of each white rat on 15-th day of experiment

Previously, input multidimensional time series have to be centered, normalized and coded to interval $[-1;1]^{[n]}$.

For calculation of centers of each cluster is suitable to use arithmetical mean or median between all white rats of corresponding group in sequential mode:

$$c_{il}(k) = c_{il}(k-1) + \eta(k) \text{sign}(x_{i+1,l}(k) - c_{il}(k-1))$$

where $\eta_m(k)$ – learn rate parameter, that tuned accordingly with expression [25]

$$\eta(k) = r^{-1}(k), \quad r(k) = \alpha r(k-1) + 1, \quad 0 < \alpha \leq 1.$$

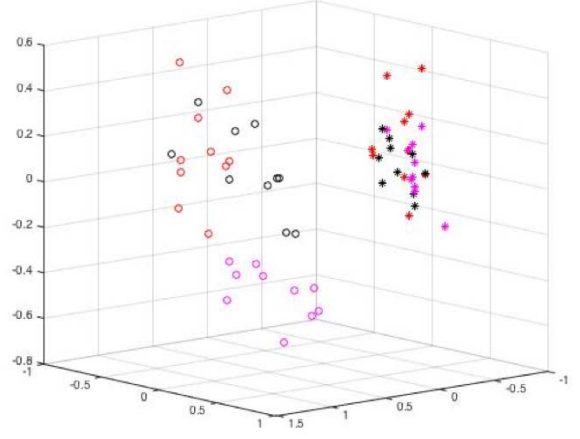


Fig. 4. PCA-visualization of each white rat on 30-th day of experiment

Distance in sense of Manhattan metrics between these centers is calculated in the form:

$$\text{dist}(c_{il}, c_{il}) = \sum_{i=1}^n \sum_{l=1}^q |c_{il} - c_{il}|$$

After that we can use a measure of distance to calculate what type of isolated influence contributes to combined one:

$$md = \frac{\text{dist}^{-1}}{\sum(\text{dist}^{-1})}$$

After calculation we obtain a result that the isolated effect of PLT has membership level $md = 0,67$ whereas the influence by isolated EMR $md = 0,33$.

At next step of our research it's needs to obtain information about most informative features. In the area of Medical Data Mining this problem is known as feature selection [18].

First eigenvector of covariance matrix should be calculated and first principal component vector should be formed. First principal component of coded multidimensional matrix $X(k) = \{x_{il}(k)\} \in R^{[n]}$ can be defined as:

$$\hat{y}^{(1)}(x_{il}(k)) = l_1 \cdot x_{il}(k)$$

where l_1 – first row of matrix L , eigenvector of covariance matrix, that corresponds to the biggest eigenvalue of this matrix.

At next step distances in the sense of Manhattan metrics between all features vectors and first principal component is calculated. A feature that has minimal distance

$$d(x_{il}(z), \hat{y}_{il}^{(1)}) = \sum_{i=1}^N \sum_{l=1}^q |x_{il}(k) - \hat{y}_{il}^{(1)}|$$

is chosen like the most informative one.

Then this feature-winner is excluded from original data matrix and system continues to process reduced matrix until all features will be turned over.

As a result the most informative features are: malonicdialdehyde (MDA), low density lipoprotein (LDL), urea, a superoxide dismutase, catalase.

IV. CONCLUSION

At this paper an isolated PLT and EMR influence and their combined effect on biological object – laboratory white rats was investigated. A contribution of isolated influence by PLT and EMR to combined influence was calculated. The most informative features in multidimensional time-series was determined.

REFERENCES

- [1] I. Belyaev, A. Dean, H. Eger, G. Hubmann, R. Jandrisovits, M. Kern, M. Kundi, H. Moshhammer, P. Lercher, K. Müller, G. Oberfeld, P. Ohnsorge, P. Pelzmann, C. Scheingraber, and R. Thill “EUROPAEM EMF Guideline 2016 for the prevention, diagnosis and treatment of EMF-related health problems and illnesses,” *Rev Environ Health*, vol. 31(3), pp. 363-397, Sep 1, 2016. doi: 10.1515/reveh-2016-0011.
- [2] V. I. Nazarenko “To the problem of modulating influence of some physical factors on biological effect of EMF 50 Hz,” *Environment and Health*, no. 4, pp. 38-43, 2009.
- [3] European Commission, Directive 2013/35/ EU of the European Parliament and of the Council of 26 June 2013 on the minimum health and safety requirements regarding the exposure of workers to the risks arising from physical agents (electromagnetic fields) (20th individual Directive within the meaning of Article 16(1) of Directive 89/391/EEC) and repealing Directive 2004/40/EC, 2013 (<http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:L:2013:179:0001:0021:EN:PDF>)
- [4] J. Kaszuba-Zwojńska, J. Gremba, B. Gałdzińska-Calik, K. Wójcik-Piotrowicz, P.J. Thor, “Electromagnetic field induced biological effects in humans”. *Przegl Lek*, 2015, 72(11), pp.636-641.
- [5] I. Perova, and Ye. Bodyanskiy, “Fast medical diagnostics using autoassociative neuro-fuzzy memory” *International Journal of Computing*, 16 (1), pp.34-40, 2017.
- [6] I. Perova, and I. Pliss “Deep hybrid System of Computational Intelligence with Architecture Adaptation for Medical Fuzzy Diagnostics,” *I.J. Intelligent System and Applications*, vol. 7, pp.12-21, 2017. DOI: 10.5815/ijisa.2017.07.02
- [7] I. Pliss, and I. Perova “Diagnostic Neuro-Fuzzy System and Its Learning in Medical Data Mining Tasks in Conditions of Uncertainty about Numbers of Attributes and Diagnoses,” *Automatic Control and Computer Sciences*, vol. 51(6), pp.391-3982, 017. DOI: 10.3103/S0146411617060062
- [8] Ye. Bodyanskiy, I. Perova, O. Vynokurova, and I. Izonin “Adaptive Wavelet Diagnostic Neuro-Fuzzy System for Biomedical Tasks,” 14th International Conference on Advanced Trends in Radioelectronics, Telecommunications and Computer Engineering (TCSET), Lviv-Slavske, Ukraine, pp.299-303, February 2018.
- [9] E. Oja “Neural Network, principal components and subspaces,” *Int. J. of Neural Systems*, vol. 1, pp. 61-68, 1989.
- [10] G. Redlarski, B. Lewczuk, A. Żak et al. “The Influence of Electromagnetic Pollution on Living Organisms: Historical Trends and Forecasting Changes,” *BioMed Research International*, 18 pages, 2015. doi:10.1155/2015/234098.
- [11] P. Bodera, W. Stankiewicz, K. Zawada, B. Antkowiak, M. Paluch, J. Kieliszek, B. Kalicki, A. Bartosiński, and I. Wawer, “Changes in antioxidant capacity of blood due to mutual action of electromagnetic field (1800 MHz) and opioid drug (tramadol) in animal model of persistent inflammatory state,” *Pharmacol Rep*, vol. 65(2), pp.421-428, 2013.
- [12] S. Manzetti, and O. Johansson, “Global electromagnetic toxicity and frequency-induced diseases: theory and short overview,” *Pathophysiology*, vol.19, iss.3, pp. 185–19, 12012.
- [13] Yu. D. Gubernsky, M. E. Goshin, N. V. Kalinina, and I. M. Banin, “Hygienic aspects of electromagnetic pollution of modern dwelling,” *Hygiene and sanitation*, no. 4. 2016. (<https://cyberleninka.ru/article/n/gigienicheskie-aspekty-elektromagnitnogo-zagryazneniya-sovremennogo-zhilischa>).
- [14] J. L. Phillips, N. P. Singh, and H. Lai, “Electromagnetic fields and DNA damage,” *Pathophysiology*, vol 16, no. 2-3, pp. 79-88, 2009. DOI: 10.1016/j.pathophys.2008.11.005.
- [15] S. Singh, and N. Kapoor “Health Implications of Electromagnetic Fields, Mechanisms of Action and Research Needs”, *Advances in Biology*, pp.1-24, 2014.
- [16] T. T. Qu, J. X. Deng, R. L. Li, Z. J. Cui, X. Q. Wang, L. Wang, and J. B. Deng “Stress injuries and autophagy in mouse hippocampus after chronic cold exposure,” *Neural Regen Res*, vol. 12(3), pp.440-446, 2017. doi: 10.4103/1673-5374.202932.
- [17] O. N. Kolosova, “Stabilization of homeostasis of rat body in cold influence with help of ethanol.” *Bulletin of experimental biology and medicine*, vol. 160, no. 9, pp. 279-283, 2015.
- [18] L. N. Maslov, S. Yu. Tsybulnikov, N. V. Naryzhnaya, V. V. Ivanov, and M. R. Tsybulnikova, “Chronic influence of cold –adaptation without stress,” *Pathological Physiology and Experimental Therapy*, no. 1, pp. 28-31, 2016.
- [19] P. Vargovic, M. Laukova, J. Ukropec, G. Manz, and R. Kvetnansky, “Prior Repeated Stress Attenuates Cold-Induced Immunomodulation Associated with "Browning" in Mesenteric Fat of Rats,” *Cell MolNeurobiol*, 38(1), pp.349-361, 2018. doi:10.1007/s10571-017-0531-z. Epub 2017 Aug 11.
- [20] I. V. Zavgorodnii, V. V. Myasoyedov, V. O. Vekshin, R. O. Bachinsky, O. S. Teslenko, D. P. Pertsev, and G. L. Nikulin, Patent on useful model “Exposure Chamber” No. 83559, applicant and patent-owner Kharkiv National Medical University No. u201305791 dated 07.05.2013
- [21] C. R. Rao “The use and interpretation of principal component analysis in applied research,” *Sankhay, A*, vol. 26, no. 4, pp. 329-358, 1964.
- [22] M. Okamoto “Optimality Principal Components Multivariate Analysis,” *Proc. 3 Int. Symp. Dayton*, 1967.
- [23] M. Okamoto, and M. Kanazawa, “Minimization of Eigenvalues of a matrix and optimality of principal components,” *Ann. Math. Statist.*, vol. 39, no. 3, pp. 1-20, 1968.
- [24] O. Nelles, *Nonlinear System Identification*. Berlin: Springer, 2001.
- [25] M. T. Vazan, *Stochastic Approximation*. Cambridge. Transactions in Mathematics and Mathematical Physics. Cambridge University Press, Cambridge, 1969.

Adaptive Kernel Data Streams Clustering Based on Neural Networks Ensembles in Conditions of Uncertainty About Amount and Shapes of Clusters

Polina Ye. Zhernova
Department of System Engineering
Kharkiv National University of Radio Electronics
Kharkiv, Ukraine
polina.zhernova@gmail.com

Yevgeniy V. Bodyanskiy
Artificial Intelligence Department
Kharkiv National University of Radio Electronics
Kharkiv, Ukraine
yevgeniy.bodyanskiy@nure.ua

Anastasiia O. Deineko
Artificial Intelligence Department
Kharkiv National University of Radio Electronics
Kharkiv, Ukraine
anastasiya.deineko@gmail.com

Vladyslav O. Riepin
Artificial Intelligence Department
Kharkiv National University of Radio Electronics
Kharkiv, Ukraine
revanmax626@gmail.com

Abstract—The neural network’s approach for data stream clustering task, that in online mode are fed to processing in assumption of uncertainty about amount and shapes of clusters, is proposed in the paper. The main idea of this approach is based on the kernel clustering and idea of neural networks ensembles, that consist of the T. Kohonen’s self-organizing maps. Each of the clustering neural networks consists of different number of neurons, where number of clusters is connected with the quality of these neurons. All ensemble members process information that sequentially is fed to the system in the parallel mode. Experimental results have proven the fact that the system under consideration could be used to solve a wide range of Data Mining tasks when data sets are processed in an online mode.

Keywords—clustering, X-means method, ensemble of neural networks, self-learning, T. Kohonen’s neural network.

I. INTRODUCTION

Data stream clustering is an important part of Data Mining. Many approaches to its solution have been developed [1, 2]. Processing of large information volumes requires, first of all, a high speed and simple numerical implementation of clustering algorithms. One of the most popular procedures is the K-means method due to its simplicity, clarity of results and possibilities for their explicit interpretation [3]. This method refers to the algorithms based on calculation of centroids-prototypes. In the frame of this approach, an initial data set (possibly growing)

$$X = \{x(1), \dots, x(k), \dots, x(N)\},$$

$x(k) = (x_1(k), \dots, x_i(k), \dots, x_n(k))^T \in R^n$, $k = 1, 2, \dots, N$ is partitioned into m clusters where their number m is defined a priori or chosen empirically.

The X-means method is alternative approach for empirical methods of clustering, but it is more bulky from computational point of view and connected with strict apriori statistical assumptions about character of initial data distribution [4, 5].

Besides, both these methods require multi epoch procedure for initial data set X , that has limited opportunities

for processing big data sets (Big Data) and data streams, when information is fed to the inputs of the clustering system sequentially observation by observation in the online mode (Data Stream Mining). In this situation number of observation k has the sense of current discrete time, but data volume N practically doesn’t limited.

In similar situations clustering self-learning artificial neural networks [6-9] show themselves rather effective. First of all, it concerns self-organizing T. Kohonen’s maps (*SOM*) [10] that can process data in sequential mode. *SOM* processing results coincide with the K-means results, wherein the number of m clusters is known apriori.

Saving capabilities of online processing using *SOM* and establishing the number of m clusters with K-means is possible, using the idea of clustering ensembles [11-14]. As elements of the ensemble it is needed to use Kohonen’s clustering neural networks *SOM^m* [15], where every network is tuned for a different number of possible classes $m=2,3,\dots,M$. Under this approach, first member of the ensemble *SOM²* in Kohonen’s layer contains only two neurons with vectors of synaptic weights w_1^2, w_2^2 . The last member *SOM^M* contains M neurons with centroids-weights $w_1^M, w_2^M, \dots, w_M^M$.

In ensemble self-learning process, all *SOM^m* are operate in parallel. As the final result is chosen clustering network-winner, which shows the most appropriate results in terms of the applied quality criterion for clustering [2,16]. Note that in every *SOM^m* at each cycle k of information processing neuron-winner is chosen exactly the same as neural network-winner is chosen in ensemble at each tact. It is the best result of clustering.

An essential restriction that reduces approach capabilities is the requirement of formed clusters linear separability and convexity. Whereas the real data have the ability to form classes of completely arbitrary form. In such a situation it can be useful to use T. Cover’s theorems of linear separation in spaces of higher dimension and J. Mercer’s kernels, which provide this increasing [17, 18]. Based on this approach so-called, kernel self-organizing maps (KSOM) were developed

[19-21]. They show quite good results under conditions of clusters arbitrary forms with a known number m of them, using fixed volume N of the processed selection. Therefore, this seems expedient to develop an ensemble of kernel clustering neural networks. It is intended for data streams

online processing under conditions of an unknown or changing number of classes.

The architecture of kernel clustering neural networks ensemble is shown on Fig. 1. It contains five layers of information processing.

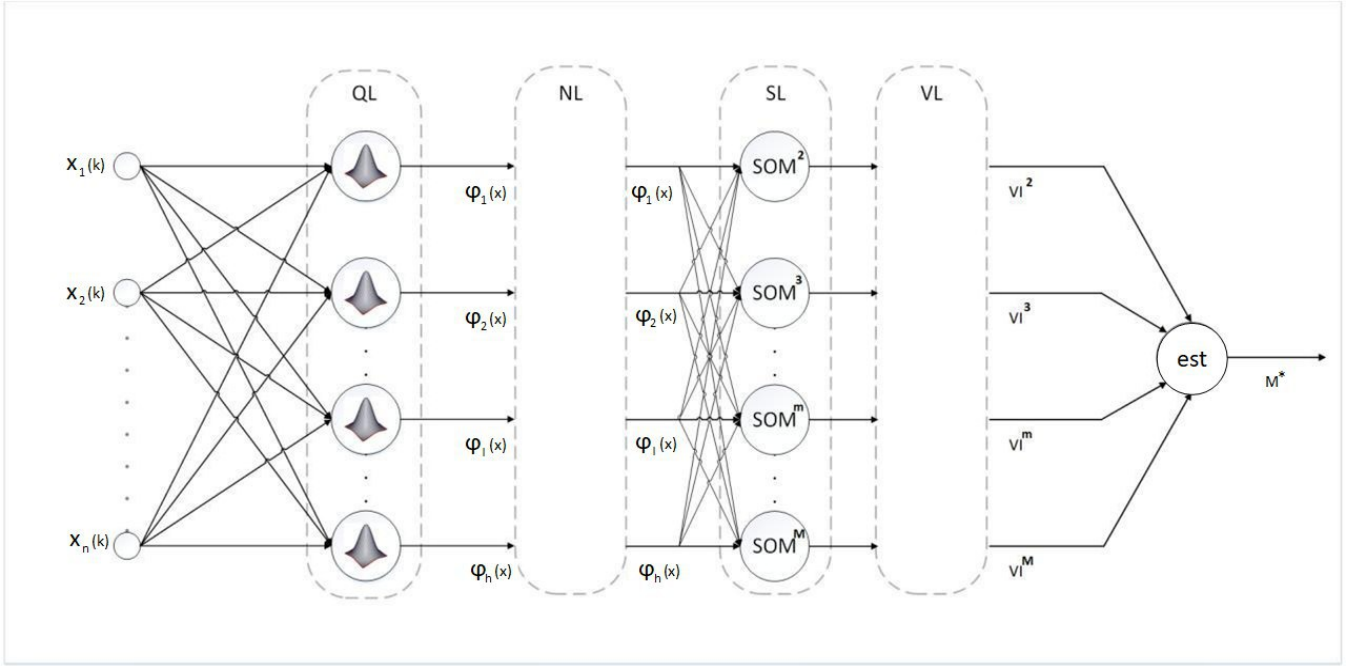


Fig. 1. The architecture of kernel clustering neural networks ensemble

The initial information to be clustered is fed to the zero (input) layer of the system as a sequence $x(1), x(2), \dots, x(k), \dots, x(N), \dots$. Then, it enters to the first hidden layer (RL) of radial-basis functions, formed by R-neurons. Right in this layer increasing in the dimensionality of the input space with the help of kernel functions system $\varphi_1(x), \varphi_2(x), \dots, \varphi_l(x), \dots, \varphi_h(x)$, $h > n$, occurs. As a functions, either Gaussians or other bell-shaped functions are used, for example,

$$\varphi_l(x) = \left(1 + \frac{\|x - c_l\|^2}{\gamma_\varphi} \right)^{-1} = \frac{\gamma_\varphi}{\gamma_\varphi + \|x - c_l\|^2}$$

where $c_l - (n \times 1)$ - vector that sets the "center" of the radial-basis function $\varphi_l(x)$, γ_φ - a scalar parameter that determines the receptive field area, which is the same as "width" of this function.

Thus, when a vector signal enters to the system input $x(k) = (x_1(k), \dots, x_i(k), \dots, x_n(k))^T \in R^n$, at the output of the first RL hidden layer, a vector signal is formed: $\varphi(x(k)) = (\varphi_1(x(k)), \dots, \varphi_l(x(k)), \dots, \varphi_h(x(k)))^T \in R^h$, $h > n$.

The second NL hidden layer realizes an elementary signal $\varphi(x(k))$ normalization operation in the form

$$\tilde{\varphi}(x(k)) = \frac{\varphi(x(k))}{\|\varphi(x(k))\|}$$

which is needed for effective work of the third SL hidden layer. It was formed at the expense of $(M-1)$ Kohonen's self-organizing maps SOM^m , each of which works under the assumption that in the data sample being processed, there are m classes. Clustering quality is provided by each SOM^m and is estimated using one or another validation index [2] in the fourth VL hidden layer. It calculates corresponding indices $VI^2, VI^3, \dots, VI^m, \dots, VI^M$ for every possible $m=2,3,\dots,M$.

Finally, in the output layer containing a single node, an optimum detector, the particular SOM^m is determined. It provides best clustering quality, wherein assumed that in the analyzing data array there are m clusters.

II. THE KERNEL CLUSTERING SYSTEM AND ITS SELF-LEARNING BASED ON NEURAL NETWORKS ENSEMBLE

Self-learning process of considered system is realized in the first hidden layer RL, where the centers $c_l, l=1,2,\dots,h$ of kernel functions $\varphi_l(x)$ are tuned. Also, it realized in a third hidden layer SL, where the synaptic weights w_j^m , $m=2,3,\dots,M$, $j=1,2,\dots,m$ are estimated for each neural network of SOM^m ensemble.

Let's consider the tuning process for the centers of kernel functions, consisting the following steps [22]:

Step 0: set threshold value Δ that determines the indiscernibility level of two neighboring kernel functions.

After that, the maximum possible number h of these functions and receptive fields parameter γ_φ .

Step 1: when the first vector-observation $x(1)$ is fed to the system input, the first center c_1 and radial-basis function are being formed.

$$\varphi_1(x) = \frac{\gamma_\varphi}{\gamma_\varphi + \|x - c_1\|^2}$$

where $c_1 = x(1)$.

Step 2: when the second vector-observation $x(2)$ is fed to the system input, inequality has been checked

$$\|x(2) - c_1\| \leq \Delta$$

and if it is satisfied, then $x(2)$ does not form a new center. And if the following condition is satisfied

$$\Delta < \|x(2) - c_1\| \leq 2\Delta,$$

then c_1 is being corrected in accordance with the T. Kohonen's self-learning rule "Winner takes all" [10]:

$$c_1(2) = c_1(1) + \eta(2)(x(2) - c_1(1))$$

where $c_1(1) = x(1)$, $0 < \eta(2) < 1$ is learning rate parameter. If the condition

$$2\Delta < \|x(2) - c_1\|$$

is satisfied then a new kernel function is formed

$$\varphi_2(x) = \frac{\gamma_\varphi}{\gamma_\varphi + \|x - c_2\|^2} = \frac{\gamma_\varphi}{\gamma_\varphi + \|x - x(2)\|^2}.$$

After every new observation $x(k)$ this process is realized.

If on the step N h radial-basis functions are generated then in the future amount of them doesn't grow. Refinement centers c_l , $l = 1, 2, \dots, h$ that were already generated can be provide only according to the condition (1) and the self-learning rule (2).

The adaptation process also consists of three steps [10]: competition, cooperation and synaptic adaptation for every SOM^m in ensemble, wherein synaptic weights vectors w_j^m describe h -dimensional centroids of formed clusters.

On the competition step input signal of second hidden layer NL $\tilde{\varphi}(x(k)) \in R^h$ is fed to every input of all SOM^m where they are compared with each of synaptic weights vectors $w_j^m(k-1)$ in the sense of distance

$$D(\tilde{\varphi}(x(k)), w_j^m(k-1)) = \|\tilde{\varphi}(x(k)) - w_j^m(k-1)\|, \quad (3)$$

$$j = 1, 2, \dots, m; \quad m = 2, 3, \dots, M.$$

Because $\|\tilde{\varphi}(x(k))\| = 1$, instead of the Euclidean metric (3) more easier is to use cosine similarity measure

$$\text{sim}(\tilde{\varphi}(x(k)), w_j^m(k-1)) = \tilde{\varphi}^T(x(k))w_j^m(k-1)$$

by the help of which for every SOM^m its neuron-winners are determined, for that

$$\tilde{\varphi}^T(x(k))w_j^{m*}(k-1) = \max_j \tilde{\varphi}^T(x(k))w_j^m(k-1)$$

On the cooperation step all neurons-winners of the ensemble generate topological neighborhoods areas, in which not only winners tuned, but and their nearest neighbors.

This area is described by the membership function $\varphi(j, l)$, that are similar to the radial-basis functions of the first hidden layer:

$$\varphi(j, l) = \frac{\gamma}{\gamma + \|w_l^m(k-1) - w_j^{m*}(k-1)\|^2}$$

The synaptic centroids-weights specification of every SOM^m is occurs on the synaptic adaptation step by the T. Kohonen's self-learning rule "Winners takes more":

$$w_l^m(k) = w_l^m(k-1) + \eta(k)\varphi(j, l)(\tilde{\varphi}(x(k)) - w_l^m(k-1)) \quad (4)$$

It's easy to see, that for winner w_j^{m*} (4) coincides with the learning rule (2). It has to be noted, that in the self-learning rule (4) learning rate parameters $\eta(k)$ and γ usually are selected according to the empirical considerations and must be decrease monotonically in the tuning process.

This process is easy to organize by the system of the recurrent relations

$$\begin{cases} \eta(k) = r^{-1}(k); & r(k) = \alpha r(k-1) + \|\tilde{\varphi}(x(k))\|^2 = \alpha r(k-1) + 1, \\ \gamma(k) = \eta(k)\gamma(k-1), & 0 < \alpha \leq 1, \end{cases}$$

that at the $\alpha = 1$ automatically is transformed to the stochastic approximation procedure. It's easy to see too, that first and third layers of the system in fact are tuned according to the same type procedures like WTA and WTM [10].

III. TUNING OF THE FOURTH HIDDEN LAYER

The estimation of the clustering quality is produced in the fourth hidden layer by the validation index VI^m [1], wherein this index is calculated for every of the T. Kohonen's maps SOM^m , $m = 2, 3, \dots, M$.

As the such index it's useful to implement Davies-Bouldin criterion [23], with the help of which clustering quality can be estimated even in the case of non-convex classes.

In the case of the m clusters this index can be written in the form

$$DB(m) = \sum_{j=1}^m \max_{\substack{1 \leq q \leq m \\ q \neq j}} \frac{s(w_j^m(k), u_j(k), \tilde{\varphi}(x(k))) - s(w_q^m(k), u_q(k), \tilde{\varphi}(x(k)))}{D(w_j^m(k), w_q^m(k))}$$

where $D(w_j^m(k), w_q^m(k))$ - distance between centroids

$$D(w_j^m(k), w_q^m(k)) = \|w_j^m(k) - w_q^m(k)\|,$$

$s(w_j^m(k), u_j(k), \tilde{\varphi}(x(k)))$ - the intracluster scattering characteristics for j -th cluster:

$$s(w_j^m(k), u_j(k), \tilde{\varphi}(x(k))) = \left(\frac{\sum_{k=1}^N u_j(k) \|\tilde{\varphi}(x(k)) - w_j^m(k)\|^2}{\sum_{k=1}^N u_j(k)} \right)^{\frac{1}{2}},$$

$u_j(k)$ - crisp membership function of the vector $\tilde{\varphi}(x(k))$ to the j -th cluster type:

$$u_j(k) = \begin{cases} 1, & \text{if } \tilde{\varphi}(x(k)) \text{ belongs to } j\text{-th cluster,} \\ 0 & \text{otherwise.} \end{cases}$$

As the optimal number of clusters m^* value, providing minimum of the $DB(m)$, is selected:

$$DB(m^*) = \min_m \{DB(2), DB(3), \dots, DB(M)\},$$

that is calculated in the output layer.

In the situation then non-stationary data are processed in online mode, is necessary to modify $DB(m)$ index for processing data on the “sliding-window” mode of dimension $1 < s < N$. Wherein only intercluster distance characteristics, that are calculated on the “sliding-window”, are modified by expression

$$s(w_j^m(k), u_j(k), \tilde{\varphi}(x(k)), s) = \left(\frac{\sum_{\tau=k-s+1}^k u_j(\tau) \|\tilde{\varphi}(x(\tau)) - w_j^m(k)\|^2}{\sum_{\tau=k-s+1}^k u_j(\tau)} \right)^{\frac{1}{2}}$$

when the data volume N isn't limited and grows with time $k = 1, 2, \dots, N, N+1, \dots$

IV. EXPERIMENTAL RESULTS

We have tested proposed method with two different training data sets. The first data set is artificial generated so

that it contains 3 clusters, 300 observations were every observation has 3 features. The second data set “Iris” is taken from UCI-Repository [24]. This data set consists of 150 observations that are divided into 3 classes where every observation has 3 random features. The clusters are clearly visible in the artificial generated data set and shown in Figure 2.

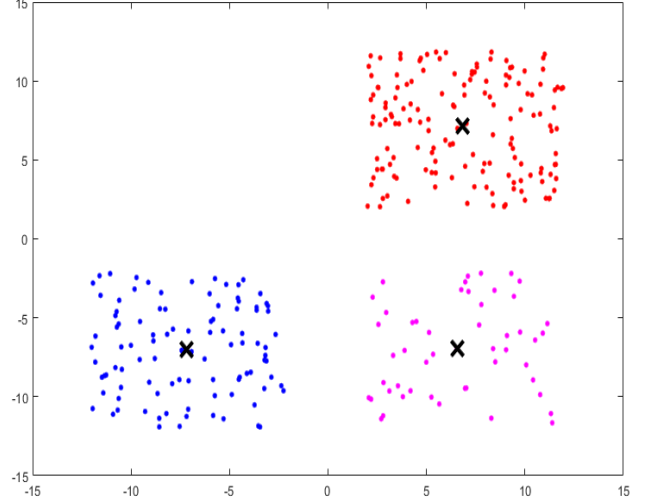


Fig. 2. The artificial generated data set

The computational accuracy of proposed method was compared with known K-means algorithm. These clustering results were estimated by the Davies-Bouldin criterion. The clustering accuracies for a series of 50 experiments are shown in Table I and Table II.

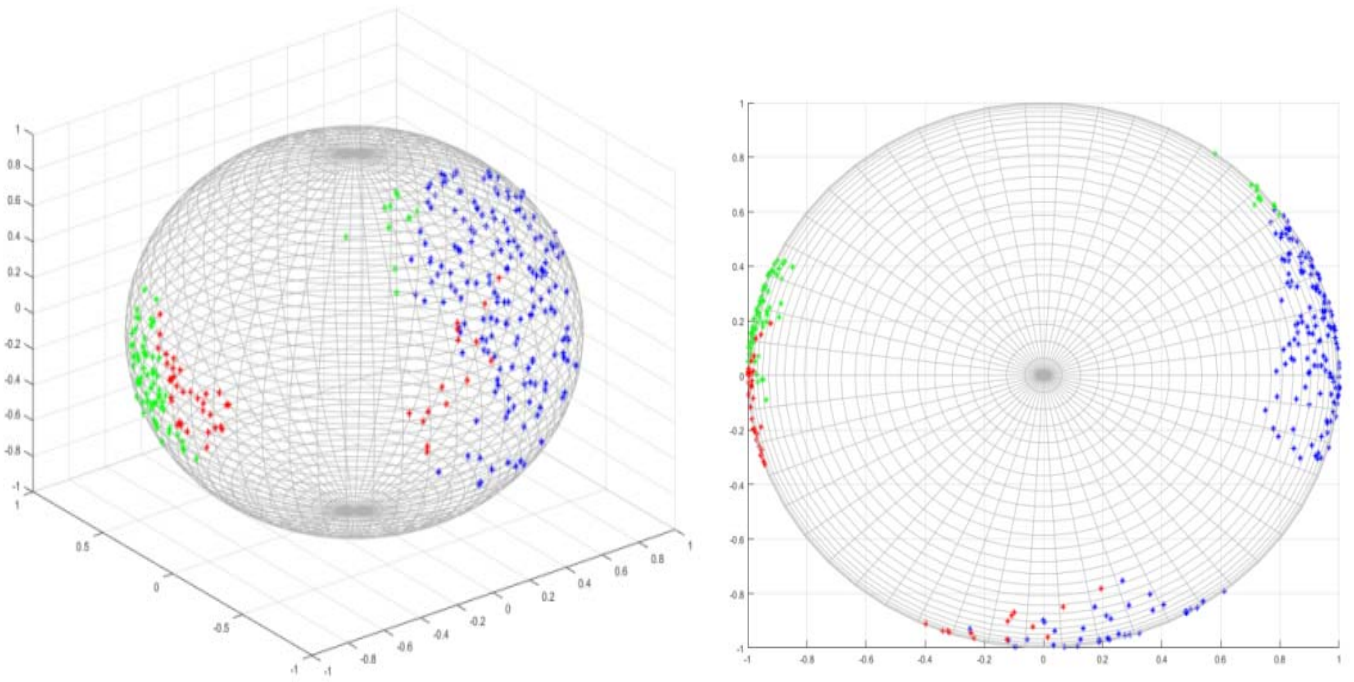
TABLE I. THE MEAN CLUSTERING ACCURACIES FOR THE DIFFERENT NUMBERS OF CLUSTERS (THE ARTIFICIAL GENERATED DATA SET)

Method	SOM ^m	k-means
clustering accuracies for 2 clusters	0,71	0,70
clustering accuracies for 3 clusters	0,89	0,76
clustering accuracies for 4 clusters	0,68	0,67

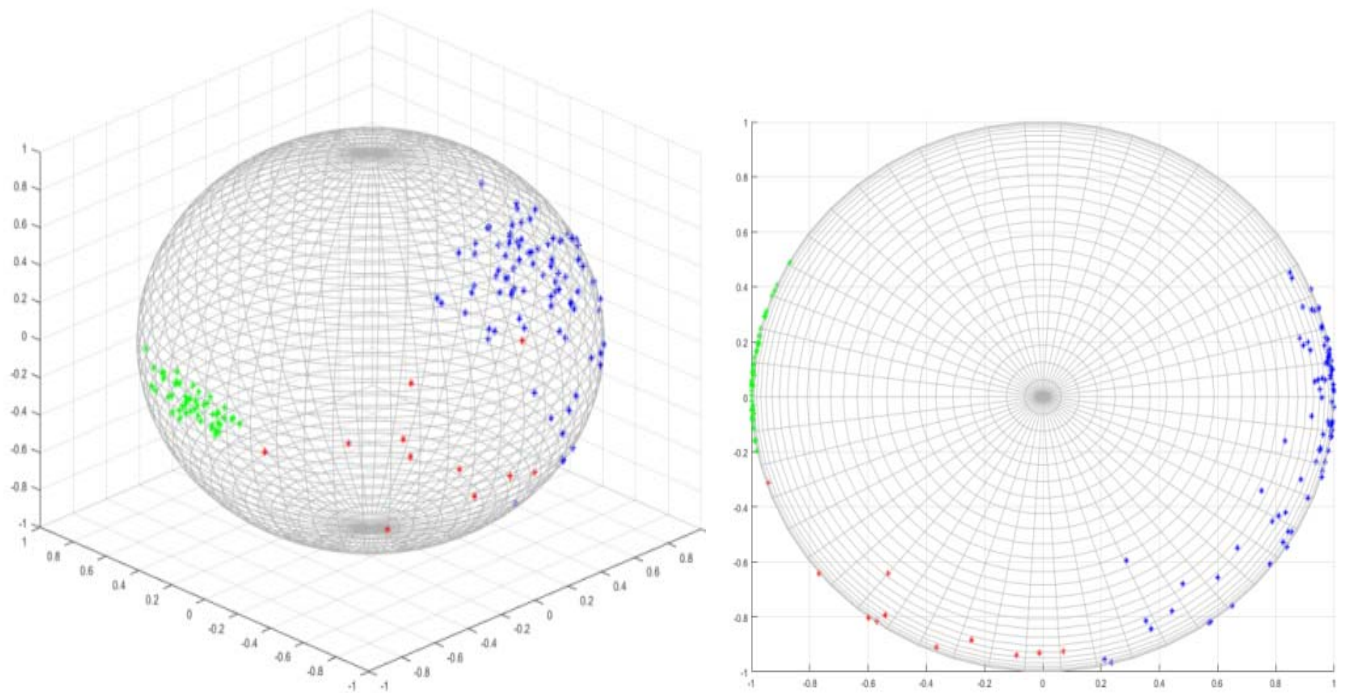
TABLE II. THE MEAN CLUSTERING ACCURACIES FOR THE DIFFERENT NUMBERS OF CLUSTERS (IRIS)

Method	SOM ^m	k-means
clustering accuracies for 2 clusters	0,84	0,83
clustering accuracies for 3 clusters	0,91	0,87
clustering accuracies for 4 clusters	0,72	0,73

For visualization, taken data sets were projected by the PCA (principal component analysis) method to three principal components. Visualization results of the proposed ensemble are shown in (Fig. 3).



a) The artificial generated data set



b) Data set "Iris"

Fig. 3. Visualization results of the proposed ensemble

V. CONCLUSION

The neural network approach for data stream clustering task, that in online mode are fed to processing in assumption that, neither the number of clusters nor their shape are known, is proposed in the paper. The main idea of this approach is based on the kernel clustering and neural networks ensembles, that consist of the T. Kohonen's self-organizing maps.

The proposed system is characterized by the simplicity of numerical implementation, high speed, and can be used for

solving different tasks of processing data streams in conditions of apriori uncertainty of their properties.

REFERENCES

- [1] G. Gan, Ch. Ma and J. Wu, *Data Clustering: Theory, Algorithms and Applications*. Philadelphia: SIAM, 2007.
- [2] R. Xu and D. C. Wunsch, *Clustering*. IEEE Press Series on Computational Intelligence. Hoboken, NJ: John Wiley & Sons, Inc., 2009.
- [3] C. C. Aggarwal and C. K. Reddy, *Data Clustering. Algorithms and Application*. Boca Raton: CRC Press, 2014.

- [4] D. Pelleg, and A. Moor, "X-means: extending K-means with efficient estimation of the number of clusters," 17th Int. Conf. on Machine Learning, Morgan Kaufmann, San Francisco, pp.727-730, 2000.
- [5] T. Ishioka, "An expansion of X-means for automatically determining the optimal number of clusters," 4th IASTED Int. Conf. Computational Intelligence, Calgary, Alberta, pp. 91-96, 2005.
- [6] L. Rutkowski, Computational Intelligence. Methods and Techniques. Berlin-Heidelberg: Springer-Verlag, 2008.
- [7] C. Mumford and L. Jain, Computational Intelligence. Collaboration, Fuzzy and Emergence. Berlin: Springer-Verlag, 2009.
- [8] R. Kruse, C. Borgelt, F. Klawonn, C. Moewes, M. Steinbrecher and P. Held, Computational Intelligence. A Methodological Introduction. Berlin: Springer, 2013.
- [9] K.-L. Du and M. N. S. Swamy, Neural Networks and Statistical Learning. London: Springer-Verlag, 2014.
- [10] T. Kohonen, Self-Organizing Maps. Berlin: Springer-Verlag, 1995.
- [11] A. Strehl, J. Ghosh, "Cluster ensembles – A knowledge reuse framework for combining multiple partitions," Journal of Machine Learning Research, pp. 583-617, 2002.
- [12] A. Topchy, A. K. Jain, and W. Punch, "Clustering ensembles: models of consensus and weak partitions," IEEE Transactions on Pattern Analysis and Machine Intelligence, pp. 1866-1881, 2005.
- [13] H. Alizadeh, B. Minaei-Bidgoli, and H. Parvin, "To improve the quality of cluster ensembles by selecting a subset of base clusters", Journal of Experimental & Theoretical Artificial Intelligence, pp. 127-150, 2013.
- [14] M. Charkhabi, T. Dhot, and S. A. Mojarad, "Cluster ensembles, majority vote, voter eligibility and privileged voters", Int. Journal of Machine Learning and Computing, vol. 4, no. 3, pp. 275-278, 2014.
- [15] Ye. V. Bodyanskiy, A. A. Deineko, P. Ye. Zhernova, and V. O. Riepin, "Adaptive modification of X-means method based on the ensemble of the T. Kohonen's clustering neural networks," VI Int. Sci. Conf. "Information Managements Systems and Technologies", Odessa, Ukraine, pp. 202-204, 2017.
- [16] J. C. Bezdek, J. Keller, R. Krishnapuram and N. Pal, Fuzzy Models and Algorithms for Pattern Recognition and Image Processing. The Handbook of Fuzzy Sets. Kluwer, Dordrecht, Netherlands: Springer, vol. 4, 1999.
- [17] T. M. Cover, "Geometrical and statistical properties of systems of linear inequalities with applications in pattern recognition," IEEE Trans. on Electronic Computers, no. 14, pp. 326-334, 1965.
- [18] M. Girolami, "Mercer kernel-based clustering in feature space", IEEE Trans. on Neural Networks, vol. 13, no. 3, pp. 780-784, 2002.
- [19] D. MacDonald and C. Fyfe, "Clustering in data space and feature space," ESANN'2002 Proc. European Symp. on Artificial Neural Networks, Bruges (Belgium), pp. 137-142, 2002.
- [20] M. Girolami, "Mercer kernel-based clustering in feature space," IEEE Trans. on Neural Networks, vol. 13, no. 3, pp. 780-784, 2002.
- [21] F. Camastra, and A. Verri, "A novel kernel method for clustering," IEEE Trans. on Pattern Analysis and Machine Intelligence, no. 5, pp. 801-805, 2005.
- [22] Ye. V. Bodyanskiy, A. A. Deineko, and Y. V. Kutsenko, "On-line kernel clustering based on the general regression neural network and T. Kohonen's self-organizing map," Automatic Control and Computer Sciences, 51(1), pp. 55-62, 2017.
- [23] D. L. Davies, and D. W. Bouldin, "A Cluster Separation Measure", IEEE Transactions on Pattern Analysis and Machine Intelligence. No. 4, pp. 224-227, 1979.
- [24] P. M. Murphy, and D. Aha, UCI Repository of machine learning databases. URL: <http://www.ics.uci.edu/mllearn/MLRepository.html>. Department of Information and Computer Science

MEMS-based Inertial Sensor Signals and Machine Learning Methods for Classifying Robot Motion

Igor Nevlydov

Department of Computer-Integrated Technologies, Automation and Mechatronics

Kharkiv National University of Radio Electronics

Kharkiv, Ukraine

igor.nevliudov@nure.ua

Mariya Volkova

Department of Computer-Integrated Technologies, Automation and Mechatronics

Kharkiv National University of Radio Electronics

Kharkiv, Ukraine

mariia.volkova@nure.ua

Oleksandr Filipenko

Department of Computer-Integrated Technologies, Automation and Mechatronics

Kharkiv National University of Radio Electronics

Kharkiv, Ukraine

oleksandr.filipenko@nure.ua

Ganna Ponomaryova

Department of Computer-Integrated Technologies, Automation and Mechatronics

Kharkiv National University of Radio Electronics

Kharkiv, Ukraine

ganna.ponomaryova@nure.ua

Abstract—Robot state classification using machine-learning methods and MEMS sensors data is proposed in the paper. An experiment was performed with a three-axis MEMS gyroscope rigidly fixed to the robot body. In it we investigated the possibilities of various machine-learning methods for solving classification task.

Keywords—robot, MEMS, classification, control

I. INTRODUCTION

MEMS sensors play a major role in the robotics and mechatronics due to their miniature size, low cost and sensitivity. The use of these sensors opens possibilities of classification features of the robot movement, balance control system [1-2].

Development of our project "PromoRobot" is caused by the need for high-quality information support, promotion (promotion) of services in places of mass presence of people, including airports, railway stations, business centers, hotels, libraries, exhibitions, educational institutions, government institutions, etc. and takes into account the public interest in robotics, new information technologies, artificial intelligence tools.

Generally, the "PromoRobot" control system corresponds to the concept of an intelligent robotic agent with a feedback control. But in some cases, the data about the robotics system is not enough to correctly work out the task. First of all, it is the task of moving on complex surfaces: up or down a slope, uneven surfaces, etc.

The solution of the task should be to create an algorithm for classifying such states of the robot that corresponded to these complexities. This will allow to take into account these features in the autonomous robot control system.

II. EASE OF USE

A. Robot Specifications

The robotic platform has a two-wheeled chassis whose elements are shown in Fig. 1. For our experiments, we use module MPU-9265. The MPU-9265 devices combine a 3-axis gyroscope, 3-axis accelerometer and 3-axis compass in

the same chip together with an onboard Digital Motion Processor capable of processing the complex MotionFusion algorithms [2]. The output signals of the accelerometers (A_x , A_y , A_z) and the gyros (w_x , w_y , w_z) are converted directly by an Analog to Digital Converters inside the microcontroller ATMEGA32. This microcontroller has 8 channels of 10-bit Analog to Digital Converters, a USART (Universal Asynchronous serial Receiver and Transmitter) port and a sampling rate - 200 Hz.

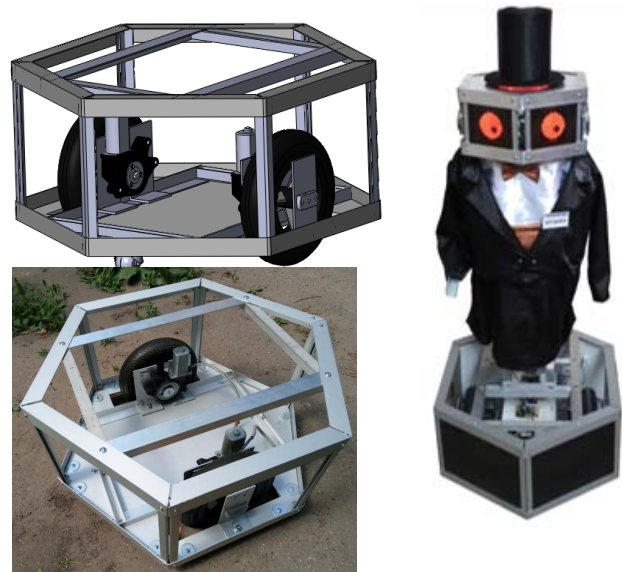


Fig. 1. "PromoRobot" project

B. Classification task

Formally, the problem of robot states classification can be represented as follows: let X be the set of data on the work state obtained from the MEMS gyroscope (along the O_x , O_y , O_z axes). Y is a finite set of classes (8 states in the work): calmness - state "0"; forward motion on the slope - state "1"; backwards motion from the slope - state "2"; backwards motion on the slope - state "3"; forward motion from the slope - state "4"; forward motion - state "5"; rotation counter-clockwise - state "6"; clockwise rotation - state "7".

There is an unknown target addiction – reflection $y^*:X \rightarrow Y$. The value is known only on known states of the robot on the training set $X_m = \{(x_1, y_1), \dots, (x_m, y_m)\}$.

It is necessary to develop an algorithm $a: X \rightarrow Y$ for classifying the robot's state Y according to sensor's reading $x \in X$ in real time-domain. In our case, the set of classes is $Y = \{0, 1, 2, 3, 4, 5, 6, 7\}$.

A number of experiments were carried out with a three-axis MEMS gyroscope rigidly fixed to the robot body. Gyroscope allows you to track the robot's precise execution of the prescribed actions, possible features of its movement. Every action is a certain state.

III. RESULTS & ANALYSIS

The results of measurements are shown in Fig.2-4 and visualize the sensors measure in the process of robot moving. To find the clustering algorithm which is support real-time work we consider three axis of the gyroscope.

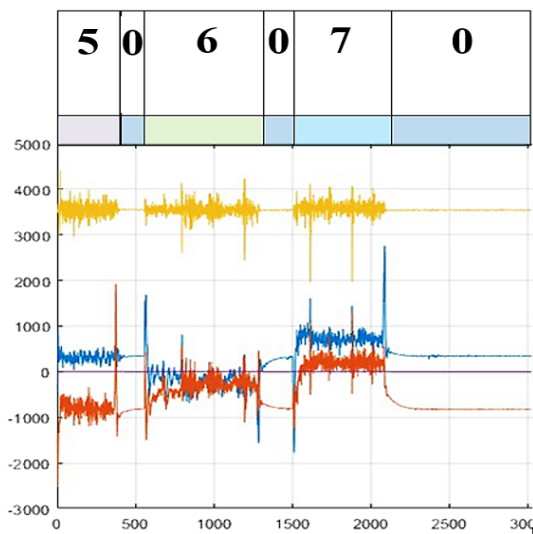


Fig. 2. " Procedure "moving forward - stop - rotation counterclockwise - stop - clockwise rotation - stop"

Figure 2 illustrates the gyroscope readings captured by the three axes. A detail visual analysis of figure 2 reveals that the relative magnitudes of the sub readings of the gyroscope could be used for event classification. For example from point 1 to point 400, robot was moving forward. And from 401 to 520 robot standing – this is indicated by the very low gyroscope values.

However it is also clear that coming up with manually defined thresholds for three sensor readings that will allow the classification of the seven events will be still a complex task. Further the raw signals captured by the sensors are noisy and will therefore have to be cleaned prior to further analysis.

The robot motion activity recognition system has to decide which of the seven events have effectively caused the measured values of the features based on real signals, which are fed from sensors. This is a general classification problem that can be dealt with by a large range of algorithms, such as logics, k-nearest Neighbor approaches, Support Vector Machines (SVMs), Artificial Neural networks [3,4], Decision Trees or Bayesian Techniques [5]. Our work

focused on finding an algorithm which are realize a classification of robot motion in real-time domain.

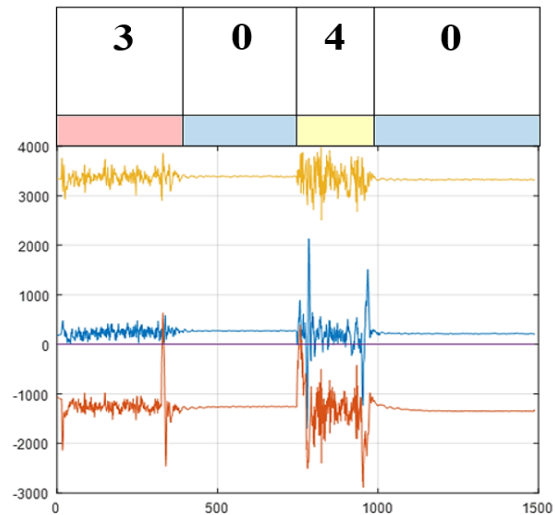


Fig. 3. Procedure "backwards motion on the slope – stop – forward motion from the slope, stop»"

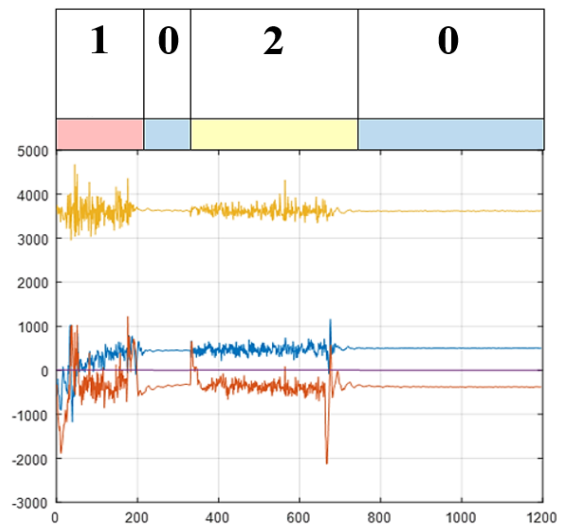


Fig. 4. Procedure "forward motion on the slope – stop – backwards motion from the slope - stop"

The three time-domain features (three axis of the gyroscope) were used to train machine learning algorithm. The mean, standard deviation, minimum, and maximum of signal in the running window also were used as features. But using more features did not improve the quality of the classification

We examined the performance of some supervised learning algorithms and singled out most appropriate among them: Support Vector Machines (Linear SVM) [3], k-nearest neighbors algorithm (Medium KNN, Weighted KNN) [4], Boosting algorithm [6], Classification Trees (Simple Tree, Medium Tree) and Ensemble (Bagged trees) [7, 8].

Gyroscope signals from robot are sufficient to classification. Weighted KNN and Bagged trees performed slightly better than other three algorithms (the classification accuracy about 89%).

The evaluation of the quality of the trained models was carried out by such criteria as accuracy, confusion matrix,

Parallel Coordinates plot ROC Curve. Results of model test can be classified according to sensitivity and specificity. Sensitivity is the ability to detect an abnormality, while specificity is the ability to distinguish an abnormality by type. Diagnostic test should also identify the frequency of false positive (FP) and false negative (FN) or true positive (TP) and true negative (TN).

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + TN + FN}$$

Table I presents the results of the accuracy calculation for various methods.

TABLE I. RESULTS OF THE STUDY BY VARIOUS METHODS

Metod	Accuracy,%
Linear SVM	66.6
Simple Tree	69.7
Medium Tree	76.7
Medium KNN	88,1
Weighted KNN	88.7
Boosted Tree	84,1
Bagged Trees	89,6

The data in an ROC analysis is used to decide which traits produce the greatest separation of two probability curves which show the likelihood of choosing wrong or right states. The standard way to interpret the data from an ROC test is to draw a ROC-curve and then measured the area under the curve. The test with the greatest area is the most accurate. The best results were shown by the Weighted KNN method.

In Fig. 5 and Fig. 6 graphs of confusion matrix are provided, which help to identify areas in which the classifier works poorly. In the first case, the lines show the current state of work, and the column shows the cjjnd classes. As can be seen from Fig. 5 the classifier works worse for determining the class "2" (backwards motion from the slope) and class "3" (backwards motion on the slope).

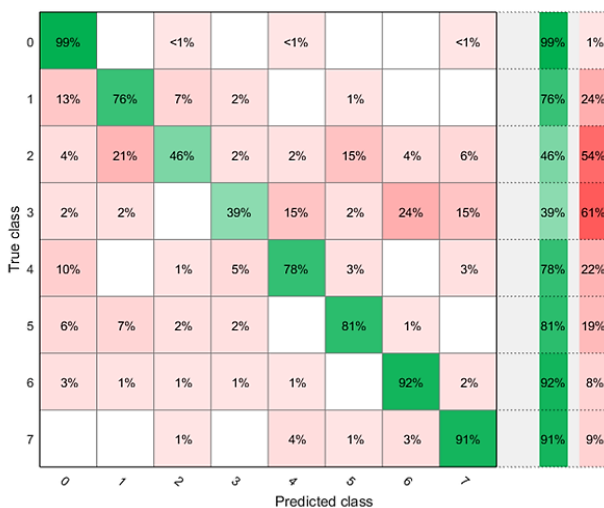


Fig. 5. Confusion matrix Weighted KNN

In fig. 3.9 in the Confusion matrix are shown false classifier actions, under the matrix green, the correct prediction is shown in each class, and the false values are

shown in red. The marker on the Fig.7 shows the performance of the currently selected classifier. For our classifier false positive rate (FPR) of 0.05 indicates that the current classifier assigns 5% of the observations incorrectly to the positive class. A true positive rate of 0.99 indicates that the current classifier assigns 99% of the observations correctly to the positive class.

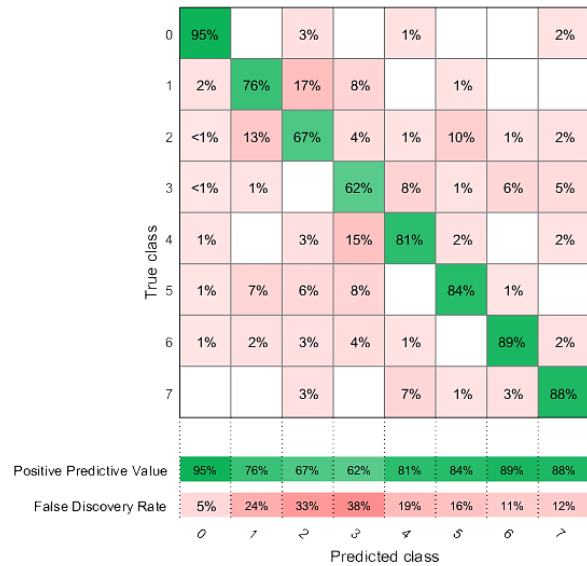


Fig. 6. Confusion matrix Weighted KNN

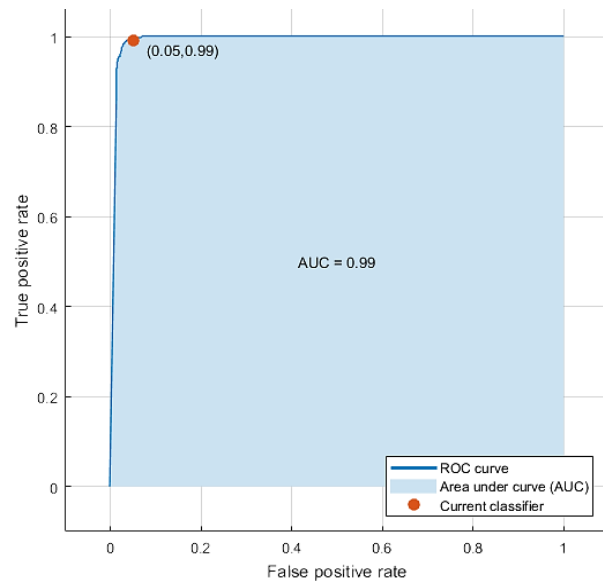


Fig. 7. ROC Curve

IV. CONCLUSIONS

The paper is devoted to solving tasks of classifying robot motion by machine learning methods. The experiments was performed based on real signals are fed from MEMS sensors on the robot board in real-time domain.

The analyzze of classificator learning results showed the possibility of using k-nearest neighbors algorithm to classify the state of a robot with 88% accuracy. An algorithm is developed based on measurements of a three-axis gyro without any pre calculations.

We are currently working on integrating a number of other in-built sensors and algorithms in the above process, allowing more detailed and complex scenarios to be identified accurately. Further development of the proposed approach can be carried out in the direction of implement the classifier in decision-making system of robot on Asus Tinker Board.

TABLE II. RESULTS OF CLASSIFIER WORK

State of robot	Weighted KNN		Bagged trees	
	Positive Predictive Value	True Positive Rate	Positive Predictive Value	True Positive Rate
calmness	95%	99%	96%	98%
forward motion on the slope	76%	76%	84%	84%
backwards motion from the slope	67%	46%	64%	56%
backwards motion on the slope	62%	39%	50%	34%
forward motion from the slope	81%	78%	81%	81%
forward motion	84%	81%	80%	82%
rotation counter-clockwise	89%	92%	91%	93%
clockwise rotation	88%	91%	91%	92%

REFERENCES

- [1] F. Coito, A. Eleutério, S. Valtchev, and F. Coito, "Tracking a Mobile Robot Position Using Vision and Inertial Sensor," 5th IFIP WG 5.5/SOCOLNET DoCEIS 2014, Costa de Caparica, Portugal, AICT-423, Springer, pp. 201-20, April 7-9, 2014.
- [2] Inven Sense. MPU-9250 Product Specification, Revision 1.1, InvenSense Inc. [ONLINE] Available at: <https://www.invensense.com/wp-content/uploads/2015/02/PS-MPU-9250A-01-v1.1.pdf>. [Accessed 26/12/2017].
- [3] K. Noda, Y. Hashimoto, Y. Tanaka, and Ichiro Shimoyama, "MEMS on robot applications," TRANSDUCERS 2009 International Solid-State Sensors, Actuators and Microsystems Conference [ONLINE] Available at: <https://ieeexplore.ieee.org/document/5285608/>
- [4] K. Frank, J. Vera Nadales, P. Robertson, and M. Angermann, "Reliable Real-Time Recognition of Motion Related Human Activities Using MEMS Inertial Sensors," [ONLINE] Available at: https://pdfs.semanticscholar.org/89ca/d05d53302b4b8c465f3fe9b9eec924aff567.pdf?_ga=2.160054457.507168460.1529006792-191559443.1528161624/
- [5] S. B. Kotsiantis, "Supervised machine learning: A review of classification techniques," Informatica, vol. 31, pp. 249–268, 2007.
- [6] C. Cortes, and V. Vapnik, "Support-vector network," Machine Learning, vol. 20, issue 3, pp. 273-297, Sept. 1995. [Online]. Available: <https://doi.org/10.1023/A:1022627411411>.
- [7] J. Zhu, S. Rosset, H. Zou, and T. Hastie, "Multiclass AdaBoost," Technical report, Stanford Univ, 2005. Available at <http://www-stat.stanford.edu/hastie/Papers/samme.pdf>.
- [8] Y. Freund, and R. E. Schapire, "Experiments with a New Boosting Algorithm," in L.Saitta, ed., 'Proceedings of the Thirteenth International Conference on Machine Learning (ICML'96)', Morgan Kaufmann, 1995, pp. 148–156.
- [9] E. Gatnar, "Fusion of Multiple Statistical Classifiers", in C. Preisach, H. Burkhardt, L. Schmidt-Thieme and R. Decker, eds, "Data Analysis, Machine Learning and Applications," Studies in Classification, Data Analysis, and Knowledge Organization, Springer, Berlin/Heidelberg, 2008, pp. 19–27.

Data Science in Open-Access Research On-line Resources

Dmytro Lande
*Institute for information recording
National academy of science of Ukraine*
Kyiv, Ukraine
dwlände@gmail.com

Valentyna Andrushchenko
*State fund for fundamental research of
Ukraine*
*Institute for information recording
National academy of science of Ukraine*
valentyna.andrushchenko@gmail.com

Iryna Balagura
*Institute for information recording
National academy of science of Ukraine*
Kyiv, Ukraine
balaguraira@gmail.com

Abstract— The data science methods are widely used in different areas of nowadays life. This paper is dedicated to forming of new approaches, in particular – development of unique model, to provide scientometric research of abstracts from open source pre-print service to process data on subject domains and directions. The objective of research is to analyze degree of presence and importance of data science in different research fields. A new way of working with the information system of the library of the University of Cornelius - the resource of the pre-prints arXiv is proposed in the work. The authors reviewed the abstract information of the resource, which is the result of the search for relevant publications for the given concept. The main attention of the authors was focused on the distribution of publications in the identified scientific areas and the relevant sub-groups provided by the resource. The result of the work is a visual representation and interpretations of the network of subject areas for the concepts - big data, neural networks, deep learning.

Keywords—*scientometric, big data, data science, concept, subject domain, network, scientific papers.*

I. INTRODUCTION

Today the problem of recording, processing and storage of information is actual for every field [1]. Big Data has become important for organization everyday wellbeing and using satellite data for forecasting weather, traffic jams, nature disasters [2]. Data science and Big data influence business and sales. Big data could be used for politic companies and for prediction of the stock fluctuating of a certain company [3, 4]. And even farming processes transformed into Smart Farming with machines that are equipped with smart sensors and devices and produce big amounts of data that provide unprecedented decision-making capabilities [5]. It attracts more consumers focused on innovations in goods production and services. Now we have possibilities to use smart transport without drivers, smart houses, Internet of things and Cloud Computing and fill more comfortable with data science development [6]. Data science is growing but still contain challenges: Data challenges (e.g. data volume, variety, velocity, veracity, volatility, quality, discovery and dogmatism); process challenges; management challenges (privacy, security, governance and ethical aspects) [7]. The Big Data Analytics requires new advanced algorithms such as text analytics, machine learning, predictive analytics, data mining, statistics, and natural language processing [1]. There are no statistical research of data science usage and real state of big data evolving in science [7]. This paper is dedicated to scientometric research of abstracts from open source pre-print service in main fields

detection of data science usage. The objective of research is to analyze degree of presence and importance of data science in different research fields. The source of data is open-access on-line recourse arxiv (www.arxiv.org). It includes 1,372,745 e-prints in Physics, Mathematics, Computer Science, Quantitative Biology, Quantitative Finance, Statistics, Electrical Engineering and Systems Science, and Economics. The source allows to design visualization of subfields links for selected concepts with maps using. The concept map of subject domains is an useful instrument for identification of related topics in scientific research, detection of trends in research, definition of its terminology, correct usage and correct application of keywords in scientific works.

II. METHODOLOGY

We propose to visualize data science integration to different research with networks theory, which was created with Euler (1736) the Konigsberg Bridge problem and presented as mathematical notation of nodes and edges [8]. In scientometrics network theory is widely used for mapping of science, among them: co-citation, co-author and co-word networks. Co-word and co-author networks could be used for identification and description of scientific groups and research topics, the most communicative researchers and main principles of science communication [9]. We propose to use maps for fields and subfields connection according to certain concepts which is necessary in datascience.

A. Data input

To provide the correct analysis of obtained information from the point of view of completeness and variety of research fields and direction the open access archive of preprint arXiv was chosen. ArXiv is the largest archive of electronic publications and their open preprints.

The archive was created in 1991. Initially all the publication on archive were allocated in frames of one subject domain - "Physics", but today the resource presumes arrangement of publications within other directions.

ArXiv is an information tool for hundreds and thousands of scholars. Among the users are more than 50 Nobel laureates, winners of prestigious scientific awards. Resource is an actual tool for users from countries with limited access to scientific information. Today, the archive contains 8 sections, where you can post your own materials: Computer Science (42 areas), Economics (1 direction), Electrical Engineering and System Science (3 courses), Mathematics (32 directions), Physics (13 units), Quantitative Biology (10

directions), Quantitative Finance (9 destinations), Statistics (6 destinations).

We will use abstract information which contains the following data:

- Number of article, identifier in the system, in the form:
arXiv: XXXX.XXXXXX [***], where the HTML code is the publication number in the system, *** - the list of available file formats for download;
- Topic of the publication;
- Author (s);
- Comments - contains information about the number, pages of publication, number of drawings and other items (not urgent);
- Journal-ref – contains information about the paper (available for publications that have already been published);
- Subject - the subject area or specific information on the scientific direction within the scope of the subject area (according to how the author of the publication noted during the presentation of the publication to place it on the resource).

B. Algorithm

Under the concept we will understand the meaningful verbal unit, or a combination of units, which defines the framework of scientific perception of the meaning of a particular notion that is appropriate to one or more subject areas [10]. The network of subject areas is a way of presenting a model of subject areas by defining generalized descriptions of the domain, represented by their proper name and the names of subordinate units of it, the scientific directions that more specifically describe the subject area defined by the information system on the basis of which the given network is constructed or on the basis of the proposed systematization of subject areas [11]. The search is provided for the concept which can be represented as a word or word combination through the array of resource publications. The algorithm for constructing a subject areas network for a given concept involves the definition of subject areas and scientific directions for which the given concept is appropriate. The implementation of the algorithm is realized by processing search results. We use abstract of paper, key-words and sub-fields. So, we will define nodes as fields and sub-fields and edges as co-occurrence of sub-fields in one paper (Fig.1).

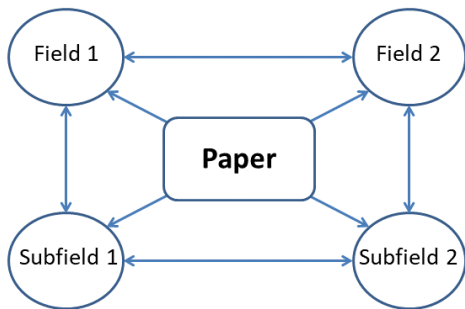


Fig. 1. Example of concept network building and fields connections

The algorithm of subject domain network building consist of next stages:

- The concept definition for the search.
- Extraction of abstract information.
- The scientific direction detection (sub-field) and subject domain (field), which is indicated in the abstract information.
- The scientific direction is the next node graph and connection with appropriate subject domain.
- If a scientific direction node has already been constructed - the name of the subject area, then only the node is constructed - the name of the scientific direction, which is connected with the node - the corresponding subject area.
- If a corresponding node has already been built for the name of the scientific direction, then the transition to upper steps. If the name of the scientific direction nodes has not yet been made, then upper steps (Fig.2)
- If there is no suched results, the network is considered to be built.

We use networks characteristics for networks analysis. Number of nodes, edges and density of network could be applied for detection of widely used terms in different fields. The density of network is the ratio of existing links to the total number of possible links. For a network of N nodes, the network link density is

$$\rho = \frac{2e}{n(n-1)} \quad (1)$$

where e – number of edges, n – number of nodes.

The (maximal) link density of a completely connected network is 1. We will admit that the lower the density of the network - the higher the polythematism.

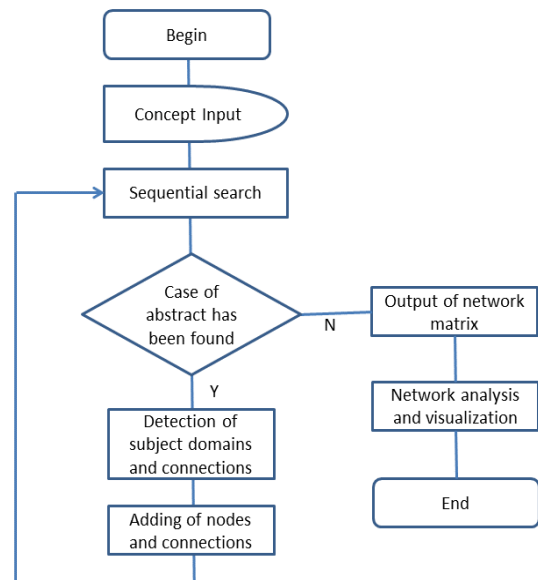


Fig. 2. The algorithm of subject domain network building

III. MAIN RESULTS

Developed software and the algorithm which was proposed above we built the networks for the concepts: big data, neural networks, deep learning, and complex network. Main parameters of the networks are shown in table 1. The largest number of nodes and smallest density among proposed networks are for concept “neural network”. It confirms the prevalence of the method in different fields. The visualizations were provided with the Gephi software (gephi.org). The concept “big data” refers mostly to computer science (CS) and statistics (Stat) fields, few articles in mathematics (math) and physics (fig.3.). Main connected sub-fields: Mashine learning (CS.LG); Distributed, Parallel, and Cluster Computing (CS.DC); Data Structures and Algorithms (CS.DS); Computational Engineering, Finance, and Science (CS.CE); Social and Information networks (CS.SI); Artificial Intelligence (CS.AI); Information retrieval (CS.IR); Networking and Internet architecture (CS.NI); Computation in statistics (Stat.CO); Methodology in statistics (Stat.ME) and others. The amount of sub-field is not low – 35. But it was expected to observe “big data” concept applied to such a wide spread fields as biology, finance, astronomy research. We can assume the reasons of absence of such research directions could be caused by author key-words missing, actively development of data science theoretical and fundamental laws, insufficiency of data base or practical industry usage etc.

The concept “neural networks” is represented in different sub-fields among them are: computer science (CS); statistics (Stat); Physics; Mathematics (Math); Quantitative Biology (q-bio); Electrical Engineering and Systems Science (eess); Quantitative Finance (q-fin); Econometrics (Econ.EM). So we can draw a clear conclusion that concept “neural networks” is widely used instrument in different fields. “Neural networks” and “deep learning” currently provide the best solutions in image recognition, speech recognition, and natural language processing. We determined the range of scientific directions related to the given concept by scanning the largest resource in the global network of preprints, containing a large amount of publications both prepared for printing and placed in the leading scientific publications. Developed applications by the proposed algorithm will allow using the network of subject areas as an additional tool for finding collaborators, expanding the use of the concept within different scientific areas and thus obtaining the opportunity for expanding collaborations and attracting specialists from various scientific fields.

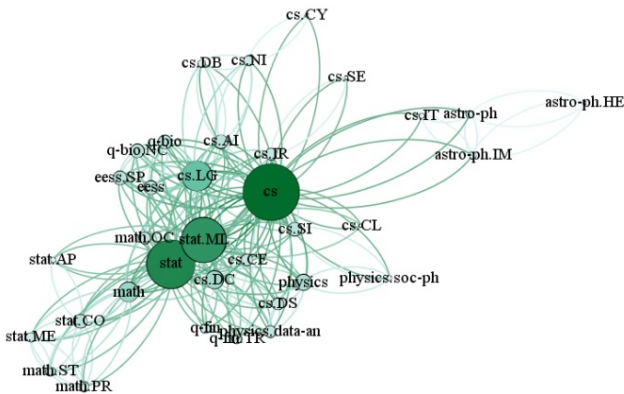


Fig. 3. Subject domain network for Big data concept

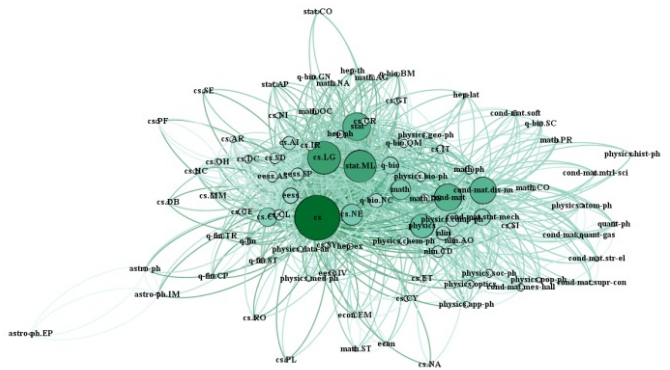


Fig. 4. Subject domain network for Neural network

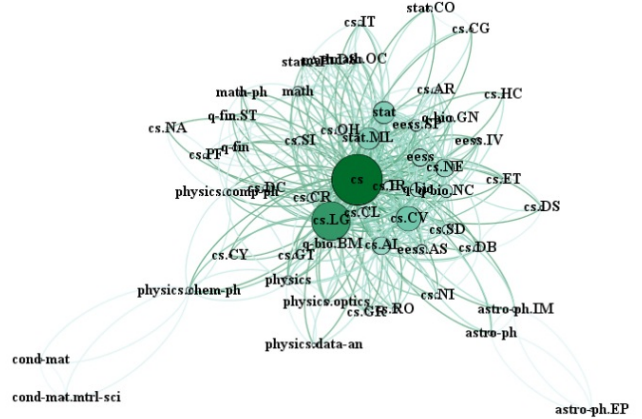


Fig. 5. Subject domain network for Deep learning

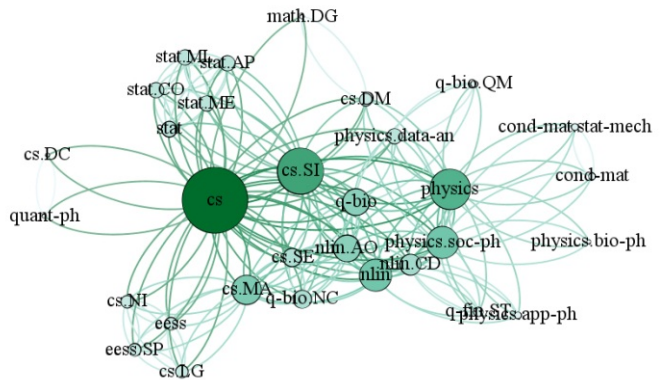


Fig. 6. Subject domain network for Complex networks

TABLE I. MAIN PARAMETERS OF THE SUBJECT DOMAIN NETWORKS

Concept	Density of nodes	Number of nodes	Number of edges
Neural network	0.136	90	1090
Deep learning	0.159	54	454
Big data	0.213	35	254
Complex network	0.234	31	218

For the development of the proposed approaches for the search, processing and interpretation of scientific information through the implementation of these algorithms, it is possible to construct a more developed network by grouping the names of scientific areas within dictionaries, as well as calculating network parameters.

IV. CONCLUSIONS

The algorithms, which based on subject domain mapping for certain concept are proposed. In maps we used nodes as fields and sub-fields and edges as co-occurrence of sub-fields in one paper. We offer to use the concept map of subject domains for identification of related topics in scientific research, detection of trends in research, searching for the ambiguity of terminology, correct usage of terms and describing science structure. Provided algorithm isn't strongly connected to the field of research or source of data and could be continued with other examples.

This paper is dedicated to forming of new approaches, such as new unique models for providing the scientometric research based on the open access archive of preprint to isolate and process the data connected to the subject domains of the publications and appropriate research directions. The objective of research is to analyze the level of representativeness and importance of data science in different research fields. We showed that the data science integration to different research with example of concepts "big data", "deep learning", "neural networks", "complex networks" using one of the biggest open access archive. We used density of complex network for estimation of the widest in the sense of concepts usage. Main subject area for selected concepts is computer science. The most common concept is "neural networks" which is used in 90 different sub-fields, other concepts mostly used in computer science. The theory of complex networks decreased inherency in comparing with neural networks and contains in 31 research fields.

REFERENCES

- [1] M.M. Najafabadi, F. Villanustre, T.M. Khoshgoftaar et al., "Deep learning applications and challenges in big data analytics," *Journal of Big Data*, Springer, vol. 2, iss. 1, pp. 1-21, 2015. <https://doi.org/10.1186/s40537-014-0007-7>
- [2] P. Thakuria, N. Y. Tilahun, and M. Zellner, "Big Data and Urban Informatics: Innovations and Challenges to Urban Planning and Knowledge Discovery," *Seeing Cities Through Big Data*, Springer, NY, pp.11-45, 2017.
- [3] Q. Li, Y. Chen, J. Wang, Y. Chen, and H. Chen, "Web Media and Stock Markets : A Survey and Future Directions from a Big Data Perspective," *IEEE Transactions on Knowledge and Data Engineering*, vol. 30, pp.381-399, Feb. 1 2018.
- [4] Eitan D. Hersh, and Brian F. Schaffner, "Targeted Campaign Appeals and the Value of Ambiguity," *The Journal of Politics*, The University of Chicago Press, vol. 75, iss. 2, pp.520-534, April 2013.
- [5] Sjaak Wolfert, Lan Ge, Cor Verdouw, and Marc-Jeroen Bogaardt, "Big Data in Smart Farming – A review," *Agricultural Systems*, vol.153, pp. 69-80, 2017.
- [6] M. Mazhar Rathore, Awais Ahmad, Anand Paul, and Seungmin Rho, "Urban planning and building smart cities based on the Internet of Things using Big Data analytics," *Computer Networks*, vol. 101, pp. 63-80, 2016.
- [7] U. Sivarajah, M. Mustafa Kamal, Z. Irani, and V. Weerakkody, "Critical analysis of Big Data challenges and analytical methods," *Journal of Business Research*, vol. 70, pp. 263-2862, 2017.
- [8] V. Kale, *Big data computing: A Guide for Business and Technology Managers*. Taylor and Francis Group, CRC Press, 2017.
- [9] D. V. Lande, I. V. Balagura, and V. B. Andrushchenko, "The detection of actual research topics using co-word networks," *Open Semantic Technologies for Intelligent Systems : proceedings*, Minsk, BNUIR, pp. 207-210, 2018.
- [10] J. C. Hayes, and D. J. M. Kraemer, "Grounded understanding of abstract concepts: The case of STEM learning." *Cognitive Research*, vol. 2, iss.1, 2017. doi:10.1186/s41235-016-0046-z.
- [11] D. V. Lande, V. B. Andrushchenko, and I. V. "Balagura Formation of the Subject Area on the Base of Wikipedia Service," *Open Semantic Technologies for Intelligent Systems : proceedings*, Minsk, BNUIR, pp. 211-214, 2017.

Building the Semantic Similarity Model for Social Network Data Streams

Svitlana Petrasova
National Technical University "Kharkiv
Polytechnic Institute"
Kharkiv, Ukraine
svetapetrasova@gmail.com

Nina Khairova
National Technical University "Kharkiv
Polytechnic Institute"
Kharkiv, Ukraine
khairova@kpi.kharkov.ua

Włodzimierz Lewoniewski
Poznan University of Economics and
Business
Poznan, Poland
wlodzimierz.lewoniewski@ue.poznan.pl

Abstract— This paper proposes the model for searching similar collocations in English texts in order to determine semantically connected text fragments for social network data streams analysis. The logical-linguistic model uses semantic and grammatical features of words to obtain a sequence of semantically related to each other text fragments from different actors of a social network. In order to implement the model, we leverage Universal Dependencies parser and Natural Language Toolkit with the lexical database WordNet. Based on the Blog Authorship Corpus, the experiment achieves over 0.92 precision.

Keywords— social network; data stream; collocations; semantic similarity; blogs; corpus; Universal Dependencies; WordNet

I. INTRODUCTION

In the last years, social media became a source of communication, data distribution, and an aspect of formation of an informal information space. Many business companies and intelligence agencies have turned to computer processing to monitor these social streams [1].

Main objects of the modern information society are social networks, forums, blogs, etc. Processing such data streams as these, the following factors should be considered: instability of content quality, e.g. spam and fake accounts, and problems with the privacy of users' personal data. All of this requires constant improvement of algorithms for analysis and processing of social data streams.

One of the approaches for studying online social structures is Social Network Analysis. Its main objectives are investigation of interactions between social actors and identification of the conditions for the emergence of these interactions [2, 3]. This way, the network of social interactions consists of a finite set of social actors and a set of links between them [4].

Nowadays, the main methods for analyzing social networks are: (1) methods of graph theory for studying the structural relationships of an actor; (2) methods for determining the equivalence of actors; (3) probabilistic models; (4) topological methods that represent the network in the form of some formalized complex of elements and links, etc.

However, we suppose that the use of NLP approaches is important for processing social data streams represented by actors' text information. To date analysing texts of social networks is one of the biggest challenging tasks in NLP. Despite existent NLP applications for IE [5], it is difficult to

extract relevant information from the streams of informal natural language sources.

In the scope of semantic processing of such texts stream as posted by people in public forums (Facebook, Twitter, LinkedIn, Google+), blogs, etc., we aim to obtain a sequence of semantically related to each other text units from different actors of a social network. In order to solve the issue, we suggest extracting semantically similar units of various levels of the language, i.e. analyzing not only syntactical relations between words or sentences but also semantic correlations between words, phrases and collocations. However, there are currently enough studies concerning the problems of computing words similarity, but relatively few researches are carried out into extracting semantic similar phrases or collocations from natural language texts.

A collocation means a combination of two or more words often used together and both syntactically and semantically integrated. In contrast to certain words that are polysemantic and have synonyms, collocations include more particular semantic information. Therefore, semantic similarity of collocations may better identify semantically similar text fragments of the different social actors.

This paper addresses the problem of searching similar collocations in English texts in order to determine semantically connected text fragments for Social network data streams analysis.

II. RELATED WORK

Nowadays, there are a few approaches to extracting semantically similar collocations from texts. At the stage of determining semantic similarity of collocations they mainly use statistical laws, (recurrent) neural networks (e.g. LSTM networks encode patterns of collocations as vector representations) [6], or syntactic characteristics of collocations.

For instance, in the paper [7] English synonymous collocation pairs are extracted using translation information. This method gets candidates of synonymous collocation pairs based on a monolingual corpus and a thesaurus, and then selects the appropriate pairs from the candidates using their translations in a second language. The other method [8] collects sets of words and paraphrases via pairwise alignment of sentence fragments. Reference [9] presents a corpus-based method for automatic extraction of paraphrases using multiple English translations of the same source text.

Generally, all of these studies work on texts of certain domains and take semantic information from thesauri that

result in a quite low precision of extraction of semantically similar collocations.

In our research we suggest the technology for extraction of semantically similar collocations considering the combination of statistical, syntactic, and semantic information for three main types of collocations.

III. THE MODEL FOR SEMANTIC SIMILARITY OF TEXT FRAGMENTS

We introduce a technology for automatic extraction of semantically similar collocations in the English language using both a method for extraction of paradigmatic correlations (tolerance and equivalence) and a logical-linguistic model of identification of synonymous collocations.

According to previous studies [10-11], the proposed logical and linguistic model formalizes semantically similar collocations by means of semantic and grammatical characteristics of collocation words. Basic mathematical means of our model are logical-algebraic equations of the finite predicates algebra.

With reference to the algebra of finite predicates, the set of word forms that make up a collocation is denoted with $M = \{m_1, \dots, m_n\}$, where n is the number of word forms. The word forms from the set M match semantic-syntactic relations using subject variables [12].

We define a set of grammatical and semantic characteristics of collocation words using two subject variables. The subject variable a denotes grammatical characteristics of words in collocations:

$$a^{NSub} \vee a^{NObj} \vee a^{NSubOf} \vee a^{NObjOf} \vee a^{VTr} \vee a^{VIntr} \vee a^{AAtt} \vee a^{APr} = 1, \quad (1)$$

where a^{NSub} is a noun, subject, a^{NSubOf} is a noun, subject, with the preposition "of", a^{NObj} is a noun, object, a^{NObjOf} is a noun, object, with the preposition "of"; a^{AAtt} is an adjective, attribute, a^{APr} is an adjective, predicative; a^{VTr} is a verb, transitive, a^{VIntr} is a verb, intransitive.

The subject variable c denotes semantic roles of the words in collocations:

$$c^{Ag} \vee c^{Att} \vee c^{Pac} \vee c^{Adr} \vee c^{Ins} \vee c^M = 1, \quad (2)$$

where c^{Ag} – an agent, c^{Att} – an attribute, c^{Pac} – a patient, c^{Adr} – an addressee, c^{Ins} – an instrument, c^M – location or content.

We determine predicate $P(x)$ which defines a set of possible semantic and grammatical characteristics for a main collocation word, and predicate $P(y)$ which defines a set of possible semantic and grammatical characteristics for a dependent collocation word. The two-place predicate $P(x, y)$ describes a binary relation of collocation words x and y , which is a subset of the Cartesian product of $P(x) \wedge P(y)$. The predicate determines a correlation of semantic and grammatical information about word forms of two-word collocations:

$$P(x, y) = (x^{NSubAg} \vee x^{NSubOfAg} \vee x^{VTr})(y^{NObjAtt} \vee y^{NObjPac} \vee y^{AAtt} \vee y^{APr}), \quad (3)$$

Using equation (3), we define the predicate of semantic equivalence between two two-word collocations as:

$$P(x_1, y_1) \times P(x_2, y_2) = \gamma_i(x_1, y_1, x_2, y_2) \wedge P(x_1, y_1) \wedge P(x_2, y_2)$$

where: \times indicates semantic similarity, \wedge is the Cartesian product, the predicate γ_i eliminates collocations between which semantic equivalence cannot be identified. The values of the predicate for three main types of collocations are shown in Table I.

TABLE I. THE PREDICATE OF SEMANTIC SIMILARITY

Type of collocations	The predicate γ_i
Adjective-Noun	$\gamma_1(x_1, y_1, x_2, y_2) = y_1^{AAtt} x_1^{NSubAg} \wedge x_2^{NSubAg} y_2^{APr} \vee y_1^{AAtt} x_1^{NSubAg} \wedge x_2^{AAtt} y_2^{NSubAg} \vee x_1^{NSubAg} y_1^{APr} x_2^{NSubAg} y_2^{APr}$ <p>guaranteed outcome ~ assured result</p>
Noun-Noun	$\gamma_2(x_1, y_1, x_2, y_2) = x_1^{NSubOfAg} y_1^{NObjAtt} \wedge x_2^{NObjAtt} y_2^{NSubAg} \vee x_1^{NSubOfAg} y_1^{NObjAtt} x_2^{NSubOfAg} y_2^{NObjAtt} \vee x_1^{NObjAtt} y_1^{NSubAg} y_2^{NObjAtt} x_2^{NSubAg}$ <p>access control ~ admission monitoring</p>
Verb-Noun	$\gamma_3(x_1, y_1, x_2, y_2) = x_1^{VTr} y_1^{NObjPac} \wedge x_2^{VTr} y_2^{NObjPac}$ <p>receive commands ~ obtain instructions</p>

Combining the predicates γ_1 , γ_2 and γ_3 that determine the semantic equivalence of Adjective-Noun, Noun-Noun, and Verb-Noun collocations, the predicate of semantic equivalence between collocations can be defined as follows:

$$\gamma(x_1, y_1, x_2, y_2) = x_1^{VTr} y_1^{NObjPac} x_2^{VTr} y_2^{NObjPac} \vee (x_1^{NSubOfAg} \vee x_1^{NSubAg}) y_1^{NObjAtt} (x_2^{NSubOfAg} \vee x_2^{NSubAg}) y_2^{NObjAtt} \vee x_1^{NSubAg} y_1^{AAtt} \vee x_2^{NSubAg} y_2^{AAtt} \vee x_1^{APr} x_2^{NSubAg} (y_2^{AAtt} \vee y_2^{APr})$$

Thus, using the algebra of predicates, the logical and linguistic model allows formalizing semantic equivalence between natural language constructions, i.e. collocations.

However, lexemes in these constructions must be semantically similar or synonymous in pairs. In order to obtain synonymous collocation words, we use a measure of semantic similarity that is defined as the ratio between the set-theoretic intersection and union of sets of terms of their definitions from glossaries. The measure is based on the

Lesk algorithm [13], according to which two words can be considered as synonyms if they have common words in their dictionary definitions.

For our experiment, as a basis for obtaining definitions of words, we use WordNet, which is the largest lexical database for English. In turn, in order to gain grammatical and semantic features of the words, we use Universal Dependencies (UD) parser.

IV. EXPERIMENTS AND EVALUATION RESULTS

Our dataset is based on the Blog Authorship Corpus [14]. The corpus collected posts of 19,320 bloggers gathered from blogger.com one day. The bloggers' age is from 13 to 47 years. For our purposes, we extract texts of all bloggers (authors) of three age groups: "10s" bloggers (ages 13-17), "20s" bloggers (ages 23-27), and "30s" bloggers (ages 33-47). In order to obtain plain texts of the corpus in the first phase of text pre-processing, we have cleared texts from tags, some other specific characters and made lemmatization¹. As a result, in the phase of pre-linguistic processing, we have gained the text corpus which contains 145 891 559 words (1 280 634 of them are unique). In the next phase, we exploited UD parser, produced by Stanford University, for POS-tagging and syntactic parsing of the corpus texts. In this UD formalism, the syntactic structure of a sentence is described in terms of the words in a sentence and an associated set of directed binary grammatical relations held among words.

Figure 1 shows the example of a sentence dependency parser which is obtained using a special visualization tool for dependency parse - Dependencee².

For our logical-linguistic model, we distinguish six types of the dependency structure which are drawn from a fixed inventory of grammatical relations. These are *compound*, *nmod*, *nmod:possobj*, *obj (dobj)*, *amod* and *nsbj* UD labels.

The first three types of the dependency structure denote directed relations between two nouns. Grammatical and semantic characteristics represented through labels of these types of UD correspond to equation (5).

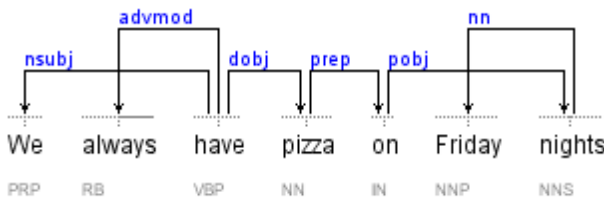


Fig. 1. The example of graphical representation of Universal Dependencies for the sentence from the Blog Authorship Corpus. Source: Dependencee.

Typically, *obj* and *dobj* relations link a verb to a noun. Grammatical and semantic characteristics represented through labels of these relations correspond to equation (4). Finally, *amod* and *cop* (with *nsbj*) relations show directed relations between a noun and an adjective. Grammatical and semantic characteristics represented through these relations correspond to equation (6).

The results achieved through the POS-tagging and UD parser discussed above enabled obtaining grammatical characteristics and semantic roles of the subject variables in equations (1), (2). In this way, we have gained more than 6 mln collocations from the corpus, grammatical and semantic characteristics of which words correspond to equation (3).

In the next step, we used WordNet in order to get synonyms of each word of distinguished collocations. Natural Language Toolkit (NLTK)³ is a free platform that provides interfaces to different corpora and a lexical resource to work with language data. Inclusion of such popular lexical database as WordNet⁴ makes it possible to use the information about similarities between word meanings. In WordNet nouns, verbs, adjectives and adverbs are grouped into sets of cognitive synonyms. These properties render the toolkit suitable for measuring the semantic similarity or relatedness between a pair of concepts (or word senses) [15].

The results of such an approach produced 43 299 words of a total, which have binary relations of synonymy according to WordNet. Based on gained synonymy words and our equations (4) – (6), we generated more than 345 million of potential semantic similarity collocations, about 197 million of which belong to Noun-Noun collocations group, about 111 million belong to Adjective-Noun collocations group and last 35 million belong to Verb-Noun collocations group.

At the last stage, we searched for matches between potential semantically similar collocations, which were generated in the previous stage, and real text items of the corpus. Table II shows a quantitative distribution of semantically similar collocations, which were found in corpora, by the types of collocations.

TABLE II. SEMANTIC SIMILARITY COLLOCATIONS, FOUND IN CORPORA

Type of collocations	Unique collocations in the corpus (mln)	Generated collocations (mln)	Similar elements in the corpora (mln)
Adjective-Noun	1 215 016	111 934 244	7 072 639
Noun-Noun	4 271 435	197 900 280	8 384 758
Verb-Noun	539 087	35 452 948	563 468
Total	6 025 538	345 287 472	16 020 865

Generally, the metrics used in the evaluation of texts classification systems or information retrieval systems are precision and recall. In order to evaluate the results of our experiments, we use precision, which is denoted as the ratio of the number of correctly found semantically similar collocations to the total amount of similar collocations that are found in the result of the experiment. We could not evaluate the recall of the experiment results due to the size of our corpus as well as a number of unstructured texts generated by Social network data streams.

We compared the precision of our method of semantically similar collocations extraction with the results of three other existent methods for semantically similar

¹ http://www.nltk.org/_modules/nltk/stem/wordnet.html

² <http://chaoticcity.com/dependensee-a-dependency-parse-visualisation-tool/>

³ <http://www.nltk.org/>

⁴ <https://wordnet.princeton.edu/>

collocation extraction. They are Wu and Zhou's method [7], Pasca and Dienes's method [8], and Barzilay and McKeown's method [9]. Table III shows the comparative values of the average precision of our approach and three methods that have been described above.

TABLE III. COMPARATIVE VALUES OF THE AVERAGE PRECISION

Our approach			Method by Hua Wu, Ming Zhou	Method by M. Pasca, P. Dienes	Method by R. Barzilay, K. McKeown
<i>Adjective-Noun</i>	<i>Noun-Noun</i>	<i>Verb-Noun</i>			
0.934	0.961	0.972	0.739	0.457	0.716

It can be seen that the average values of precision of Methods 1-3 are much lower than that of our approach.

V. CONCLUSION AND FURTHER WORK

The semantic similarity model, built in this paper, and obtained lists of semantic similarity collocations can be used in computer processing to monitor social networks data streams. The model is based on logical-algebraic equations, which leverage a set of grammatical and semantic characteristics of words to determine semantic similarity of text fragments from public forums. To implement the model, we leverage (1) UD parser to gain the grammatical and semantic features of words, (2) NTLK with lexical database WordNet to obtain synonyms of words.

Experimental results indicate that our model extracts semantically similar collocations with average precision over than 92 %. As far as we can see the precision of our experimental result significantly outperforms precision of the existent methods. Consequently, the use of the model allows increasing the effectiveness of natural language processing tasks, including social network data streams analysis. In future studies we are planning to use our model to analyze data streams from other popular social sources, such as Wikipedia. Despite its popularity, this free encyclopedia is often criticized for pool quality of information. Proposed models can help to analyze semantic similarity of Wikipedia articles of various topics. Together with other metrics, a proposed approach can help to build models for a more comprehensive analysis of the text quality, which can be used to enrich less developed language versions of Wikipedia [16].

REFERENCES

- [1] M. Adedoyin-Olowe, M. M. Gaber, and S. Frederic, "A Survey of Data Mining Techniques for Social Media Analysis," *Journal of Data Mining & Digital Humanities*, 2014.
- [2] J. Golbeck, *Analyzing the Social Web*. M. Kaufmann, 2013.
- [3] J. Scott and P.J. Carrington, *The SAGE Handbook of Social Network Analysis*. SAGE Publications, 2011.
- [4] D. M. Boyd and N. B. Ellison, "Social Network Sites: Definition, History, and Scholarship," *Journal of Computer-Mediated Communication*, vol. 13 (1), pp. 210-230, 2007.
- [5] M. J. F. Rodrigues and A. J. S. Teixeira, *Advanced Applications of Natural Language Processing for Performing Information Extraction*. Springer, 2015.
- [6] A. Pesaranghader, A. Pesaranghader, S. Matwin, and M. Sokolova, "One Single Deep Bidirectional LSTM Network for Word Sense Disambiguation of Text Data," *Springer, Advances in Artificial Intelligence: 31st Canadian Conference on Artificial Intelligence, Canada*, pp. 96-107, 2018.
- [7] Hua Wu and Ming Zhou, "Synonymous Collocation Extraction Using Translation Information," *41st Annual Meeting on Association for Computational Linguistics (ACL'03)*, Stroudsburg, PA, USA, vol.1, pp. 120-127, 2003.
- [8] M. Pasca and P. Dienes, "Aligning Needles in a Haystack: Paraphrase Acquisition Across the Web," *Second International Joint Conference: Natural Language Processing (IJCNLP 2005)*, Korea, pp. 119-130, 2005.
- [9] R. Barzilay and Kathleen R. McKeown, "Extracting Paraphrases from a Parallel Corpus," *39th Annual Meeting on Association for Computational Linguistics (ACL '01)*, Stroudsburg, PA, USA, pp. 50-57, 2001.
- [10] S. Petrasova and N. Khairova, "Automatic Identification of Collocation Similarity," *10th International Scientific and Technical Conference: Computer Science & Information Technologies (CSIT'2015)*, Lviv, pp. 136-138, 2015.
- [11] S. Petrasova and N. Khairova, "Using a Technology for Identification of Semantically Connected Text Elements to Determine a Common Information Space," *Cybernetics and Systems Analysis*, Springer, vol. 53 (1), pp. 115-124, 2017.
- [12] M. Bondarenko and Yu. Shabanov-Kushnarenko, *The intelligence theory*. Kharkiv, SMIT, 2007. (in Russian).
- [13] P. Basile, A. Caputo, and G. Semeraro, "An Enhanced Lesk Word Sense Disambiguation Algorithm through a Distributional Semantic Model," *International Conference on Computational Linguistics*, pp. 1591-1600, 2014.
- [14] J. Schler, M. Koppel, S. Argamon, and J. Pennebaker, "Effects of Age and Gender on Blogging," *2006 AAAI Spring Symposium on Computational Approaches for Analyzing Weblogs*, pp. 191-197, 2006.
- [15] T. Pedersen, S. Patwardhan, and J. Michelizzi, "WordNet: Similarity - Measuring the Relatedness of Concepts," *Demonstration Papers at HLT-NAACL*, pp. 38-41, 2004.
- [16] W. Lewoniewski, "Enrichment of Information in Multilingual Wikipedia Based on Quality Analysis," *In International Conference on Business Information Systems*, Springer, Cham, pp. 216-227, 2017.

Big Data Real Time Ingestion and Machine Learning

Gautam Pal
Department of Computer Science
University of Liverpool
Liverpool, UK
gautam.pal@liverpool.ac.uk

Gangmin Li
Department of Computer Science and
Software Engineering
Xi'an Jiaotong-Liverpool University
Suzhou, China
Gangmin.Li@xjtlu.edu.cn

Katie Atkinson
Department of Computer Science
University of Liverpool
Liverpool, UK
K.M.Atkinson@liverpool.ac.uk

Abstract— Data arrives in all shapes and sizes. Many time data are acquired sequentially – as an infinite ever growing stream. This real time stream data needs to be processed sequentially by taking the data source and splitting it up along temporal boundaries into finite chunks or windows. Take examples from stock market, sensors or Twitter feed data. Rather waiting for data to be collected as a whole at a long periodic interval, streaming analysis let us identify patterns – and make decisions based on them – as data start arriving. When data are non-stationary, and patterns change over time, streaming analyses adapt. At scales, where storing raw data becomes impractical, streaming analysis let us persist only smaller, more targeted representations. This work describes machine learning approaches to analyze streams of data with an intuitive parameterization. Linear regression and K-means clustering concepts are redefined to the context of streaming.

Keywords— Real Time Data Analytics, Big Data, Real Time Data Ingestion, Real Time Machine Learning, K-means clustering

I. INTRODUCTION

Streaming is a data processing paradigm that is designed with infinite dataset in mind. Introduced as a new category of open source project-scalable stream processing, by Twitter's Nathan Marz, creator for Apache storm, a distributed real time computation framework [1].

Analyzing and predicting at a real time through a machine learning approach on a huge volume of heterogeneous data pool requires a novel distributed parallel computing approach-call it *Big Data Online Learning*.

Thus, the two-main focus area of this research are (a) the presentation of a live streaming data ingestion and processing mechanism through Flume, Kafka and Spark Streaming framework. We take a case study of capturing click-stream data to illustrate the real time data retrieval. (b) Provide an insight into real time Big Data machine learning models like streaming regression and streaming K-means clustering.

II. PRELIMINARIES

A. Processing Time Series Data

Considering how fast we want response should return, data processing paradigm is broadly categorized into following three types:

1) *Batch Processing*: This is a high-latency/high-throughput option. Processing time ranges from few minutes to hours. Data is loaded in huge batches at a certain

intervals, reports are generated, users look at the same reports until the next data load occurs. Tools and solutions like Hadoop Map Reduce, Spark core, Spark (SQL, MLLib and GraphX) are the popular choice for Big Data analytical processing using batch mode. Schedulers for batch include Apache Oozie, Spring Batch and Unix Corn.

2) *Request response*: This is known as Online Transaction Processing (OLTP). This is the lowest latency paradigm, with response times ranging from submilliseconds to a few milliseconds. Point-of-sale systems, credit card processing, and time-tracking systems typically work in this paradigm.

3) *Stream Processing*: This is a continuous nonblocking option which is for low latency messaging and event process responding to user request real time or near real time. Most operations on streams are windowed operations—operating on slices of time: moving averages for stock process every hour, top products sold this week, etc. Popular choice for stream processing tools include Apache Kafka, Apache Flume, Apache Storm, Spark Streaming, Apache Flink, Amazon Kinesis etc [2].

B. Data Processing Patterns

Following are the two data processing patterns are observed in a time series data:

1) *Bounded Data*: It's a finite set of data, possibly full of entropy, runs through data processing engine such as Map Reduce to transform gradually into more structured dataset. Typically batch jobs are run as bounded data patterns at a periodic intervals.

2) *Unbounded Data*: It represents an infinite ever growing data stream which has no beginning and end. Real time stream processing is the framework for processing this unbounded type.

C. Windowing

Windowing is a notion of taking a data source and splitting it up along temporal boundaries into finite chunks for processing. Three types of window operations possible [4].

1) *Fixed Window*: Fixed windows splits up time into chunks with a fixed-size temporal length and aggregates

This work is supported by Xi'an Jiaotong-Liverpool University
(Ref: RDF 15 - 02 - 35)

them. Fixed window doesn't overlap with each other's time boundaries. Example: Count the blue elements in the stream every 1 min intervals. See Fig. 1.

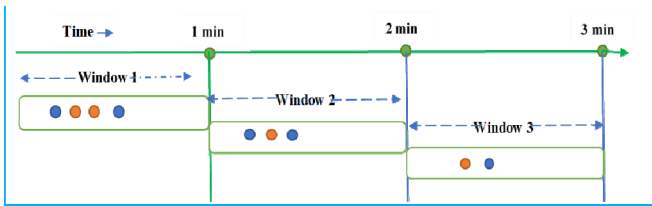


Fig. 1. Fixed window

2) *Sliding Window*: Sliding window is made of a fixed length and fixed period. If period is less than length then windows overlap. Example: Every 30 sec find the number of blue elements over the last 1 min. See Fig. 2.

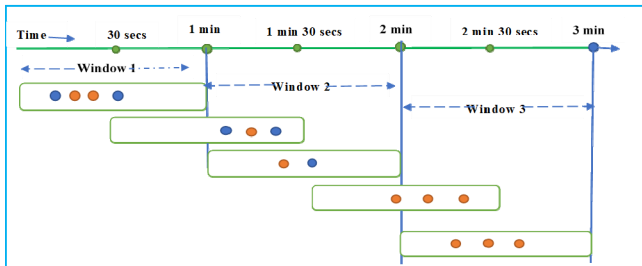


Fig. 2. Sliding window

3) *Sessions*: Windowing over sessions groups together number of related events over distinct sessions. Example: Count the blue elements in each unique sessions. See Fig. 3.

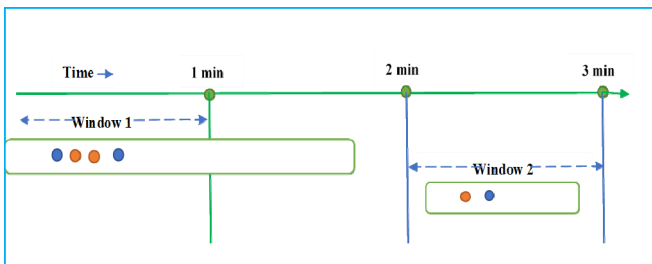


Fig. 3. Window based on sessions

D. Watermarking

In the event time windowing we take the finite set of data based on the time they occurred. Most of the older version of stream processing systems did not consider the event generation time. If the event arrives late, application just considers the processing time. Accepting processing time as event generation time can lead to major errors in the insights. To allow the processing time to be considered for the late arriving events, the engine should maintain state and allow late data to update the window state until a threshold time. This threshold is defined by *watermarks* to manage how long we wait for the late events. See [3] for deep dive into the concepts of watermarking.

E. Spark D-Streams

DStreams or Discretized stream are Apache Spark stream processing abstraction in which computations are structured as a series of stateless, deterministic batch computations at a small amount of time intervals [4].

DStreams is the abstraction on top of Spark Resilient Distributed Dataset (RDD). Series of RDDs (of same type) are put together as one single DStream which is processed and computed at a user-defined time interval. DStreams are created from data sources like file systems, sockets etc. or from other DStreams as well.

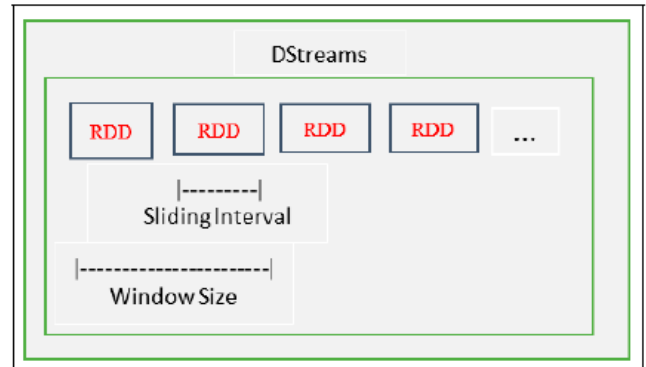


Fig. 4. DStreams as a collection of Spark RDDs

III. STREAMING DATA PIPELINE AND PROCESSING

Data ingestion pipeline is as important as the data itself. Therefore, the need arises for frameworks capable of delivering streams of events in a reliable distributed manner. Systems should be scalable, load distributing across cluster nodes. Ingestion framework persists incoming data into memory and disk to get consumed at a later point of time – much like a producer consumer system. For this work, we have experimented on two popular Big Data ingestion frameworks – Kafka and Flume. These tools acts as a message bus in the integration pipeline without transforming anything on the data. Spark is taken as a stream processing framework and Cassandra, a columnar NoSQL database, acts as a data storage system. Each of the tools are briefly described below:

1) *Apache Kafka*: Kafka is a popular choice for real-time data retrieval. It is capable of Ingesting high velocity large volume of data which requires fast, fault tolerant, distributed pipelines. Kafka being massively distributed client-server-oriented publisher-subscriber messaging system replaces the traditional message queue systems like Rabbit MQ, IBM MQ because of its higher throughput, reliability and replication capability [5]. It acts as a central hub for real-time processing when using along with Spark stream processing APIs [4].

2) *Apache Spark and Apache Storm*: Spark and Storm are the two popular distributed stream processing computation framework. See [6][1] to learn more about Spark and Storm real time frameworks.

3) *Apache Cassandra*: Cassandra is a columnar NoSQL database for storage amounts of data across many commodity servers, providing high availability with no single point of failure. Refer Datastax documentation [7] for more about Cassandra.

4) *Flume*: Is an agent based framework which enables information gathering from multiple sources and integrate them to collate in an enterprise data lake. Flume is a high

available system for collecting, aggregating large amount of data from different sources to a centralized datastore [8]. Flume agents are JVM processes that host the components through which events flow from external sources to next destination(hop). Each agent comprises of three components: (i) source (origin of data) (ii) sink (final storage for the

events) and (iii) channel (passive store that keeps the events until it is consumed) [5].

Refer Fig. 5 for set of Big Data tools used for architecting four layered Big Data architecture.

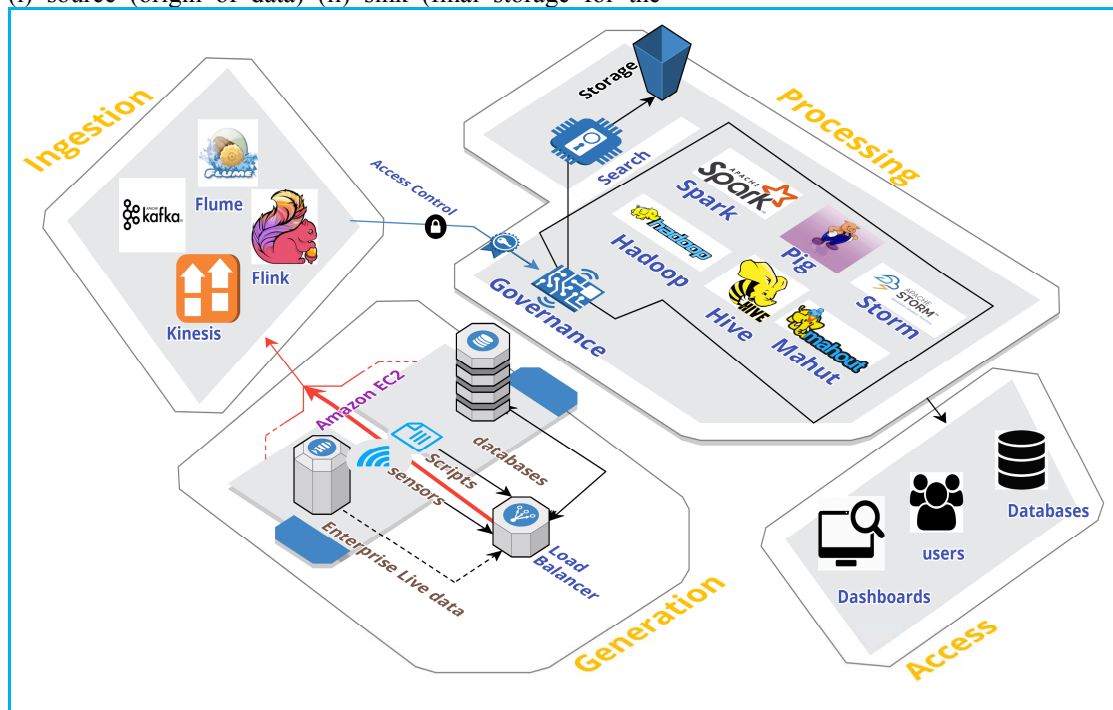


Fig. 5. Four layered Big Data architecture.

IV. EXPERIMENTS

This section describes the experiment with a use case for live click-stream data using Kafka and Flume and provides a comparative study between them.

A. Real Time Clickstream Data Ingestion

A clickstream is the recording of parts of the screen a user clicks on while web browsing [9]. The user action is captured in client-side browser. Therefore, clickstream data is the URL generated from each user click data. For example, in online shopping portal a clickstream data may look like this:

<http://xxx.com:8080/electronics/products?userID=id3&productName=iPhone8&price=500&location=london>

Each time user opens a new session, a new user context is captured. The context is basically a uniquely derived object created from session object created at JavaScript layer. ContextID would then be appended to each user's click data.

B. Capturing Clickstream Data

There are two distinct approaches for capturing live click-stream data.

Approach 1: Simplest solution to the problem is to collect the web browser log files and then push them into HDFS.

Disadvantage: No real-time processing. The approach is inherently batch oriented. As the Hadoop/Spark batch processing would require data to be picked up from HDFS/NoSQL for further processing this would add delays to the response.

As the raw clickstream events are unstructured, direct insertion will also require further parsing, cleansing and re-inserting into datastore.

Approach 2: This approach doesn't rely on server-side logs to be processed in batches. Rather, an event is generated on the client side and delivered to the separate back-end service to handle the event processing and logging while still in transit without the need for storing.

Advantages: Real time processing. Allows custom events fired from JavaScript to be processed in-flight before ingesting into HDFS. This makes processing response real time or near real time. We adapt approach 2 for our ingestion frameworks.

C. Retrieval of Clickstream Data using Apache Kafka

We used a simulation which generates high rate of simulated click data and subsequent writes into Kafka topic. Spark acts as a consumer to the Kafka and processes the stream and persists the results into Cassandra. The end-to-end flow is depicted in the Fig. 6:

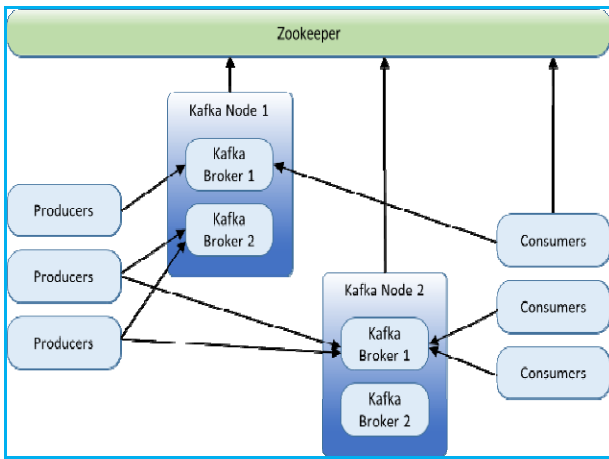


Fig. 6. Multi-node, multi-broker Kafka cluster

1) Test Bed

System configuration for each node in Amazon Cloud Platform is uses basic setup as follows:

2 vCPU, 6.5 GB RAM, Centos 7 OS

We started with 1 node cluster and went up to 3 nodes cluster size with Kafka, Spark and Cassandra installed in each of the node. Multi-node, multi-broker cluster setup is shown in the Fig. 5.

2) *Execution*: First, Create a Kafka messaging queue capable of handling streaming text(clickstream) using String serializer/deserializer. Spark acts as a consumer to Kafka queue and processes data (counting the number of events) at a window interval of 1 minute. Once the streaming job is started Spark will process and continually ingest data into Cassandra DB every 1 minute at a pre-created keyspace and column family. See Algorithm 1.

D. Retrieval of Clickstream Data using Apache Flume

Moving data on a real-time basis over the network can cause the network being loaded all time. If millisecond range response is not required, we can choose to go for batch processing mode using tools like *Apache* Flume. Also, moving small amount of real-time data can be particularly a problem for HDFS storage since Hadoop is designed for large files. Therefore, for sources with lots of small files, Flume agents could collect them and flush data in a batch mode as a large single file.

Algorithm 1: Live Data Ingestion and processing using Kafka and Spark

Input: Stream data, Kafka broker address, Cassandra server address

Output: Processed data saved into Cassandra datastore

1. **Begin**
 2. Create a Kafka stream with broker configuration
 3. Subscribe Kafka from Spark Stream
 4. Create a stream pipeline from Spark to Cassandra
 5. Spark consumes from Kafka, processes and commits to Cassandra
 6. **End**
-

Configuration: Config File for live data ingestion using Flume

```

1. agent1.sources = source1
2. agent1.sinks = sink1
3. agent1.channels = channel1
4. # Describe/config source1
5. agent1.sources[source1].type = netcat
6. agent1.sources[source1].bind = hostname.com
7. agent1.sources[source1].port = 7000
8. agent1.sources[source1].eventSize = 1
9. # Describe sink1
10. agent1.sinks[sink1].type = hdfs
11. agent1.sinks[sink1].hdfs.path =
    /var/clickdata
12. agent1.sinks[sink1].channel = memory-channel
13. agent1.sinks[sink1].hdfs.writeFormat=Text
14. agent1.sinks[sink1].hdfs.fileType =
    DataStream
15. # Use a channel which buffers events in
    memory
16. agent1.channels[channel1].type = memory
17. agent1.channels[channel1].capacity = 1000
18. agent1.channels[channel1].transactionCapacity
    = 100
19. # Bind the source and sink to the channel
20. agent1.sources[source1].channels = channel1
    agent1.sinks[sink1].channel = channel1

```

1) Flume Spool Directory

We create a data source as spool directory which lets to insert small files as they appear. Flume source keeps tracking this directory and fetch all files in spool directory after a configurable batch size reached. It combines all the incoming files into one large size file and move to the destination (Flume sink).

2) Test Bed

System configuration for each node in Amazon Cloud Platform is as follows:

2 vCPU, 6.5 GB RAM, Centos 7 OS

3) Execution

Single node Flume installation was used with two nodes Hadoop and Hive installation. Flume used 5 sources with each event size of 400 kb.

Following Flume configuration file was used:

2) Result

3) Remarks

Cassandra consistency level is kept to ONE which is lowest level of consistency setting to achieve maximum write performance possible. The overall throughput scales up linearly extremely well with added number of nodes in the cluster (Fig. 7).

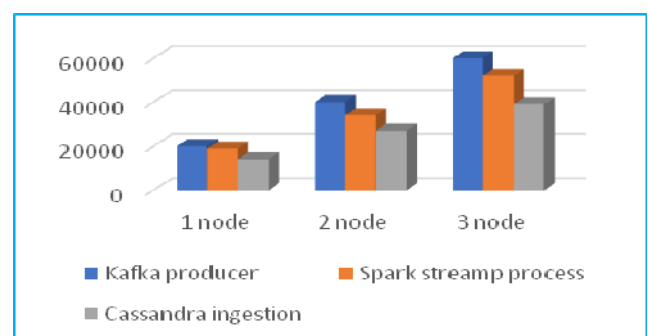


Fig. 7. Write Throughput vs Cluster Size

Line 6-7: Flume agent creates a server socket and continually listening at the specified host and port.

Line 11: Final storage location into HDFS

Line 17: Specifies capacity for maximum number of event storage.

4) Results

As depicted in the Fig. 8, sink throughput largely depends on sink type (hdfs, hive etc.) and level of parallelism (number of sources, sinks).

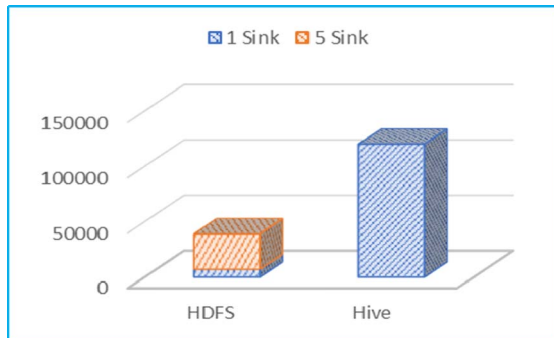


Fig. 8. Flume throughput on different sinks

5) Remarks

Since flume doesn't support distributed client server architecture a single Flume agent can't be distributed across multiple nodes. However multiple agents can be created to load distribute the overall input stream data. Fig. 8 shows the experimental results for a single agent flume client. Write throughput largely varies over the destination sink we choose from.

E. Comparing Kafka and Flume

While Kafka and Flume can achieve the same objective, there are few key differences observed:

Firstly, Kafka can take only pull based approach i.e. it can accumulate data until client initiate read request and client provides the Kafka address and pulls data from it. Flume on the other hand, can perform both pull and push based approach as shown in the following scala code snippet:

```
val stream = FlumeUtils.createStream(ssc,
"hostname.com", 7788) //Push based

val stream = FlumeUtils.createPollingStream(ssc,
"hostname.com", 7788) //Pull based.
```

From developer's point of view Flume would be easier to implement being a configuration based contrary to Kafka's programming-based integration.

Another key advantage observed with Flume is, some level of data transformation or processing is possible through Flume interceptors. While Kafka needs to integrate with other data processing framework like Apache Storm or Apache Spark.

F. Using both Kafka and Flume

Latest trend in data ingestion paradigm is to use both Kafka and Flume together. Cloudera distribution of Hadoop

(CDH) Flume latest distributions (version 5 onwards) accepts data from Kafka via Kafka Sink [10].

V. STREAMING BIG DATA MACHINE LEARNING

Batch machine learning is applied on a fixed set of data. Typically, these techniques are also iterative, and we perform multiple passes over training data to converge to an optimal solution [11].

In contrary, online learning predicts on each passing window of time frame. In an incremental way the model continually updates as new information is received.

However, online learning model can be used along with batch setting. Like we can use stochastic gradient descent (SGD) optimization to train classification and regression model after each training example. However, we still make use of multiple passes over the training data in order to converge into a better result[11].

A. Streaming Regression

Defined in the following two phases:

1) *Training*: Takes the labeled data points. Model gets trained on every batch of the input stream. It can be called repeated time to train on different stream.

2) *Predict*: It also take labeled data points and tells the model to make prediction on the input stream.

On each passing window, model variable gets updated and exposes the latest trained model. This enables user to use the model in other applications or save at an external location. Similar to batch execution, streaming model can be conFig.d with step size and number of iterations. At the start of the training, initial weight vectors are set to zero vector or as a random vector [10].

In the following section we discuss a key stream data learning technique like streaming regression model and streaming k-means clustering model. We start with pseudocode for generating data for the learning models.

Algorithm 2: Stream Data Producer

Input: events/second, number of features
Output: Stream of training data

7. **Begin**
8. maxEvents=n1
9. numberOfFeatures=n2
10. Generate weight vectors(w) attached to each n2 features
11. Run infinitely
12. For max events
13. Stream data=feature vector × weight vector
14. Write to socket
15. End for
16. **End**

A streaming regression model get trained and predicts on top of data produced by the above stream producer.

Algorithm 3: Streaming Regression Model

Input: Events. number of features
Output: Predict values

1. **Begin**
2. numberOfFeatures=same as producer (n2)
3. Generate stream of labeled points
4. Train model on the stream
5. Make predictions
6. **End**

Algorithm 2 and 3 is presented together in the Fig. 9:

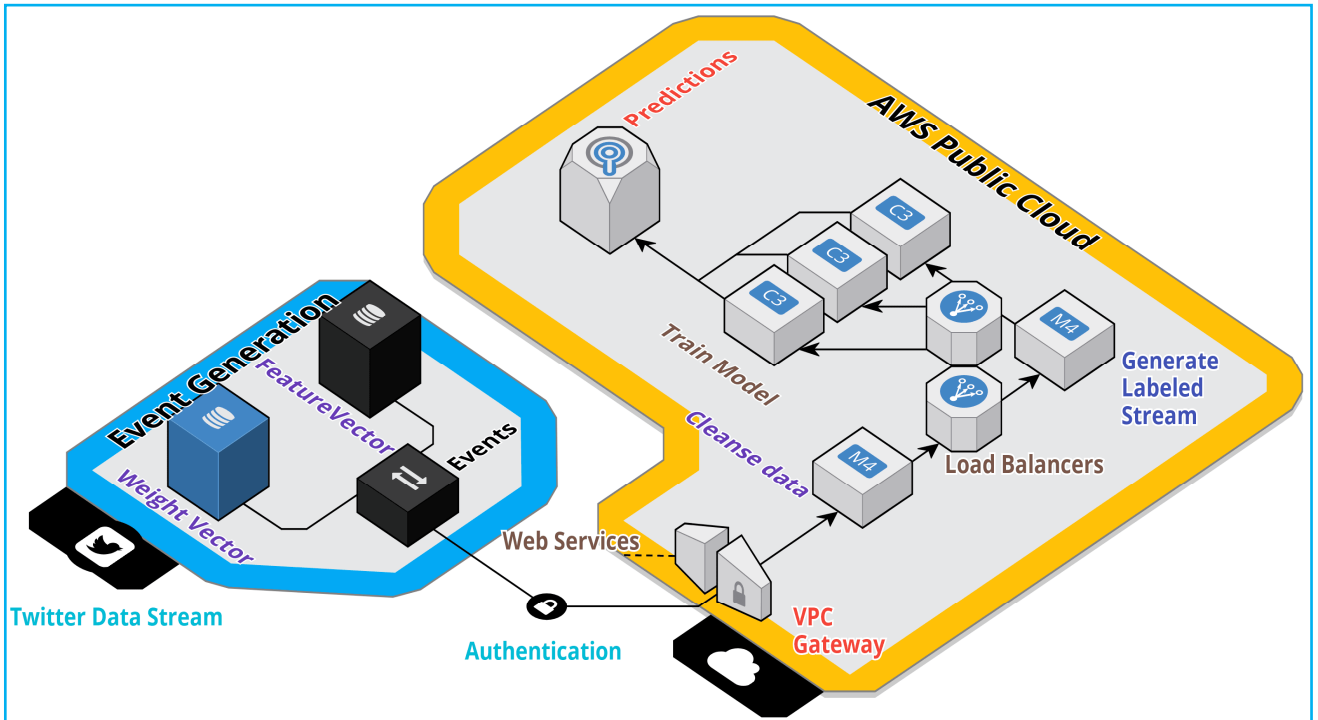


Fig. 9. Stream processing and lambda architecture

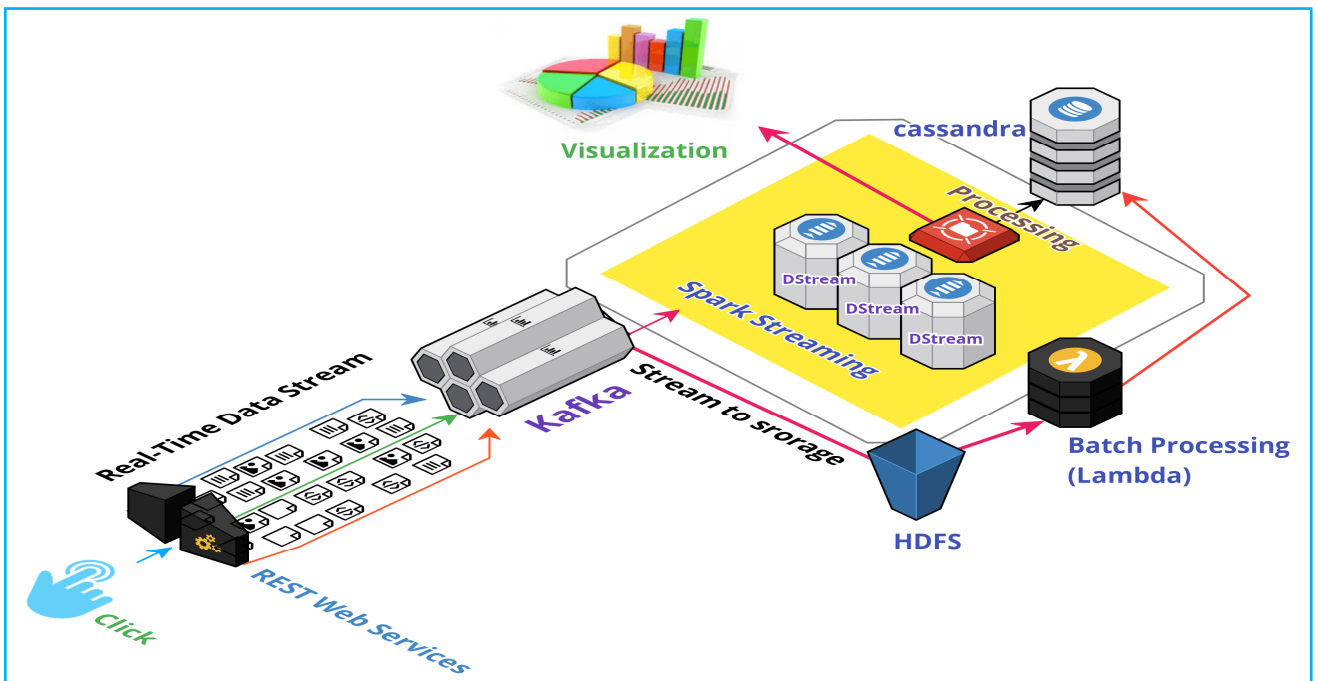


Fig. 10. Event generation and learning

B. Streaming K-Means Clustering

This concept is taken from latest release of Apache Spark release [12]. In streaming K means clustering, model is updated with each passing window used on a combination between cluster centers computed from the previous batches and the current batch. Algorithm starts with assigning data points to their nearest cluster. For each new iteration, when

new data comes, compute new cluster centers, then update each cluster using:

$$c_{t+1} = \frac{c_t n_t \alpha + x_t m_t}{n_t \alpha + m_t}$$

$$n_{t+1} = n_t + m_t$$

Where n_t is the old data points and c_t is the old cluster center. m_t is the new data points and x_t is the new cluster center. α is the decay factor. If $\alpha = 1$ all data will be used from the beginning; with $\alpha = 0$ only the most recent data will be used. Spark MLlib includes a streaming version K-Means clustering called streaming K-Means.

VI. CONCLUSION AND FUTURE WORK

This work sought to put lights on scalable stream processing framework and its three focus areas – data ingestion, processing and learning. With respect to ingestion, Apache Kafka and Flume are the two popular choices for high velocity Big Data ingestion due to their horizontal scalability and robust failover. In the data processing layer, Spark Streaming and Storm are the two efficient Stream processing frameworks. Spark streaming has number of programming language support (Scala, Java, Python, R, SQL) and does in-memory processing. Finally, we explored machine learning in the streaming context with respect to streaming regression and streaming K-means clustering approaches. In future, this work can extend to study Lambda architecture [13] to combine the both batch and stream processing approaches. As depicted in the Fig. 10, all data ingested to the system is dispatched to both the batch layer and the stream layer for processing. Stream layer serves only low-latency queries. Data gets merged for other type of query which requires historical data.

REFERENCES

- [1] D. Xiang, Y. Wu, P. Shang, J. Jiang, J. Wu, and K. Yu, "RB-storm: Resource Balance Scheduling in Apache Storm," in 2017 6th IIAI International Congress on Advanced Applied Informatics (IIAI-AAI), pp. 419-423, 2017.
- [2] T. L. Akidau, Slava Chernyak, and Reuven Lax, "chapter 1," in Streaming Systems: o'reilly media, 2017
- [3] T. L. Akidau, Slava Chernyak, and Reuven Lax, "Chapter 3. Watermarks," in Streaming Systems, 2017
- [4] R. Dua, M. Singh Ghotra, and N. Pentreath, "An introduction to Spark Streaming," in Machine Learning with Spark, 2017.
- [5] Apache Kafka. Available: <https://kafka.apache.org/>
- [6] "Spark Streaming."
- [7] Datastax. Apache Cassandra Available: <https://docs.datastax.com/en/cassandra/3.0/>
- [8] A Apache Flume. Available: <https://flume.apache.org/>
- [9] R. Hanamanthrao and S. Thejaswini, "Real-time clickstream data analytics and visualization," in 2017 2nd IEEE International Conference on Recent Trends in Electronics, Information & Communication Technology (RTEICT), pp. 2139-2144, 2017.
- [10] Flume or Kafka for Real-Time Event Processing. Available: <https://www.linkedin.com/pulse/flume-kafka-real-time-event-processing-lan-jiang/>
- [11] R. Dua, M. Singh Ghotra, and Nick Pentreath, "Online learning with Spark Streaming," in Machine Learning with Spark, 2nd ed., 2017.
- [12] Spark Streaming K-means. Available: <https://spark.apache.org/docs/latest/mllib-clustering.html>
- [13] J. Heidrich, A. Trendowicz, and C. Ebert, "Exploiting Big Data's Benefits," IEEE Software, vol. 33, no. 4, pp. 111-116, 2016.

A Method to Solve Uncertainty Problem for Big Data Sources

Andrii Berko
Information Systems and Networks Department
Lviv Polytechnic National University
Lviv, Ukraine
andrii.y.berko@lpnu.ua

Vladyslav Alieksieiev
Applied Mathematics Department
Lviv Polytechnic National University
Lviv, Ukraine
vladyslav.i.aliexsieiev@lpnu.ua

Abstract— Big Data analysis and processing is a popular tool for Artificial Intelligence and Data Science based solutions in various directions of human activity. It is of a great importance to ensure a reliability and a value of data source. One of the key problems is the inevitable existence of uncertainty in stored or missing values. Any uncertainty in a source causes its disadvantageous, complexity or inapplicability to use. That is why it is crucial to eliminate uncertainty or to lower uncertainty influence. Here in this research, we offer ontology-based method to solve an uncertainty problem for big data sources.

Keywords— big data; data sources; data uncertainty; ontology; uncertainty elimination

I. INTRODUCTION

Nowadays there are many areas requiring to solve problems with artificial intelligence solutions and tools supplemented with the necessity to use big data sources. It concerns many tasks in business, finance, medicine, politics, ecology and ecological surveillance, and many other all requiring artificial intelligence.

These tasks need to take into consideration such features of big data like volume, velocity, and variety. Meanwhile, significance and reliability of the data should be kept. Special preparations should be made with big data sources before use. Those can be ETL (extract, transform, load) processes, normalization, aggregation etc. This is the step of data source processing when the problem of uncertainty appears.

An anomaly appears in some missing values, incomplete data, inaccuracy, inconsistency, unreliability etc. Generally, this lowers the value of the big data source, reliability of final results, makes it difficult or impossible to work with the resource. These are the reasons to consider the importance of the problem of elimination or lowering the influence of uncertainty level in big data sources.

The problem of data uncertainty has been discussed for a long time. Solutions for the problem and its different aspects were offered in [1–5]. Also, there were researches on some particular cases of custom IoT based monitoring system like a problem of data losses [6] and a problem of aggregation of obsolete data [7].

Difficulties of solving the uncertainty problem for big data sources are explained with its features: huge volume, high level of velocity and variety. Due to these features, all standard tools are not applicable. And this is the motivation to develop some new approaches oriented to interact with sources of big data.

II. PREREQUISITES OF BIG DATA UNCERTAINTY PROBLEM

A. Why Big Data?

The primary question is to understand the peculiarity of uncertainty problem in big data. First, the answer comes from its key features. Basic characteristics differing big data from other sources types are so named “triple V” – Volume, Variety, and Velocity. These are the specific features responsible for the appearance of effect and problem of uncertainty in big data sources. Now, let’s discuss the influence of these features.

1st. Big and huge volumes of incoming source require its distribution. In the meantime, different parts can be managed with different tools. This kind of architecture does not allow to maintain the global integrity of the data. Inability to ensure source integrity causes inevitably to the appearance of inconsistency in data, data losses and data distortion. As a result, one gains the uncertainty of some part of a big data source.

2nd. The variety of big data requires using different schemas, descriptors, and another mechanism to describe data within the global resource. Thereafter, this causes to appear an inconsistency, duplication, incompleteness, ambiguity, and different interpretation of data units. In an example, in one source the data may exist, and in another source, similar by meaning, it may be absent. Another inconsistency is to have the same presentation for different data or different presentation for the same data in different parts of the global source. Again there is a reason for an effect of uncertainty within a global big data source.

3rd. The velocity of big data combined with no control of integrity makes any changes asynchronous and inconsistent. There may also happen, that some values of data endure some changes but a corresponding or similar values remain unchanged. Besides, fast and unsynchronized changes make it appear incorrect and inconsistent values, or unpredictable loss of relevance of some data, etc. This is one more reason for uncertainty to appear in big data sources.

Thus, the conclusion can be made, that the effect of uncertainty is natural for big data. Uncertainty seems to exist almost always in big data sources, due to the basic features – Volume, Variety, and Velocity.

There are some more factors causing the uncertainty in big data sources. These are the requirements for the resource known as “another double V”, what means Value and Veracity [5]. Value means the cost and applicability of data

to solve some definite set of problems. Veracity is the relevance, accuracy, and reliability of the data. It is obvious, that the high-level uncertainty makes it is impossible to fulfill the requirements. Some traditional data sources (Databases, Data Warehouses, formatted documents, XML etc.) to fulfill the requirements due to a low level of uncertainty, and some other data sources (private sources, texts, pictures, social networks) has no such requirements to be followed strictly. At the same time, those big data sources the problem of uncertainty needs a special way for solution.

Usage peculiarities of big data can be defined as a third aspect, establishing the specifics of uncertainty problem. While the traditional data sources can be considered available and ready to use, the big data sources have a need to implement some preparation procedures. This preparation procedure is typically made as ETL (extract–transform–loading), normalization, aggregation etc. Nevertheless, in all cases, a data cleaning should be made to prepare big data to use. And one of the most important steps of data cleaning is to eliminate the uncertainty or to lower the uncertainty level to reach some appropriate level.

B. Types of data uncertainties

C. J. Date [1] asserts not all the uncertainties to be the same. It can be divided by the origin factors, origin nature, and abilities of interpretation. Really, there is a difference between those cases, when the value does not exist due to impossibility, when the value exists but remains unknown, and when the value exists and known, but it is inconsistent or ambiguous. According to C. J. Date concept [1], there are the following types of uncertainties:

- **inexistent** value for the data element,
- value **not formed** yet at the moment,
- value exists, but **unknown**,
- value exists, but not received (**obtained**),
- value is invalid (**unacceptable**),
- value not **determined**,
- value is inconsistent (**corrupted**),
- value is **ambiguous**,
- value is not accurate enough,
- value is an empty, etc.

The list is incomplete and there more items could be added. Depending on contents and peculiarities of the source of big data, there can appear some different causes for data uncertainty. Why should we categorize data uncertainty in some resource? First, the way of uncertainty elimination depends on the nature and factors of origin of the uncertainty. For example, if the value exists, but remains unknown, then it can be queried again; if the value is inaccurate, then it can be refined; if the value is not formed, then it can be received later; etc. Second, when uncertainty elimination is impossible, then the process of the source processing could be planned accordingly; if the value is inappropriate, or invalid, or empty, then it can be excluded from processing; if the value is inaccurate or inconsistent, then the level of accuracy or confidence can be changed, etc. That is why the

categorization of uncertainties is recognized as a key element to solve the uncertainty problem for big data source.

C. Approaches to data uncertainty problem solution

There are a number of approaches to solve the problem of uncertainty in big data sources. The primary target is to eliminate uncertainty or to lower uncertainty level to support effective resource usage. The most popular approaches to uncertainties elimination are:

- **repeat a request** to receive a value,
- refine inconsistent or inaccurate values (**rectification**),
- eliminate the origin of inconsistency, repeat or ambiguity of values,
- **replace** uncertainty with some aggregate value (average value, probable value, standard or default value, initial value, some calculated value, estimated value, expert value, etc.),
- use of **fictitious** value as artificial surrogate marks instead of uncertainty,
- **remove** of uncertain value or data element from the resource,
- **ignore** the data uncertainties while processing the resource,
- create **special tools** to process uncertainties.

Besides those mentioned, there are some other approaches to eliminate uncertainties that can be used according to peculiarities of a big data source. The choice of appropriate approach is rather a difficult task. The appropriate approach should match the following conditions:

- 1st. Match the type of uncertainty.
- 2nd. Match the contents and peculiarities of big data source.
- 3rd. Ensure a correct result.
- 4th. Determine the best approach to reach the aim.

The solution of a problem of choosing the appropriate approach to eliminate uncertainties is one of the important steps to prepare a big data source for the use.

III. ONTOLOGY-BASED SOLUTION OF DATA UNCERTAINTY PROBLEM

A. Data uncertainty and ontology

As it was shown above, there is a direct tie between the approach to eliminate data uncertainty and the nature of the uncertainty. Meanwhile, the question comes to define the best approach to match the uncertainty type. The answer is ambiguous. Using an expert approach there were defined some cases of conformity. The cases are presented in Table I.

Obviously, to use an appropriate approach to eliminate data uncertainty it can be not enough only to define the nature of uncertainty.

Generally, the process of elimination of data uncertainty can be described as a formation of a new value for data element to replace uncertain value. Thus, the new value becomes explicit, definite, exact, unambiguous (consistent) and acceptable.

TABLE I. THE RELATION BETWEEN UNCERTAINTY TYPES AND METHODS OF ITS ELIMINATION

Uncertainty Type \ Method of Elimination	Not exists	Not formed	Unknown	Not obtained	Unacceptable	Not determined	Corrupted	Ambiguous
Repeated request	+	+	+	+	-	+	+	+
Rectification	-	-	-	-	-	-	+	+
Replacement	-	-	+	-	-	+	+	+
Factitious value	-	-	+	-	+	+	-	-
Remove	-	-	-	-	+	-	-	+
Ignore	+	-	-	-	-	+	-	-
Special tools	+	+	+	-	+	-	-	-
No Action	+	-	+	-	-	+	+	-

The value v_{ij} of some data unit V_i , which was formed to eliminate uncertainty, depends on uncertainty category U_k and elimination approach S_l . The model of new value formation to be described as consequent transitions: “data unit – uncertainty – elimination – new value”, or as

$$V_i \rightarrow U_k \rightarrow S_l \rightarrow v_{ij} \quad (1)$$

or as a function

$$v_{ij} = \Phi(V_i, U_k, S_l) \quad (2)$$

Basic assignment of the model is to answer the question: which method of uncertainty elimination, for which data unit, of what uncertainty type, when and how should be implemented. One solution of the problem is to implement a special knowledge base within cleaning tools for big data. The knowledge base should include some expert and synthetic knowledges like:

- structure and contents of a big data source;
- types of uncertainties in a big data source;
- approaches to eliminate uncertainties in big data source;
- correspondence between data units, uncertainties types, and elimination approaches.

The key tasks for that knowledge base are 1st – to accumulate expert knowledges about approaches for elimination of different types of uncertainties for particular data units of the big data source; 2nd – to develop and to supplement knowledge base with new knowledges; 3rd – to use knowledges for the purpose of data uncertainty elimination. The use of knowledge base allows to exclude or to reduce the influence of human factor and to increase the quality of preparation results for big data source.

The basis of knowledge base is the special kind of ontology. Generally, the ontology can be defined as

$$O^o = \langle C^o, R^o, F^o \rangle, \quad (3)$$

where $C^o = \{C^V, C^U, C^S\}$ is the set of concepts (classes), which are

C^V – data units of the resource,

C^U – types of uncertainties of the resource,

C^S – approaches to eliminate data uncertainties.

$R^o = \{R^{VU}, R^{US}, R^{VS}\}$ is the set of relations between ontology classes, respectively,

R^{VU} – relation between data units and uncertainties types,

R^{US} – relation between uncertainties types and approaches for uncertainties elimination,

R^{VS} – relation between data units and approaches for uncertainties elimination.

F^o – the set of rules (axioms) of data uncertainties elimination. Each rule defines the approach to eliminate particular uncertainty type within a particular data unit. Unlike to those conformities from Table I, the result of such rule implementation should be definitely unambiguous.

The common structure of the ontology for elimination of uncertainty in big data sources is described at Figure 1.

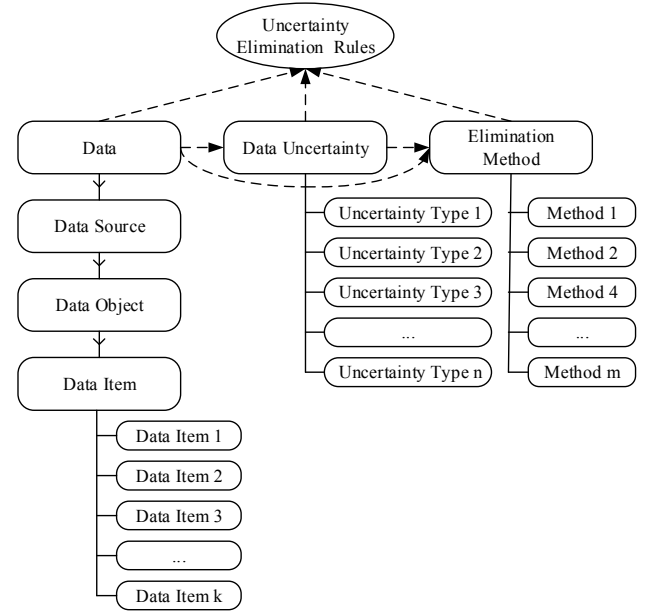


Fig. 1. Common ontology structure for Data uncertainty elimination in Big data Sources

The given ontology has the following concepts (classes) defined:

- The class “Data” describes the common content of big data source. Class is hierarchical and includes subclasses. 1) “Data Source” – local resource, like database, document repository, web-resource, content storage etc. 2) “Data Object” – sub source of a local resource, like a table, file, document,

container. 3) “Data Item” – class object, the elementary data unit having its own interpretation, like a column, field, XML-element, JSON-element and so on.

- Class “Data Uncertainty” describes data uncertainties of big data source. These uncertainties to be defined as it was shown above. The content of these uncertainties to be specific for each resource.
- Class “Elimination Method” consists of set of methods to be implemented for elimination of data uncertainty at a particular resource.

Relations defined in the ontology describe: 1) between data items and data uncertainty types; 2) between data uncertainty and related methods of its elimination; 3) between data items and possible data uncertainty elimination methods.

The defined ontology classes are connected to a set of rules to use methods of uncertainty elimination for particular units of big data source. Each rule can be a production rule of the form “if-then-else”. It defines the approach of elimination of uncertainty in big data source following the principle

FOR <Data Item i> **IF** <Data Uncertainty>
IS <Uncertainty Type j> **THEN**
 < Elimination Method k> **WOULD BE USED.**

This is the way the ontology allows to answer definitely to an earlier posed question: which method of uncertainty elimination to use with what kind of data while preparing some resource of big data.

B. The algorithm of ontology-based data uncertainty problem solution for big data sources

With the ontology for big data source to eliminate uncertainties a specialized algorithm is offered. The algorithm consists of three stages. The stages are to describe the solutions for the types of tasks respectively: 1) elimination (lower the level) of uncertainty of big data source; 2) analysis and identification of problem situations that appeared during the data uncertainties elimination process; 3) learning and improvement of ontology. For the purpose of best effectiveness stages 1 and 2 to be done first. Stage 3 can be executed after the elimination of uncertainties in big data source and after fixing problem situations. Now there is a description of the algorithm.

/* Part 1 – Data Uncertainty Elimination */

Step 1. Research on resource (sub-resource, unit) of data. The step is to reveal data uncertainties. In case of success (uncertainty found) Step 2 should be made, otherwise, (no uncertainties found) current step to be repeated for the next data unit of the resource (sub-resource). When research is finished for all the data units, then move to Step 8.

Step 2. Uncertainty categorization. The uncertainty to be appraised to find out its type according to ontology defined types. If the type is found, then Step 3 should be made, otherwise – Step 6.

Step 3. Search for an approach to solve uncertainty. This step is to find out whether the ontology has the definition of the required approach to eliminate the particular uncertainty type. If there is an appropriate approach defined, then Step 4 should be made, otherwise move to Step 6.

Step 4. Making a decision to eliminate uncertainty. Among the set of rules, there should be found a rule to conform data unit and uncertainty type and its elimination approach. If the rule was found, then Step 5 should be made, otherwise (no rule found in the ontology) move to Step 6.

Step 5. Execution of uncertainty elimination according to the rule. This step is processed with the call to some specific procedure, which was previously defined. If the uncertainty is eliminated, then the changes to the data are made. After that move to Step 1 for the next data unit. If the uncertainty was not eliminated, then move to Step 6.

/* Part 2 – Problem Analysis */

Step 6. Processing the problem situation. This step is to recognize and fix the problem situation, appeared during the attempt of data uncertainty processing. These problem situations can be categorized according to its origin:

- no description of data unit (sub-resource) in the ontology;
- no description of uncertainty type for the data unit in the ontology;
- no description of approach to eliminate some type of uncertainty or in some type of data unit;
- no description of rule to eliminate uncertainty for particular data unit;
- implementation of rule to eliminate uncertainty did not effect.

Step 7. Fix the problem situation. If the situation is recognized and categorized it can be fixed in a special log (or register) using some particular format. After fixing the problem situation move back to Step 1.

/* Part 3 – Ontology Learning and Evolution */

Step 8. If the result of Steps 1–7 there is no fixed problems, or there are no unprocessed records in the log (register), then the ontology-based algorithm is finished. If there are records in the log, then move to Step 9.

Step 9. Unsupported Learning. A specially prepared procedure for autonomous improvement of the ontology should be made. It is supposed to make supplements with new descriptions of the data unit, uncertainty types, uncertainty elimination approaches, relations and rules to eliminate uncertainties. Some service resources (catalogs, vocabularies, thesauruses, etc.) should be used for that purpose. If the supplement is successful, then the Step 8 is to be repeated. Otherwise, if unsupported learning failed, then Step 10 should be made.

Step 10. Supported Learning. The problem situation to be analyzed by an expert to make a decision on how to eliminate uncertainty. According to that decision, there should be made some changes to the ontology of big data source. If the expert is unable to solve the problem situation, then the situation is marked like unsolvable. The description

of the problem situation should be made in the ontology. And then move back to Step 8.

This is the end of the algorithm. If the result of execution of all algorithm steps is that the data is cleared to the required level, then the algorithm considered to succeed. If the level of uncertainty for the data is higher than an accepted level, then an algorithm can be executed again.

Finally, the algorithm allows to achieve two aims: 1) the big data source can be cleared from incomplete, ambiguous, or other non-quality data; 2) the ontology, as a core of knowledge base, can be learned and improved. The development of the ontology is recognized to be the key to improve the results of solving the uncertainty problem of big data sources.

IV. AN EXAMPLE OF ONTOLOGY-BASED PROCESSING OF DATA UNCERTAINTY

As an example of ontology-based method elimination of data uncertainty using, input data stream of news portal is considered. Such data stream is the time-serialized sequence of news data block. News data block is formed by a robot, news aggregator etc. Usually, news data block may contain uncertain data such as absent, incorrect, invalid, or unreliable values because data were obtained from various sources. So, procedures of data uncertainty processing are needed to be performed before download news data block into repository of news portal.

In considered case, each news data block contains its ID and set of records. Each record is structured according to the document-oriented model and is called "document". Document contain description of one of news message about any new event, presented in JSON format. Proposed structure of document is the next:

News ID – unique identifier of document using for its identifying and search in the news repository,

News Category – classify message by predefined category (so as Politics, Society, Culture, Sport etc.),

News Priority – describe event value in the general news context,

News Date&Time – when presented event occurred,

News Place – where (country, region city) presented event occurred,

News Object – persons, organizations, institutions or etc., to which the event relates,

News Subject – describe what the event means,

News Source – define where did the news come from,

News Text – contains text of message about presented event and its details.

Each block of news may contain some uncertainties in the data, because it formed by special program tools using information sources of various format. Such data uncertainty types as value absence, invalid values, out of range values, and unreliable values has been defined as possible uncertainties for news data stream. Each type of data uncertainty is associated with a certain condition, which lets detect this uncertainty in data set (Table II).

Some method of data uncertainty elimination in news data stream has been defined as well in considered example. These are such as (1) repeated request for a message; (2) set default value; (3) reject message; (4) no any action.

TABLE II. THE ASSOCIATION BETWEEN UNCERTAINTY TYPES AND CONDITIONS OF ITS DETECTION IN NEWS DATA STREAM

Uncertainty type	Uncertainty condition
absent value	Is Null
invalid value	not in <i>ValueDomain</i>
value out of range	not between <i>MaxValue</i> and <i>MinValue</i>
unreliable value	not in <i>ReliableValuesSet</i>

Problem of defining what method should be applied for elimination of uncertainty of any type for certain data item processing. For solution this problem using developed method special ontology has been developed. Such tools as Protégé ontology editor and OWL ontology model were applied for ontology development.

Ontology for data uncertainty elimination in news data stream include main class **NewsDataStream** divided into three classes:

- (1) class **NewsBlock** which include entries **NewsBlockID** and **NewsMessageDoc**;
- (2) class **UncertaintyType** includes entries **InvalidValue**, **NoValue**, **OutOfRange**, **UnReliable** which corresponds to types of uncertainties in the input news stream,
- (3) class **UncertaintyProcessing** contains entries **Request**, **SetDefault**, **Reject**, **NoAction** which corresponds to uncertainty processing methods developed for input news stream data.

Entry **NewsMessageDoc** also is class includes entries corresponding to each data item of news message according to the structure of the document (see Fig.3).

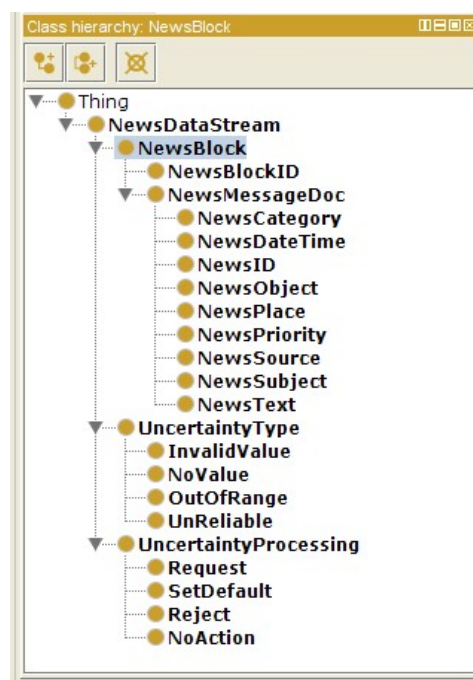


Fig. 2. Ontology of news portal input data stream structure created by Protégé

Relations between ontology classes of classes has been defined by using object properties tools. Two class of properties defined in developed ontology:

- (1) object property class **IsUncertaintyOfType** is defined for establishing correspondence between the data items (class **NewsMessageDoc**) and data uncertainty types (class **UncertaintyType**);
- (2) object property class **ProvesingMethod** is defined for establishing correspondence between the data uncertainty types (class **UncertaintyType**) and data uncertainty elimination method (class **UncertaintyProcessing**).

Each property defined to describe directed functional dependency between items of certain classes. Using of such dependencies allows define a rule for choose of uncertainty eliminated method for each given data item.

TABLE III. THE RESULTS OF EXPERT EVALUATION OF ONTOLOGY-BASED METHOD OF DATA UNCERTAINTY ELIMINATION

Evaluation criteria	Manual Uncertainty Elimination	Ontology-Based Uncertainty Elimination
Uncertainty Detection Level	low	high
Correctness of the method selection	middle	high
Performance	low	high
Absent value	high	high
Invalid value	middle	middle
Value out of range	middle	high
Unreliable value	high	middle

Experimental application of ontology has been evaluated by expert way in quality score: "low", "middle", "high". The results obtained after expert evaluation of processing about 500 news documents are presented in table III.

V. UNSOLVED PROBLEMS AND DISCUSSIONS

There are some problems appeared as a result of this research which require further discussion and investigation.

1st problem is to build an initial ontology of big data source. The problem requires some expert knowledge of data processing and detailed description of the resource. Typically, there is no such description for big data. That is

why the building of the initial ontology is not enough defined problem.

2nd problem is to qualify the type of uncertainty. The problem has no formal solution often. That is the motivation for further researches in areas of machine learning and artificial intelligence to be able to solve the problem effectively.

3rd problem is the learning and development of the ontology. Unsupported learning requires using some additional methods and algorithms based on experience and analysis of numerous precedents of data uncertainty problem solving.

The solution of these problems can be a separate task for scientific researches. Making these researches can strongly improve the method presented in the paper.

VI. CONCLUSIONS

An approach to solve an uncertainty problem for big data sources was offered in the paper. The peculiarity of the solution is to use an ontology as a core of knowledge base. The ontology can be considered as a special type of metadata. The developed approach allows, first, to make better data clearing in a big data source, and, second, to accumulate for further use a knowledge and an experience to solve data uncertainty problem.

REFERENCES

- [1] C. J. Date, *Database in Depth: Relational Theory for Practitioners*. O'Reilly, CA, 2005.
- [2] K. Aliksieieva, and A. Peleshchynshyn, "Application of incomplete and inexact data for commercial web-project management," *Scientific announcements of Lviv Polytechnic National University, Lviv, Ukraine*, no. 805, pp.345-353, 2014.
- [3] N. Marz , and J. Warren, *Big Data: Principles and best practices of scalable realtime data systems*. Manning Publications, 2015.
- [4] J. Chen, D. Dosyn, V. Lytvyn, and A. Sachenko, "Smart Data Integration by Goal Driven Ontology Learning. *Advances in Big Data*," *Advances in Intelligent Systems and Computing*, Springer International Publishing AG, pp. 283-292, 2016.
- [5] D. Losin, *Big data analytics*. Elsevier Inc., Waltham, MA, USA, 2014.
- [6] V. Aliksieiev, and O. Gaiduchok "About the problem of data losses in real-time IoT based monitoring systems," *Mathematical Modeling, STUME, Sofia, BULGARIA, Year I, issue 3*, pp. 121–122, 2017.
- [7] V. Aliksieiev, G. Ivasyk, V. Pabyrivskyi, and N. Pabyrivska, "Big data aggregation algorithm for storing obsolete data," *Industry 4.0 – STUME, Sofia, BULGARIA, Year III, issue 1*, pp.20–22, 2018.

Computational Library of the Direct Analytic Models for Real-Time Fuzzy Information Processing

Yuriy Kondratenko
Intelligent Information Systems Dept.
Petro Mohyla Black Sea National University
Mykolaiv, Ukraine
y_kondrat2002@yahoo.com

Nina Kondratenko
Darla Moore School of Business
University of South Carolina
Columbia, USA
nina.kondratenko@grad.moore.sc.edu

Abstract—This paper reveals the computational library of the analytic models for the results of fuzzy arithmetic operations with fuzzy sets. In particular, the focus is on the synthesis of the universal inverse and direct models for maximum of triangular fuzzy numbers with different masks of their parameters. The results of the study verify the efficiency of the suggested computational library with soft computing models for fuzzy information processing in real-time control and decision making.

Keywords—computational library, fuzzy number, arithmetic operation, maximum, fuzzy information processing

I. INTRODUCTION

The development of the efficient methods for big data analysis and dynamic information processing in the real-time is one of the most salient responsibilities in the signal processing, as well as control and decision making in uncertainty [1-4].

The big volume of data and high speed of its appearance requires using special mathematical approaches developed in the theory of artificial intelligence and computational optimization [5]. In some cases when the complexity of developing analytical models to ensure efficient functioning of processes and systems in the conditions of uncertainty, it is necessary to advance and develop new mathematical methods and algorithms [6]. One of these approaches, flexible to solving real-world problem, is a theory of fuzzy sets and fuzzy logic, initially developed in 1965 by Lotfi A. Zadeh [7].

Ever since, the theory of fuzzy sets has set grounds for compelling scientific and technological developments, in particular, in terms of its mathematical methods and their diverse applicability. These theoretical advancements in the theory of fuzzy sets and fuzzy logic receive a substantial attention from the global academic community [8-13].

We further proceed with analyzing a fuzzy set A as pairs $(x, \mu_A(x))$, specified on the universal set [7,14-16] and any element $x, x \in E$ of the fuzzy set A , that corresponds to the specific value of the membership function (MF) $\mu_A(x) \in [0, 1]$.

Fuzzy sets and fuzzy logic allow solving different tasks in uncertain conditions in the field of complex systems

control and decision-making in economics, management, engineering and logistics [5,6,8,17-19], in particular, in marine transportation [19-21], investment [6], finances [22] and other fields. Special attention is paid to data analysis using fuzzy mathematics and soft computing [5,23,24].

In many cases, developing the solution to the problems require fulfilling diverse fuzzy arithmetic operations, such as addition, subtraction, multiplication division, minimum and maximum calculations [4,17,25-28].

Research cautions using the inverse (horizontal) models of resulting membership functions (MFs) based on using α -cuts. It appears that using these models in solving control tasks in real time often results in compromised quality of computing operations performance [2,8,29-31].

Hence, our study aims to offer advancements in the field of universal direct analytic models that grant improvement in operating speed and accuracy of fuzzy arithmetic operations. This paper contributes to the literatures on the fuzzy information processing and data analysis [4,27]. We further proceed with developments in one of the most complex fuzzy arithmetic operations, an operation of maximum of the fuzzy numbers (FNs-maximum).

II. PROBLEM STATEMENT

Arithmetic algorithms for the FNs- maximum operations based on the α -cuts [14-16] possess high computational complexity, considering it is executed in turn for all α -levels with $\Delta\alpha$ discreteness level, which value strongly influences the computational processes' accuracy and operating speed [4,15,27,28].

Therefore, α -cuts of the fuzzy set $A \in R$ is ordinary subset $A_\alpha = \{x \mid \mu_A(x) \geq \alpha\}$, $\alpha \in [0, 1]$, that contains (Fig. 1) elements $x \in R$ whose degree of membership to a set A is not less than α . The subsets A_α та B_α that determine the appropriate α -cuts of fuzzy sets $A, B \in R$ can be written as

$$A_\alpha = [a_1(\alpha), a_2(\alpha)], B_\alpha = [b_1(\alpha), b_2(\alpha)], \alpha \in [0, 1],$$

where R is real numbers set.

The apiration of this work is to grant the synthesis of the computational library of universal analytical models of resulting MFs for the FNs-maximum of triangular fuzzy

numbers (TrFNs) with different combinations of their parameters (Fig. 1) in order (a) to increase operating speed and (b) to lower the volume, complexity and accuracy of fuzzy information processing. The TrFNs $\underline{A} = (a_1, a_0, a_2)$ and $\underline{B} = (b_1, b_0, b_2)$ have MFs $\mu_{\underline{A}}(x)$ and $\mu_{\underline{B}}(x)$ with parameters $\mu_{\underline{A}}(a_1) = \mu_{\underline{A}}(a_2) = \mu_{\underline{B}}(b_1) = \mu_{\underline{B}}(b_2) = 0$. $\mu_{\underline{A}}(a_0) = \mu_{\underline{B}}(b_0) = 1$. The inverse A_α , B_α and direct $\mu_{\underline{A}}(x)$, $\mu_{\underline{B}}(x)$ models of the TrFNs $\underline{A}, \underline{B} \in R$ can be

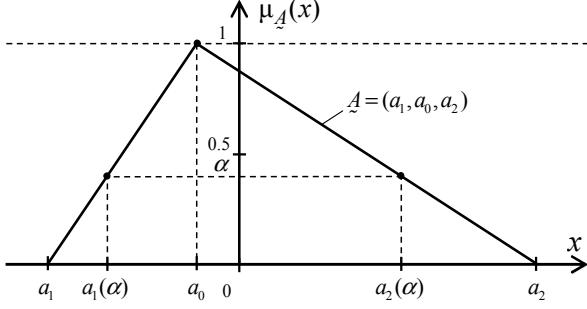


Fig. 1. Triangular Fuzzy Number \underline{A} , $\underline{A} \in R$

determined [4,14-16,27,28] by the corresponding dependencies (1)-(4):

$$A_\alpha = [a_1(\alpha), a_2(\alpha)] = [a_1 + \alpha(a_0 - a_1), a_2 - \alpha(a_2 - a_0)], \quad (1)$$

$$B_\alpha = [b_1(\alpha), b_2(\alpha)] = [b_1 + \alpha(b_0 - b_1), b_2 - \alpha(b_2 - b_0)], \quad (2)$$

$$\mu_{\underline{A}}(x) = \begin{cases} 0, \forall (x \leq a_1) \cup (x \geq a_2) \\ F_{Al}(x, a_1, a_0), \forall (a_1 < x \leq a_0), \\ F_{Ar}(x, a_0, a_2), \forall (a_0 < x < a_2) \end{cases} \quad (3)$$

$$\mu_{\underline{B}}(x) = \begin{cases} 0, \forall (x \leq b_1) \cup (x \geq b_2) \\ F_{Bl}(x, b_1, b_0), \forall (b_1 < x \leq b_0), \\ F_{Br}(x, b_0, b_2), \forall (b_0 < x < b_2) \end{cases} \quad (4)$$

where $F_{Al}(x, a_1, a_0) = (x - a_1) / (a_0 - a_1)$;

$$F_{Bl}(x, b_1, b_0) = (x - b_1) / (b_0 - b_1);$$

$$F_{Ar}(x, a_0, a_2) = (a_2 - x) / (a_2 - a_0);$$

$$F_{Br}(x, b_0, b_2) = (b_2 - x) / (b_2 - b_0).$$

Using Max-Min or Min-Max convolutions for FNs-maximum realization [14-16] in some cases results in increased complexity and lower speed of processing or to the resulting fuzzy sets with violation of the convexity and normality properties.

The operation of FNs- maximum ($\underline{C} = \underline{A}(\vee)\underline{B}$) based on α -cuts [14] can be written as

$$C_\alpha = A_\alpha(\vee)B_\alpha = [a_1(\alpha), a_2(\alpha)](\vee)[b_1(\alpha), b_2(\alpha)] = [a_1(\alpha) \vee b_1(\alpha), a_2(\alpha) \vee b_2(\alpha)] = [c_1(\alpha), c_2(\alpha)]. \quad (5)$$

III. SYNTHESIS OF INVERSE AND DIRECT RESULTING MODELS

Firstly, let us analyze the separate intersections of the left branches of TrFNs

$$F_{left}(x, a_1, a_0) \cap F_{left}(x, b_1, b_0): \underline{A}, \underline{B} \in R \quad (6)$$

and right branches of TrFNs

$$F_{Ar}(x, a_0, a_2) \cap F_{Br}(x, b_0, b_2): \underline{A}, \underline{B} \in R. \quad (7)$$

If the condition (6) exists for $\alpha \in [0, 1]$ then

$$a_1(\alpha) = b_1(\alpha) = c_1(\alpha), \quad (8)$$

and we can write

$$a_1 + \alpha(a_0 - a_1) = b_1 + \alpha(b_0 - b_1), \quad (9)$$

taking into account that

$$a_1(\alpha) = a_1 + \alpha(a_0 - a_1)$$

and

$$b_1(\alpha) = b_1 + \alpha(b_0 - b_1).$$

From (8), (9) we can find intersection parameter

$$\alpha = (b_1 - a_1) / (a_0 - a_1 - b_0 + b_1). \quad (10)$$

The value (10) corresponds to the vertical coordinate α^* of intersection point (6)

$$\alpha^* = \mu_{\underline{A}}(x^*) = \mu_{\underline{B}}(x^*) = \mu_{\underline{C}}(x^*), \quad (11)$$

where x^* is a horizontal coordinate of the intersection point (6). In this case, two pairs of coordinates $(a_1(\alpha^*), \alpha^*)$ for inverse model and $(x^*, \mu_{\underline{A}}(x^*))$ for direct model are corresponding to the intersection point (6), where $x^* = a_1(\alpha^*)$, $\mu_{\underline{A}}(x^*) = \alpha^*$. Using (1) and (3) we can find

$$a_1(\alpha^*) = a_1 + \alpha^*(a_0 - a_1) = a_1 + \frac{(b_1 - a_1)(a_0 - a_1)}{a_0 - a_1 - b_0 + b_1}, \quad (12)$$

where $a_1(\alpha^*) \in [\max(a_1, b_1), \max(a_0, b_0)]$.

If the condition (7) exists for $\alpha \in [0, 1]$, then analyzing right branches of the TrFNs (3), (4) and intersection condition (7) we can find, in the same way, two pairs of the intersection point's (7) coordinates for inverse model $(a_2(\alpha^{**}), \alpha^{**})$ and for direct model $(x^{**}, \mu_{\underline{A}}(x^{**}))$, where $x^{**} = a_1(\alpha^{**})$ and $\mu_{\underline{A}}(x^{**}) = \alpha^{**}$:

$$\alpha^{**} = (b_2 - a_2) / (b_2 - b_0 - a_2 + a_0), \quad (13)$$

$$a_2(\alpha^{**}) = a_2 - \alpha^{**}(a_2 - a_0) = a_2 - \frac{(b_2 - a_2)(a_2 - a_0)}{b_2 - b_0 - a_2 + a_0}, \quad (14)$$

where $a_2(\alpha^{**}) \in [\max(a_0, b_0), \max(a_2, b_2)]$.

Thus, the coordinates $(a_1(\alpha^*), \alpha^*)$ and $(a_2(\alpha^{**}), \alpha^{**})$ for the intersections (6) and (7) can be calculated using universal models (11)-(14) and parameters of the TrFNs $\underline{A} = (a_1, a_0, a_2)$ and $\underline{B} = (b_1, b_0, b_2)$. In case, if $a_1 < b_1, a_0 > b_0, a_2 < b_2$, the inverse and direct models of resulting MF can be presented as

$$C_\alpha = A_\alpha(\vee)B_\alpha = [a_1(\alpha) \vee b_1(\alpha), a_2(\alpha) \vee b_2(\alpha)] = [c_1(\alpha), c_2(\alpha)] = \quad (15)$$

$$\left[\left\{ \begin{array}{l} b_1(\alpha), \forall \alpha | \alpha \in [0, \alpha^*] \\ a_1(\alpha), \forall \alpha | \alpha \in [\alpha^*, 1] \end{array} \right\}, \left\{ \begin{array}{l} a_2(\alpha), \forall \alpha | \alpha \in [\alpha^{**}, 1] \\ b_2(\alpha), \forall \alpha | \alpha \in [0, \alpha^{**}] \end{array} \right\} \right]$$

$$\mu_{\underline{C}}(x) = \begin{cases} 0, \forall (x \leq b_1) \cup (x \geq b_2) \\ F_{Bl}(x, b_1, b_0), \forall (b_1 < x \leq a_1(\alpha^*)) \\ F_{Al}(x, a_1, a_0), \forall (a_1(\alpha^*) < x \leq a_0) \\ F_{Ar}(x, a_0, a_2), \forall (a_0 < x < a_2(\alpha^{**})) \\ F_{Br}(x, b_0, b_2), \forall (a_2(\alpha^{**}) < x < b_2) \end{cases}, \quad (16)$$

where $c_1(0) = b_1; c_2(0) = b_2; c_1(1) = c_2(1) = a_0$;

$$c_1(\alpha) = \left\{ \begin{array}{l} b_1 + \alpha(b_0 - b_1), \forall \alpha | \alpha \in [0, \alpha^*] \\ a_1 + \alpha(a_0 - a_1), \forall \alpha | \alpha \in [\alpha^*, 1] \end{array} \right\};$$

$$c_2(\alpha) = \left\{ \begin{array}{l} a_2 - \alpha(a_2 - a_0), \forall \alpha | \alpha \in [\alpha^{**}, 1] \\ b_2 - \alpha(b_2 - b_0), \forall \alpha | \alpha \in [0, \alpha^{**}] \end{array} \right\}.$$

IV. COMPUTATIONAL LIBRARY OF DIRECT RESULTING MODELS

The inverse C_α (15) and the direct $\mu_{\underline{C}}(x)$ (16) models for the FNs-maximum are validated only for TrFNs $\underline{A} = (a_1, a_0, a_2)$ and $\underline{B} = (b_1, b_0, b_2)$ under the following conditions:

$$a_1 < b_1, a_0 > b_0, a_2 < b_2.$$

Simultaneously much of the real input values for fuzzy processing can be conferred as TrFNs with diverse relations \mathbb{R} , $\mathbb{R} \in \{(<), (>)\}$ between parameters:

$$a_1 \mathbb{R} b_1, a_0 \mathbb{R} b_0, a_2 \mathbb{R} b_2. \quad (17)$$

Therefore, for each special case it is necessary to develop a separate analytic model of resulting fuzzy set for the performance of "FNs-maximum" if the TrFNs $(\underline{A}, \underline{B})$ have diverse relations \mathbb{R} between parameters $(a_1, b_1; a_0, b_0; a_2, b_2)$.

Let us form the set of direct analytic models of the resulting fuzzy sets \underline{C} for execution of the "maximum" as arithmetic operation with TrFNs \underline{A} and \underline{B} for diverse combinations of the relations \mathbb{R} . For evaluation of the relations \mathbb{R} and following [4, 15, 27, 28], we can determine a mask

$$\text{Mask}(\underline{A}, \underline{B}) = \{d, g, p\} \quad (18)$$

for any pair of the TrFNs \underline{A} and \underline{B} , where indicators d, g and p are defined as

$$\begin{aligned} d &= \begin{cases} 0, & \text{if } a_1 > b_1 \\ 1, & \text{if } a_1 < b_1 \end{cases} \\ g &= \begin{cases} 0, & \text{if } a_0 > b_0 \\ 1, & \text{if } a_0 < b_0 \end{cases} \\ p &= \begin{cases} 0, & \text{if } a_2 > b_2 \\ 1, & \text{if } a_2 < b_2 \end{cases} \end{aligned} \quad (19)$$

The Mask (18) is a the basis for creating a 8-component's library of the resulting mathematical models $\{M_1 \dots M_8\}$ for FNs-maximum with all possible \mathbb{R} combinations of TrFNs $(\underline{A}, \underline{B})$. The computational library $\{M_1, M_2, \dots, M_8\}$ of the developed direct models $\mu_{\underline{C}}(x)$ is represented in the Table I.

TABLE I. LIBRARY OF THE RESULTING DIRECT MODELS

Mask	M_i	Model description
$\{1,1,1\}$	M_1	$\begin{cases} 0, \forall (x \leq b_1) \cup (x \geq b_2) \\ F_{Bl}(x, b_1, b_0), \forall (b_1 < x \leq b_0) \\ F_{Br}(x, b_0, b_2), \forall (b_0 < x < b_2) \end{cases}$
$\{1,1,0\}$	M_2	$\begin{cases} 0, \forall (x \leq b_1) \cup (x \geq a_2) \\ F_{Bl}(x, b_1, b_0), \forall (b_1 < x \leq b_0) \\ F_{Br}(x, b_0, b_2), \forall (b_0 < x < a_2(\alpha^*)) \\ F_{Ar}(x, a_0, a_2), \forall (a_2(\alpha^*) \leq x < a_2) \end{cases}$
$\{1,0,1\}$	M_3	$\begin{cases} 0, \forall (x \leq b_1) \cup (x \geq b_2) \\ F_{Bl}(x, b_1, b_0), \forall (b_1 < x \leq a_1(\alpha^*)) \\ F_{Al}(x, a_1, a_0), \forall (a_1(\alpha^*) < x \leq a_0) \\ F_{Ar}(x, a_0, a_2), \forall (a_0 < x < a_2(\alpha^*)) \\ F_{Br}(x, b_0, b_2), \forall (a_2(\alpha^*) \leq x < b_2) \end{cases}$
$\{1,0,0\}$	M_4	$\begin{cases} 0, \forall (x \leq b_1) \cup (x \geq a_2) \\ F_{Bl}(x, b_1, b_0), \forall (b_1 < x \leq a_1(\alpha^*)) \\ F_{Al}(x, a_1, a_0), \forall (a_1(\alpha^*) < x \leq a_0) \\ F_{Ar}(x, a_0, a_2), \forall (a_0 < x < a_2) \end{cases}$
$\{0,1,1\}$	M_5	$\begin{cases} 0, \forall (x \leq a_1) \cup (x \geq b_2) \\ F_{Al}(x, a_1, a_0), \forall (a_1 < x \leq a_1(\alpha^*)) \\ F_{Bl}(x, b_1, b_0), \forall (a_1(\alpha^*) < x \leq b_0) \\ F_{Br}(x, b_0, b_2), \forall (b_0 < x < b_2) \end{cases}$
$\{0,1,0\}$	M_6	$\begin{cases} 0, \forall (x \leq a_1) \cup (x \geq a_2) \\ F_{Al}(x, a_1, a_0), \forall (a_1 < x \leq a_1(\alpha^*)) \\ F_{Bl}(x, b_1, b_0), \forall (a_1(\alpha^*) < x \leq b_0) \\ F_{Br}(x, b_0, b_2), \forall (b_0 < x < a_2(\alpha^*)) \\ F_{Ar}(x, a_0, a_2), \forall (a_2(\alpha^*) \leq x < a_2) \end{cases}$
$\{0,0,1\}$	M_7	$\begin{cases} 0, \forall (x \leq a_1) \cup (x \geq a_2) \\ F_{Al}(x, a_1, a_0), \forall (a_1 < x \leq a_0) \\ F_{Ar}(x, a_0, a_2), \forall (a_0 < x < a_2(\alpha^*)) \\ F_{Br}(x, b_0, b_2), \forall (a_2(\alpha^*) \leq x < b_2) \end{cases}$
$\{0,0,0\}$	M_8	$\begin{cases} 0, \forall (x \leq a_1) \cup (x \geq a_2) \\ F_{Al}(x, a_1, a_0), \forall (a_1 < x \leq a_0) \\ F_{Ar}(x, a_0, a_2), \forall (a_0 < x < a_2) \end{cases}$

V. EXAMPLE OF COMPUTATIONAL LIBRARY APPLICATION

Let's consider an example with realisation of the arithmetic operation "maximum" for the pair $(\underline{A}, \underline{B})$ of TrFNs: $\underline{A} = (3, 10, 17)$, $\underline{B} = (5, 7, 24)$. In this case, we have: $a_1 = 3$; $b_1 = 5$; $a_0 = 10$; $b_0 = 7$; $a_2 = 17$; $b_2 = 24$.

Using (18), (19), we can automatically determine (a) the corresponding Mask $(\underline{A}, \underline{B}) = \{d, g, p\} = \{1, 0, 1\}$ for the conditions $a_1 < b_1$; $a_0 > b_0$; $a_2 < b_2$ and (b) the corresponding model M_3 from the computational library of models $\{M_1, M_2, \dots, M_8\}$ (Table I). Let's calculate the coordinates

$(a_1(\alpha^*), \alpha^*)$ and $(a_2(\alpha^{**}), \alpha^{**})$ for intersection points (6) and (7) of the given fuzzy numbers $(\underline{A}, \underline{B})$ according to (12), (11), (14) and (13):

$$a_1(\alpha^*) = 3 + \frac{(5-3)(10-3)}{10-3-7+5} = 5.8;$$

$$\alpha^* = \frac{5-3}{10-3-7+5} = 0.4;$$

$$a_2(\alpha^{**}) = 17 - \frac{(24-17)(17-10)}{24-7-17+10} = 12.1;$$

$$\alpha^{**} = \frac{24-17}{24-7-17+10} = 0.7.$$

Then (for recognized M_3) we can choose the corresponding direct model $\mu_C(x)$ from the computational library (Table I). We further present the resulting inverse $C_\alpha = A_\alpha(\vee)B_\alpha$ and direct $\mu_C(x)$ models (Fig.2) for FNsmaximum $\underline{C} = \underline{A}(\vee)\underline{B}$:

$$C_\alpha = A_\alpha(\wedge)B_\alpha = \begin{bmatrix} \left\{ \begin{array}{l} 5 + 2\alpha, \forall \alpha \in [0, 0.4] \\ 3 + 7\alpha, \forall \alpha \in [0.4, 1] \end{array} \right\}, \\ \left\{ \begin{array}{l} 24 - 17\alpha, \forall \alpha \in [0, 0.7] \\ 17 - 7\alpha, \forall \alpha \in [0.7, 1] \end{array} \right\} \end{bmatrix},$$

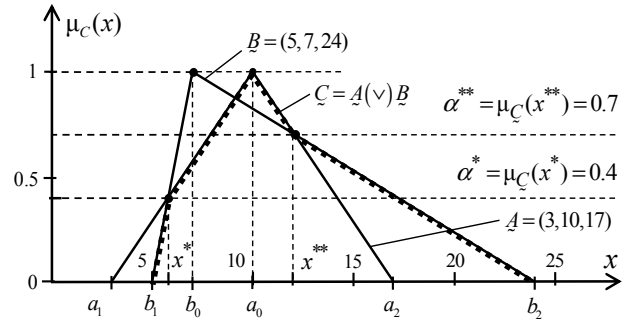


Fig. 2. FNsmaximum $\underline{C} = \underline{A}(\vee)\underline{B}$ of the TrFNs $\underline{A} \in R$ and $\underline{B} \in R$

$$\mu_C(x) = \begin{cases} 0, \forall (x \leq b_1) \cup (x \geq b_2) \\ F_{Bl}(x, b_1, b_0), \forall (b_1 < x \leq a_1(\alpha^*)) \\ F_{Al}(x, a_1, a_0), \forall (a_1(\alpha^*) < x \leq a_0) \\ F_{Ar}(x, a_0, a_2), \forall (a_0 < x < a_2(\alpha^{**})) \\ F_{Br}(x, b_0, b_2), \forall (a_2(\alpha^{**}) \leq x < b_2) \end{cases},$$

$$\mu_C(x) = \begin{cases} 0, \forall (x \leq 5) \cup (x \geq 24) \\ (x-5)/2, \forall (5 < x \leq 5.8) \\ (x-3)/7, \forall (5.8 < x \leq 10) \\ (17-x)/7, \forall (10 < x < 12.1) \\ (24-x)/17, \forall (12.1 \leq x < 24) \end{cases},$$

VI. CONCLUSION

The maximum of fuzzy sets is an essential fuzzy arithmetic operation, which is often time consuming in terms of its realization. The execution of the developed direct analytic models' library (Table I) advances current research by allowing the usage of one step automation mode for "FNs-maximum" $C = A(\vee)B$ operation. In some instances, it is necessary to aggregate different data streams (big data), which are presented as random time series or random consequences [32]. For fuzzy information processing of such random streams or consequences, we can use three steps algorithm:

Step 1. Each random stream or consequence can be evaluated by interval value and presented as fuzzy set or fuzzy number [16,32]. For example, in [16] the realization of such random sequences is presented as triangular fuzzy number of such types - "approximate A" or "between B and C";

Step 2. For each pair of triangular fuzzy numbers it is necessary to determine the mask [4,27,28], for example (18), according to the relations between considered TrFNs parameters;

Step 3. Using corresponding computational library for desired fuzzy arithmetic operation [33-38] (addition, subtraction, multiplication, division, minimum or maximum) it is possible to find the resulting MF based on TrFNs mask and TrFN parameters. For FN's-maximum we can use corresponding computational library, presented in Table I.

Modeling results confirm the efficiency of proposed universal direct analytic models for different applications. In some cases, such direct analytic models $\mu_C(x) = \mu_{A(\vee)B}(x)$ provide an efficient solution to the fuzzy processing in evaluation, control and decision-making processes, in particular, for the financial analysis [22], automatic evaluation of the student's knowledge [39], group anonymity [40] and partner selection [41,42], model design process [43], soft computing based on reconfigurable technology [44], analysis of the big data during testing of computer systems and their components [5,45], optimization of tanker or truck routes in the conditions of fuzzy demands at nodes [10,21,46], redesigning social inquiry [47], fuzzy-algorithmic reliability analysis of complex systems in economics, management and engineering [18, 48,49], fuzzy control in industrial processes and robotics [2,11,17,50], and others.

REFERENCES

[1] D. Simon, "Design and rule base reduction of a fuzzy filter for the estimation of motor currents," *International Journal of Approximate Reasoning*, 25, pp. 145-167, 2000.

[2] Y. P. Kondratenko, and E. Y. M. Al Zubi, "The Optimisation Approach for Increasing Efficiency of Digital Fuzzy Controllers," 20th Int. DAAAM Symp. "Intelligent Manufacturing and

Automation", Published by DAAAM International, Vienna, Austria, pp. 1589-1591, 2009.

[3] S. Encheva, Y. Kondratenko, M. Z. Solesvik, S. Tumin, "Decision Support Systems in Logistics," *AIP Conference Proc.* 1060, 254-256, 2008, <https://doi.org/10.1063/1.3037065>

[4] Y. P. Kondratenko, N. Y. Kondratenko, "Soft Computing Analytic Models for Increasing Efficiency of Fuzzy Information Processing in Decision Support Systems," Chapter in book: *Decision Making: Processes, Behavioral Influences and Role in Business Management*, R. Hudson (Ed.), Nova Science Publishers, New York, 2015, pp. 41-78.

[5] J. Kacprzyk, S. Zadrozny, and G. De Tré, "Fuzziness in database management systems: Half a century of developments and future prospects," *Fuzzy Sets and Systems* vol. 281, pp. 300-307, 2015, <https://doi.org/10.1016/j.fss.2015.06.011>

[6] J Gil-Aluja., *Investment in Uncertainty*. Kluwer Academic Publishers, Dordrecht, Boston, London, 1999.

[7] L. A. Zadeh, "Fuzzy Sets," *Information & Control*, 8, pp. 338-353, 1965.

[8] R. Yager, and D. Filev, *Essential of Fuzzy Modeling and Control*. Wiley, New York, 1994.

[9] M. Jamshidi, V. Kreinovitch, J. Kacprzyk, (Eds), *Advance Trends in Soft Computing. Proceedings of WCSC 2013*, San Antonio, Texas, USA, STUDEFUZZ 312, Springer, 2013.

[10] B. Werners, and Y. Kondratenko, "Alternative Fuzzy Approaches for Efficiently Solving the Capacitated Vehicle Routing Problem in Conditions of Uncertain Demands," *Complex Systems: Solutions and Challenges in Economics, Management and Engineering*, C. Berger-Vachon et al. (Eds.), *Studies in Systems, Decision and Control*, vol. 125, Berlin, Heidelberg: Springer, 2018, pp. 521-543. DOI: https://doi.org/10.1007/978-3-319-69989-9_31

[11] J. M. Merigo, A. M. Gil-Lafuente, and R. R. Yager, "An overview of fuzzy research with bibliometric indicators," *Applied Soft Computing*, no. 27, pp. 420-433, 2015.

[12] D. E. Tamir, N. D. Rishe, A. Kandel, (Eds), *Fifty Years of Fuzzy Logic and its Applications. Studies in Fuzziness and Soft Computing*, vol. 326, Cham-Heidelberg-New York-Dordrecht-London, Springer International Publishing Switzerland, 2015.

[13] Lodwick, W.A., Kacprzyk, J. (Eds): *Fuzzy Optimization. Studies in Fuzziness and Soft Computing* 254, Springer-Verlag, Berlin, Heidelberg, 2010.

[14] A. Kaufmann, and M. Gupta, *Introduction to Fuzzy Arithmetic: Theory and Applications*. Van Nostrand Reinhold Company, New York, 1985.

[15] Y. Kondratenko, and V. Kondratenko, "Soft Computing Algorithm for Arithmetic Multiplication of Fuzzy Sets Based on Universal Analytic Models," In: *Information and Communication Technologies in Education, Research, and Industrial Application. Communications in Computer and Information Science*, vol. 469, Ermolayev, V. et al. (Eds): ICTERI'2014, Springer International Publishing, 2014, pp. 49-77. DOI: 10.1007/978-3-319-13206-8_3

[16] A. Piegat, *Fuzzy Modeling and Control*. Springer, Heidelberg, 2001.

[17] L. E. S. Pereira, and V. M. da Costa, "Interval analysis applied to the maximum loading point of electric power systems considering load data uncertainties," *International Journal of Electrical Power & Energy Systems*, vol. 54, pp. 334, 2014.

[18] M. Hanss, *Applied Fuzzy Arithmetics: An Introduction with Engineering Applications*. Springer, Berlin, Heidelberg, New York, 2005.

[19] M. Solesvik, Y. Kondratenko, G. Kondratenko, et al., "Fuzzy decision support systems in marine practice," *IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*, Naples, Italy, pp. 1-6, 9-12 July 2017. DOI: 10.1109/FUZZ-IEEE.2017.8015471

[20] P. Toth, D. Vigo, (Eds), *The vehicle routing problem*. SIAM, Philadelphia, 2002

[21] D. Teodorovic, and G. Pavkovich, "The fuzzy set theory approach to the vehicle routing problem when demand at nodes is uncertain," *Fuzzy Sets and Systems*, vol. 82, pp. 307-317, 1996

[22] A. M. Gil-Lafuente, *Fuzzy Logic in Financial Analysis. Studies in Fuzziness and Soft Computing*, vol. 175, Springer, Berlin, 2005.

[23] Y. Bodyansky, O. Vynokurova, I. Pliss, and P. Mulesa, "Multilayer Wavelet-Neuro-Fuzzy Systems in Dynamic Data Mining Tasks," In: *Soft Computing: Developments, Methods and Applications*, Alan

- Casey (Ed), Series: Computer Science, technology and applications, Nova Science Publishers, Hauppauge, NY, 2016, pp. 69-146.
- [24] M. Hanss, "Fuzzy Arithmetic for Uncertainty Analysis," In: Seising R., Trillas E., Moraga C., Termini S. (Eds), *On Fuzziness. Studies in Fuzziness and Soft Computing*, vol 298. Springer, Berlin, Heidelberg, 2013, pp. 235-240. DOI https://doi.org/10.1007/978-3-642-35641-4_36
- [25] S. Gao, and Z. Zhang, "Multiplication Operation on Fuzzy Numbers," *Journal of Software*, vol. 4(4), pp. 331-338, JUNE 2009. <http://ojs.academypublisher.com/index.php/jsw/article/download/0404331338/1061>.
- [26] A. Klimke, "An efficient implementation of the transformation method of fuzzy arithmetic," 22nd International Conference of the North American Fuzzy Information Processing Society, NAFIPS 2003, pp. 468-473, 2003. doi: 10.1109/NAFIPS.2003.1226830
- [27] Y. P. Kondratenko, and N. Y. Kondratenko, "Reduced library of the soft computing analytic models for arithmetic operations with asymmetrical fuzzy numbers," In: *Soft Computing: Developments, Methods and Applications*, Alan Casey (Ed), Series: Computer Science, technology and applications, Nova Science Publishers, Hauppauge, NY, 2016, pp. 1-38.
- [28] Y. P. Kondratenko, and N. Y. Kondratenko, "Synthesis of Analytic Models for Subtraction of Fuzzy Numbers with Various Membership Function's Shapes," In: Gil-Lafuente A. et al. (Eds), *Applied Mathematics and Computational Intelligence. FIM 2015. AISC*, vol 730. Springer, Cham, 2018, pp.87-100, DOI https://doi.org/10.1007/978-3-319-75792-6_8
- [29] M. Pasięka, N. Grzesik, and K. Kuźma, "Simulation modeling of fuzzy logic controller for aircraft engines," *International Journal of Computing*, vol. 16(1), pp. 27-33, 2017. Retrieved from <http://computingonline.net/computing/article/view/868>
- [30] Y. P. Kondratenko, O. V. Kozlov, O. S. Gerasin, and Y. M. Zaporozhets, "Synthesis and research of neuro-fuzzy observer of clamping force for mobile robot automatic control system," *IEEE First International Conference on Data Stream Mining and Processing (DSMP)*, pp.90-95, 2016, DOI: 10.1109/DSMP.2016.7583514
- [31] P. Bykovyy, Y. Pigovsky, A. Sachenko, and A. Banasik, "Fuzzy inference system for vulnerability risk estimation of perimeter security," 5th IEEE International Workshop on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications, IDAACS'2009, pp. 380-384, 2009. DOI 10.1109/IDAACS.2009.5342956
- [32] Y. Bodyansky, O. Vynokurova, I. Pliss, and D. Peleshko, "Hybrid Adaptive Systems of Computational Intelligence and Their On-line Learning for Green IT in Energy Management Tasks," In: *Green IT Engineering: Concepts, Models, Complex Systems Architectures*, Studies in Systems, Decision and Control, V. Kharchenko et al. (Eds.), Vol. 74. Berlin, Heidelberg: Springer International Publishing, 2017, pp. 229-244. DOI: 10.1007/978-3-319-44162-7_12
- [33] M. Hanss, "An Approach to Inverse Fuzzy Arithmetic," 22nd International Conference of the North American Fuzzy Information Processing Society, NAFIPS 2003, pp. 474-479, 2003. doi: 10.1109/NAFIPS.2003.1226831
- [34] L. Stefanini, L. Sorini, and M. L. Guerra, "Fuzzy Numbers and Fuzzy Arithmetic," In: *Handbook of Granular Computing*, W. Pedrycz, A. Skowron, V. Kreinovich (Eds.), John Wiley and Sons, 2008, pp. 249-284.
- [35] P. Grzegorzewski, "On the Interval Approximation of Fuzzy Numbers," In: Greco, S., Bouchon-Meunier, B., Coletti, G., Fedrizzi, M., Matarazzo, B., Yager, R.R. (Eds.): *Advances in Computational Intelligence*, IPMU 2012. Communications in Computer and Information Science, vol 299. Springer, Berlin, Heidelberg, 2012, pp. 59-68. DOI https://doi.org/10.1007/978-3-642-31718-7_7
- [36] M. L. Guerra, and L. Stefanini, "Approximate fuzzy arithmetic operations using monotonic interpolations," *Fuzzy Sets and Systems*, vol. 150, iss. 1, pp. 40-55, 2005.
- [37] L. Stefanini, "A generalization of Hukuhara difference and division for interval and fuzzy arithmetic," *Fuzzy Sets and Systems*, vol. 161, iss. 11, pp. 1564-1584, 2010.
- [38] L. Coroianu, "Necessary and sufficient conditions for the equality of the interactive and non-interactive sums of two fuzzy numbers," *Fuzzy Sets and Systems*, vol. 283, pp. 40-55, 2016.
- [39] S. N. Shahbazova, "Application of Fuzzy Sets for Control of Student Knowledge," *An International Journal Applied and Computational Mathematics* 10, no. 1, pp. 195-208, 2011.
- [40] O. Chertov, and D. Tavrov, "Memetic algorithm for solving the task of providing group anonymity," In: Jamshidi M., Kreinovich V., Kacprzyk J. (Eds), *Advance Trends in Soft Computing. Studies in Fuzziness and Soft Computing*, vol. 312. Springer, Cham, 2014, pp. 281-292. DOI: https://doi.org/10.1007/978-3-319-03674-8_27
- [41] M. Z. Solesvik, and S. Encheva, "Partner selection for interfirm collaboration in ship design," *Industrial Management & Data Systems*, 110(5), pp. 701-717, 2010.
- [42] M. Solesvik, "Partner selection in green innovation projects," In: Berger-Vachon C., et al. (Eds.), *Complex Systems: Solutions and Challenges in Economics, Management and Engineering*, Studies in Systems, Decision and Control, vol. 125, Springer, Cham 2018, pp. 471-480.
- [43] A. N. Trunov, "An adequacy criterion in evaluating the effectiveness of a model design process," *Eastern-European Journal of Enterprise Technologies* 1, vol. 4 (73), pp. 36-41, 2015.
- [44] A. V. Palagin, and V. N. Opanasenko, "Reconfigurable Computing Technology," *J. Cybernetics and Systems Analysis*, vol. 43(5), pp. 675-686, 2007.
- [45] J. Drozd, and A. Drozd, "Models, methods and means as resources for solving challenges in co-design and testing of computer systems and their components," 9th Intern. Conf. on Digital Technologies 2013. Zhilina, Slovak Republic, pp. 225-230, 29-31 May, 2013.
- [46] Y. P. Kondratenko, S. B. Encheva, and E. V. Sidenko, "Synthesis of Intelligent Decision Support Systems for Transport Logistic," 6th IEEE Intern. Conf. on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications, Prague, Czech Republic, vol. 2, pp. 642-646, Sept. 15-17, 2011. DOI: 10.1109/IDAACS.2011.6072847
- [47] C. C. Ragin, *Redesigning Social Inquiry: Fuzzy Sets and Beyond*. The University of Chicago Press, Chicago, London, 2008.
- [48] A. P. Rotshtein, "Fuzzy-algorithmic reliability analysis of complex systems," *Cybern. Syst. Anal.*, vol. 47 (6), pp. 919-931, 2011. DOI <https://doi.org/10.1007/s10559-011-9371-x>
- [49] O. Pomorova, O. Savenko, S. Lysenko, and A. Kryshchuk, "Multi-Agent Based Approach for Botnet Detection in a Corporate Area Network Using Fuzzy Logic," In: Kwiecien, A., Gaj, P., Stera, P. (eds.) 20th Conference on Computer Networks, CN 2013, Lwówek 'Slaski, Poland. Communications in Computer and Information Science, vol. 370, pp. 146-156. Springer-Verlag Berlin Heidelberg, 2013.
- [50] D. Dubois, "An application of fuzzy arithmetic to the optimization of industrial machining processes," *Mathematical Modelling*, vol. 9, no. 6, pp. 461-475, 1987.

Models of Magnetic Driver Interaction with Ferromagnetic Surface and Geometric Data Computing for Clamping Force Localization Patches

Oleksandr Gerasin
*Computerized Control Systems
Department*
Admiral Makarov National University
of Shipbuilding
Mykolaiv, Ukraine
oleksandr.gerasin@nuos.edu.ua

Yuriy Zaporozhets
*Intelligent Information Systems
Department*
Petro Mohyla Black Sea National
University
Mykolaiv, Ukraine
umz-286@bigmir.net

Yuriy Kondratenko
*Intelligent Information Systems
Department*
Petro Mohyla Black Sea National
University
Mykolaiv, Ukraine
y_kondrat2002@yahoo.com

Abstract— This paper deals with the main features of electromagnetic interaction between the mobile robot's (MR) electromagnetic driver and ferromagnetic surface. In particular, the problems of clamping force calculation are discussed for robotic applications. The main attention is paid to computing of geometrical parameters of clamping forces localization patches by proposed combination-generation method. Efficiency of the method is checked by computer simulation and shown for one arc of localization patch's contour. Obtained results show rather high accuracy of this processing and its ability to be used in a functional structure of an experimental setup for fast Hall sensors' data stream computing at clamping force calculation by computerized tools.

Keywords— *clamping device, mobile robot, clamping force, simulation, control system, finite elements, data processing.*

I. INTRODUCTION

Complicated industrial complexes are widely used in modern industrial conditions, which fully or partially equipped with the objects of robotics [1,2]. Thus, industrial robots are successfully used in machine building for metalworking and assembly of prefabricated constructions. Such systems allow not only to significantly increase the productivity of the given technological operations, but also to reduce the risks to human health and life under hazardous conditions [3-5]. This is especially true for execution of works related to the release of aggressive and toxic substances indoors and outdoors or in conditions of increased danger (radiation, elevated temperature, high-altitude works) [6-9]. To accomplish such tasks, the most successful solution is the use of mobile robots (MRs), equipped with technological tools and means of vertical movement in the form of appropriate movers.

In the domestic production the need for MRs which are able to move safely along the sloping and vertical surfaces is actual in the following areas: shipbuilding and ship repair, oil and gas refining and transportation, the agrarian sector [3,4,6-9]. Cleaning, cutting, welding, polishing, painting and inspection of large areas, as well as installation of individual elements and fire extinguishing on vessels, bridge supports, tanks and large diameter pipelines, elevators are the main

tasks that can be successfully performed by such MRs. It should be noted that the majority of working surfaces in the above-mentioned applications have a ferromagnetic nature, therefore magnetic or magnetically operated clamping devices (CDs) and propulsion systems are used in such MRs' structures, which are able to provide better control of adhesion with the working surface and high speed of movement with electromagnets and permanent magnets [3-7], unlike pneumatic [2] or vacuum [9] CDs.

Magnetic and magnetically operated CDs are widely used not only in the tasks of robotics. It is difficult to do without them when carrying out black metals lifting-transport operations, assembly of prefabricated designs, machining in the constructions of magnetic tables and magnetic suspension apparatus. So, CDs can be used in almost all types of machining in single, serial and mass production [10,11]. At present, it is difficult to find an industrial enterprise where magnetic and electromagnetic plates, chucks, lifting devices, demagnetizers, etc., isn't used. However, the common problem in such systems is the complexity of determining the clamping force (CF), created by magnetic CD, and actually there are no developed systems for its control. Therefore, the widespread implementation of advanced technological devices using the energy of the magnetic field in the industry requires essential improvement of these devices and makes particularly relevant works about its research and optimization [10].

So, **the main aim of the work** is comparative analysis of applied tasks of magnet's driver and ferromagnetic surfaces interaction in context of computing of geometrical parameters of clamping forces localization patches by using of considered models of their interaction with proposed combined method for different applications in robotics, industry and services.

II. ANALYSIS OF THE MAIN MODELS OF CD'S INTERACTION WITH FERROMAGNETIC SURFACE AND CF CALCULATION

For the above applications, different types of CDs are used [3-8,10,11], in particular:

- magnetic based on permanent magnets;

- magnetically operated on the basis of electromagnets (direct and alternating current supply);
- composite, which are built on the principle of permanent magnets with electromagnets combinations or magnetic assemblies, e.g. Halbach assemblies.

Structurally, CDs can be in different ways mounted in the designs of robots, machine tools and grippers: stationary and fixedly with a fixed air gap or with several angular/linear degrees of freedom (then, most often, the gap varies during the operation) – in relation to the working ferromagnetic surface (FS) [12-15]. However, there are situations in operation conditions when the air gap between the CD and the FS can be changed in an unpredictable manner. For example, in shipbuilding and ship repair we can observe the conditions of local uncertainty in the area of the pressed magnet to the working FS: varying thickness, the presence of significant pollution, excrescences, holes or damages, high welds. Despite the fact that the surface of the ship hull is globally determined (there are drawings, the main dimensions can be measured), the local uncertainties of the ferromagnetic and non-ferromagnetic nature have a decisive effect on the distribution of magnetic induction on the working FS and the efficiency of the magnetic field energy use to create the appropriate value of the CFs within its localization [16].

Next, consider the basic existing methods of CF calculating for CDs to determine the most suitable for use in conditions of local uncertainty of the working FS and the need to control the created force.

A. Energy Approach

This approach is to determine the permeance of the gap between the surface of the magnet and the FS at different configurations of this gap and the probable paths of the magnetic flux portions. By this method, the dependence of the electromagnetic force (created by the CD) on the gap is a reversed power function with an index equal to 2 [17,18], so even a slight increase in the gap results in a significant reduction in the value of the CF. In general clamping force F can be defined considering the energy balance at the electromagnet anchor motion and air gap changing without taking into consideration dissipation flows, constant value of magnetomotive force IW and unsaturated magnetic system as

$$F = -\frac{(IW)^2}{2} \cdot \frac{dG}{d\delta}, \quad (1)$$

where G – the gap's magnetic permeance taken for the current anchor position, δ – air gap thickness.

Another case of the use of the energy approach to determining the clamping force applies to the magnetic systems that operate at constant flux linkage. Then without taking into account dissipation flows

$$F = -\frac{1}{4} \cdot \frac{\Phi_{\delta m}^2}{G^2} \cdot \frac{dG}{d\delta}, \quad (2)$$

where $\Phi_{\delta m}$ – maximum value of magnetic flux in the gap [18].

Equations (1) and (2) show that such approach assumes the constancy of magnetomotive force or the flux linkage. Therefore, in case of the necessity of the CF control or the local uncertainty of the FS, it is not suitable for further consideration.

B. Sources Interaction Method (Coulomb Approach)

This approach is to determine the interaction of field carriers – superficial and volume magnetic charges and magnetic dipoles. So, it is necessary to use the means of field theory, by which one of the spatial characteristics of the field distribution (magnetic dipoles or conditional magnetic charges) can be determined [19,20].

To implement this method, various techniques and approaches are used. In particular, the well-described and mathematically uncomplicated is the use of mirror images method, according to which a field of the charge located next to a well-conducting surface (FS in our case) can be found. In this case, the influence of the whole conducting surface (the effect of the charges induced on it) is replaced by the field of the mirror image of the given charge with a reversed sign [21]. A calculation scheme, a mathematical model, a methodology and an example of determining the basic parameters of a magnetic field for a walking MR with separate CDs are given in [5]. In the calculation process the authors calculate the normal and tangential components of the CF, as well as the main vector and the main moment acting on the clamping magnet [22]. The limitation of this approach lies in the fact that it is well suited for defining the CF of CD based on permanent magnets. When describing a permanent magnet, its magnetization remains (is) the same over its entire surface (for the rectangular magnet in [5]) in accordance with the adopted CM's model. While for an electromagnet the density of "magnetic charges" is a function of the coordinates of the pole face, which is determined by the solution of the field problem. However, the general method of finding images for any problem in the case of several boundary surfaces has not been developed, but in some examples [23] the method of successive approximations gives the correct results.

C. Approach Based on Artificial Intelligence Means

Neuro-fuzzy observers of CF, which is created by electromagnetic MR's CD, are proposed in [24,25] and for another applications – in [26-28]. According to the data on the supplying voltage and spatial position of the electromagnet, which are measured experimentally, the authors sufficiently accurately determine the CF value. Such approach can be found in applications for CDs based on a permanent magnet, as well as for combined CDs. However, the disadvantage of this method is that for each CD it is necessary to carry out a series of measurements to form the training and test samples of the proposed hybrid observers before putting MR into operation. In addition, such a technique does not take into account the peculiarities of the electromagnetic interaction of the CD with a FS in the presence of local uncertainties of the ferromagnetic nature (holes, damages, weld seams, structural elements, etc.).

At the same time, the authors select the phenomenological basis for the functioning of the MR's magnetically operated wheel mover, that consists in the CF formation of the electromagnetic CD by means of the magnetic interaction (i.e. by means of a magnetic field) of the FS's elements with the magnet's poles. Obviously that

the main factors of this interaction can be fully taken into account at computing the CDs and the value of the required guaranteed value of their strengths, if we take as a basis the Laplace equation and the method partly outlined in [5].

III. COMBINATION-GENERATION ALGORITHM FOR COMPUTING OF GEOMETRICAL PARAMETERS OF CLAMPING FORCES LOCALIZATION PATCHES

In [5] authors propose an idea of clamping force automatic control and monitoring system based on Hall sensors' measurements in the feedback. In this case, the Hall sensors measure the current values of magnetic induction on the clamping surface of the CD, by which the CF value can be indirectly determined. The informing-controlling system must process a big data stream with at least 5 sensors to estimate the dimensions of the patch (area) of the created CF localization (Fig. 1, a) and CF's value. In this case the sizes of the patch refer to the points of its breakdown into elementary area parts – the finite elements (FEs) – both for the area of the patch, and its contour. Based on practical experience, for a patch (created by one MR's CD) a partition at one coordinate must be applied up to 100÷200 points. Therefore, when considering a two-dimensional problem an array of patch's FEs will consist of 10,000 to 40,000 cells. But for solving of a field task it is required to create a square matrix of equations coefficients which may have already an order from 10^8 to 10^{10} . Such number of data items must be formed just for one of the patch's geometric parameters set (for all linear or angular parameters up to 10). The amount of data that needs to be processed at this stage is substantially expanded due to the availability of several CDs (depending on the MR's design) and the speed of the MR's movement. The last factor reduces the time of processing continuous data stream from sensors at computing CF and requires the development of fast and reliable algorithms to determine the main geometric parameters of the patch and the created strength value [29].

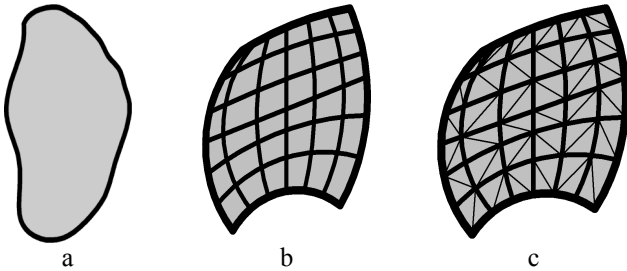


Fig. 1. An arbitrary form of the CF localization patch (a), the breakdown of the patch by arches (b), and an illustration of the formed FE after splitting (c).

The key tool for determining the clamping strengths by the method of sources interaction is the algorithm for calculating the main geometric parameters of the FS's parts – patches, on which the CF is directly created (the places of CF localization). Due to the existing local uncertainties of the working surface (ferromagnetic and non-ferromagnetic), the main area of CF creation may have an irregular shape (while neglecting the leakage flux with less than 5% of the main flow of the CD). So, the contour of this region can have some complex or arbitrary shape (Fig. 1, a), and therefore it is important to get its geometric parameters with high accuracy.

The active CF value is determined after solving the Poisson equation by finding the integral sum of the forces created by each individual FE of the FS [30]. To do this, it is necessary to split the formed patch into the grid of FEs by tracing separate lines or arcs (Fig. 1, b). Such approach is also used in problems of approximation of complex spatial surfaces, including the ship's hull [31,32]. The points of intersection of these lines will be nodal points for the elementary pieces of the patch area, on which elementary clamping strength are created.

The uniform grid of FE consisting of equilateral triangles is established as optimal in the theory of numerical methods [33,34]. But, it is not possible to satisfy such a requirement for a complex surface with variable curvature. Therefore, the criterion of the quality of the approximation should be the minimum deviation of the internal angles of triangular FEs (Fig. 1, c) from the canonical value of 60° with a minimal spread of their areas. These conditions can be formulated as

$$K = [\min \varphi_i] \cap [S_{\min} \leq S \leq S_{\max}], i = 1..3, \quad (3)$$

where φ_i – internal angles of triangular FE; S – FE's area; S_{\min}, S_{\max} – restrictions on the spread of the FEs areas. Then the optimal grid will be the FEs grid, each element of which is formed by the maximum K criterion condition within the given sample of node points. Thus, the possibility of optimizing the FEs grid is due to the presence of a sufficiently wide sample. However, traditional combinational or generational approaches impose significant limitations in this sense [31]. The first of them practically does not allow for any variations. And the second, being a locally deterministic procedure, does not guarantee the convergence of the grid to a given region on a complicated surface.

Therefore, a synthetic approach is proposed, the essence of which is the design of an isoparametric lattice covering the whole region on which an excess amount of points that belong to a given surface is formed using spline approximation. The triangular FEs (are optimal in the sense of the formulated criterion K) are formed on this set of points. Such algorithm is a globally deterministic procedure that performs local optimization of grid parameters of the FEs, and can be implemented as software.

Thus, the proposed in [5,24] approach for computing CD based on the Laplace equation can be described in several stages of determining the basic patches' geometric parameters. Each stage, in turn, consists of certain actions, some of which are separate applied geometric problems that can be used as a fully prepared mathematical apparatus for solving problems related to the field theory when processing continuous streams of input data of sufficiently large dimension. So, it is expedient to carry out the statement of the essence of such processing in the form of an algorithm that is convenient for next modeling.

Stage 1. Specification of the source data for the patch's configuration. It is first necessary to specify the contour of the working FS on which the calculations are made to compute the CDs and created CFs. For this, the number of arcs M (in the general case, of different curvatures) constituting the contour of the boundary of the patch (not less than 2, M is an integer) is given. Then give the coordinates $(X_{ip,j}, Y_{ip,j})$ of initial point of the next arc, coinciding with the coordinates of the end point of the

previous arc $(X_{ep,j-1}, Y_{ep,j-1})$, where j – arc number of the arbitrary number of arcs M . After that by specifying the curvature radii of the arcs $R_{c,j}$ we introduce the restriction: $R_{c,j}$ must be at least half the length l_j of the segment of the straight line connecting the beginning and end of the arc. If $R_{c,j} = 0,5l_j$ is specified then the arc will be a semicircle; if $R_{c,j} > 100l_j$ then the arc becomes a straight line segment. Moreover, the positive value $R_{c,j}$ indicates that when passing this arc (motion from the initial to the end point) the point of the center of curvature is situated on the right. So, the negative value of $R_{c,j}$ indicates that when passing this arc (motion from the initial to the end point) the point of the center of curvature is situated on the left.

Stage 2. Coordinates calculation of the arcs' centers of curvature, sector angles, directions of the chord and the normal of the corresponding arcs. We calculate the coordinates $(X_{cc,j}, Y_{cc,j})$ of the arcs' centers of curvature and the sector angles γ_j corresponding to each of them by using Fig. 2, which shows the coordinate plane with one arc as a separate part of the entire contour. Main designations at Fig. 2 are IP – initial point, EP – end point, CC – center of circle, CP – central point of the arc's chord, α – angle of the arc chord inclination to the vertical, β – angle of the arc chord inclination to the horizontal.

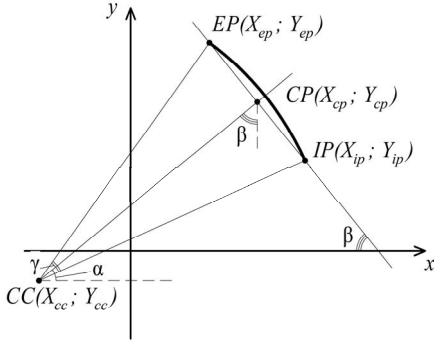


Fig. 2. Scheme for determining the main geometric parameters of the created by CD CF localization patches (for the first quadrant of the coordinate plane)

Initially, the coordinates of the point CP $(X_{cp}; Y_{cp})$ are calculated as the half-sum of each corresponding coordinate [35]. Half of the \widehat{IPEP} arc chord length is obtained from the formula for determining the length of the segment denoted by l_c . Then the height of an isosceles triangle h_c formed by a chord and two radii R_c drawn from the common center of curvature of the arc to its ends IP and EP , it is also the bisector of the sector angle γ

$$h_c = \sqrt{R_c^2 - l_c^2}, \quad (4)$$

$$\gamma = 2\arctg(l_c/h_c). \quad (5)$$

The angle of the \widehat{IPEP} arc chord inclination to the horizontal and it is the angle of inclination of the bisector of the sector angle γ to the vertical

$$\beta = \arctg \left| \frac{Y_{ep} - Y_{ip}}{X_{ep} - X_{ip}} \right| = \arctg \left| \frac{\Delta Y_{ipep}}{\Delta X_{ipep}} \right|. \quad (6)$$

Similarly, the angle of inclination of the arc chord to the vertical, and for the bisector of the sector angle γ – to the horizontal

$$\alpha = \arctg \left| \frac{X_{ep} - X_{ip}}{Y_{ep} - Y_{ip}} \right| = \arctg \left| \frac{\Delta X_{ipep}}{\Delta Y_{ipep}} \right|. \quad (7)$$

Note here that the determination of the angles α and β from (6) and (7) is valid in the case of the location of the arc in the first quadrant (Fig. 2). Features of the calculation for the other cases are making some changes in the determination of the direction of the normal and the chord of the any arc, which are taking into account by the following equations for α_{rez} and β_{rez}

$$\beta_{rez} = \begin{cases} \beta, & \text{if } \Delta X_{ipep} > 0 \cap \Delta Y_{ipep} > 0 \\ \pi - \beta, & \text{if } \Delta X_{ipep} < 0 \cap \Delta Y_{ipep} > 0 \\ -\pi + \beta, & \text{if } \Delta X_{ipep} < 0 \cap \Delta Y_{ipep} < 0 \\ -\beta, & \text{if } \Delta X_{ipep} > 0 \cap \Delta Y_{ipep} < 0 \end{cases}, \quad (8)$$

$$\alpha_{rez} = \begin{cases} \alpha, & \text{if } \Delta X_{cccp} > 0 \cap \Delta Y_{cccp} > 0 \\ \pi - \alpha, & \text{if } \Delta X_{cccp} < 0 \cap \Delta Y_{cccp} > 0 \\ -\pi + \alpha, & \text{if } \Delta X_{cccp} < 0 \cap \Delta Y_{cccp} < 0 \\ -\alpha, & \text{if } \Delta X_{cccp} > 0 \cap \Delta Y_{cccp} < 0 \end{cases}, \quad (9)$$

where $\Delta X_{cccp} = X_{cp} - X_{cc}$, $\Delta Y_{cccp} = Y_{cp} - Y_{cc}$ (according to Fig. 2). Analogous equations determine the angles α_{ip} and α_{ep} – the directions from the point of the center of arc's curvature to the initial and end points, respectively.

Further the coordinates of the point CC are determined taking into account the position of the arc on the coordinate plane and its curvature:

$$\begin{aligned} \text{if } R_c < 0 &\Rightarrow \\ \Rightarrow X_{cc} &= X_{cp} - h_c \sin \beta_{rez}; Y_{cc} = Y_{cp} + h_c \cos \beta_{rez}, \end{aligned} \quad (10)$$

$$\begin{aligned} \text{if } R_c > 0 &\Rightarrow \\ \Rightarrow X_{cc} &= X_{cp} + h_c \sin \beta_{rez}; Y_{cc} = Y_{cp} - h_c \cos \beta_{rez}. \end{aligned} \quad (11)$$

The lengths of each of the arcs l_j [35] and the total length of the contour $L = \sum l_j$ are calculated at the end of the second stage.

Stage 3. Calculation of the parameters of finite elements. The number of elementary sections m_j , into which each arc is divided, is set so that the $\Delta l_j = l_j/m_j \approx \sum l_j / \sum m_j = L/M$ ratio is approximately maintained. Then the sector angles for each elementary arc $\Delta \gamma_j = \gamma_j/m_j$ and the corresponding lengths of the elementary arc FEs as $\Delta l_j = l_j/m_j$ are calculated. So, chord length of elementary arc

$$\Delta l_{ch,j} = 2R_c \sin \Delta \gamma_j / 2. \quad (12)$$

Next, we calculate the angles of the direction of the normal $n_{j,i}$ to the elementary arc at the midpoint $\alpha_{j,i}$ and the

slope angles of the elementary arc chord to the X -axis $\beta_{j,i}$ in a similar manner, as in (6) – (9). Then the coordinates of the node points $(x_{j,i}, y_{j,i})$ of the partition of the j -th arc into FEs and collocation points $(\xi_{j,i}, \eta_{j,i})$ are calculated as

$$\begin{aligned} \text{if } R_c < 0 \Rightarrow & \begin{cases} x_{j,i} = x_{cc,j} + |R_{c,j}| \cos(\alpha_{ip,j} + i\Delta\gamma_j) \\ y_{j,i} = y_{cc,j} + |R_{c,j}| \sin(\alpha_{ip,j} + i\Delta\gamma_j) \end{cases}, \\ \text{if } R_c > 0 \Rightarrow & \begin{cases} x_{j,i} = x_{cc,j} + |R_{c,j}| \cos(\alpha_{ip,j} - i\Delta\gamma_j) \\ y_{j,i} = y_{cc,j} + |R_{c,j}| \sin(\alpha_{ip,j} - i\Delta\gamma_j) \end{cases}, \\ & \begin{cases} \xi_{j,i} = x_{cc,j} + |R_{c,j}| \cos \alpha_{j,i} \\ \eta_{j,i} = y_{cc,j} + |R_{c,j}| \sin \alpha_{j,i} \end{cases}. \end{aligned} \quad (13) \quad (14)$$

In addition, at the end of the geometric calculations, the relation $\Sigma l_{j,i} / \Sigma m_j = L/M$ is checked, which shows the accuracy of determining the arc length of the elementary part by the actual angle $\Delta\gamma_j$ and the radius of curvature R_c .

IV. MODELING THE MAIN STAGES OF THE ALGORITHM AND CHECKING THE ACCURACY OF THE RECEIVED RESULTS

The developed in Section III algorithm was translated into program code and modeled for the given values of the initial and end points of the arc, the radius of its curvature and the number of the partitions. Fig. 3, a shows the given curve and its partitions, which obtained by a computer program for designing drawings. Moreover, A(47.500000;-17.009619) – initial point (the beginning of the arc); B(27.009619;-22.500000) – end point (end of arc); $R_1 = 15.000000$ ($R_1 > 0$ – the center of curvature is on the right when passing the arc from the initial to the end point). The simulation results of the proposed approach application are presented in Fig. 3, b.

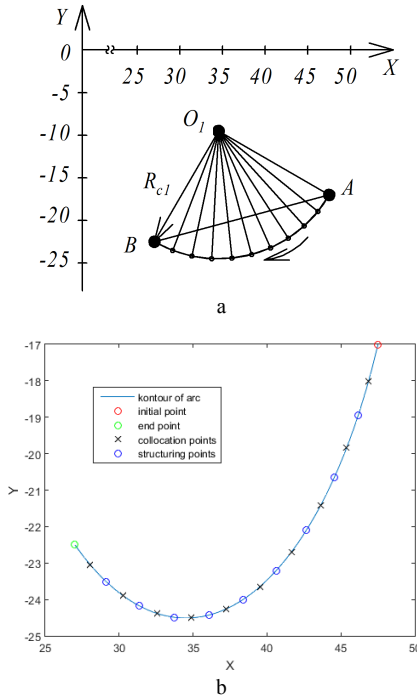


Fig. 3. Checking the proposed approach to determining the basic geometric parameters of one arc of the CF localization patch: a – a given arc, b – the results of computer simulation.

Fig. 3 shows good performance of the algorithm for solving the assigned tasks and its suitability for CF computing and control by embedded tools. The accuracy analysis of the one arc individual geometric parameters determination of the CF localization patch shows the high correspondence of the obtained results with the given values. For example, the given value of the parameter α is -1,30899693899575, and it turned out that the calculations -1,30899694088797; for α_{ip} -0,523598775598299 and -0,523598780767938; for $\beta_{1,1}$ -2,17293491873294 and -2,17293492357484 respectively (the worst cases of the obtained results are indicated). So, the computing error is less than 10^{-5} per cent.

V. CONCLUSIONS

The main features of electromagnetic interaction between the clamping electromagnetic driver and FS interaction are considered for robotic, industry and services applications. The main attention is paid to computing of geometrical parameters of CF localization patches by proposed combination-generation method as finite element method better modification.

Simulation results of computing of CF localization patches geometrical parameters by proposed combined method evidently show its high accuracy, which proves the adequacy of the developed algorithm for computing the main geometrical parameters of force localization patches. So, the method can be successfully used for practical implementation of the sensors system based on Hall transducers' data stream computing at CFs calculation [5]. However, developed in [5] CF control system strongly requires data stream processing from the sensors for indirect CF determination by field theory means. So, proposed in present paper method is practically applicable for integration equations solving in sources interaction method and another applications though some refinements in the way of computational complexity reduction yet may be accomplished.

REFERENCES

- [1] Y.G. Kozyrev, Industrial Robots, Reference book, 2nd edition, revised and enlarged. Moscow: Publisher Mashinostroenie, 1988.
- [2] C. Blanes, M. Mellado, and P. Beltran, "Novel additive manufacturing pneumatic actuators and mechanisms for food handling grippers," in Actuators, vol. 3, pp. 205–225, 2014. DOI:10.3390/act3030205
- [3] M. Taranov, J. Rudolph, C. Wolf, Y. Kondratenko, and O. Gerasin, "Advanced approaches to reduce number of actors in a magnetically-operated wheel-mover of a mobile robot," 13th Int. Conf. Perspective Technologies and Methods in MEMS Design (MEMSTECH), Polyana, Ukraine, pp. 96–100, April 20 – 23, 2017.
- [4] B. Ross, J. Bares, and C. Fromme "A semi-autonomous robot for stripping paint from large vessels," in International Journal of Robotics Research, pp. 617–626, July-August, 2008.
- [5] Y. Kondratenko, Y. Zaporozhets, J. Rudolph, O. Gerasin, A. Topalov, and O. Kozlov, "Modeling of clamping magnets interaction with ferromagnetic surface for wheel mobile robots," in International Journal of Computing, vol. 17, iss. 1, pp. 33–46, 2018. <http://www.computingonline.net/computing/article/view/947/812>
- [6] D. Souto, A. Faiña, A. Deibe, F. Lopez-Peña, and R. J. Duro, "A Robot for the Unsupervised Grit-Blasting of Ship Hulls," in International Journal of Advanced Robotic Systems, vol. 9, pp. 1–16, 2012.
- [7] D. Longo and G. Muscato, "A small low-cost low-weight inspection robot with passive-type locomotion," in Integrated Computer-Aided Engineering, vol. 11, pp. 339–348, 2004.
- [8] L. Christensen, N. Fischer, S. Kroffke, J. Lemburg, and R. Ahlers, "Cost-effective autonomous robots for ballast water tank inspection,"

- in *J. of Ship Production and Design*, August, vol. 27, no. 3, pp. 127–136, 2011.
- [9] D. Souto, A. Faiña, F. Lypez-Peca, and R. J. Duro, “Lappa: a new type of robot for underwater non-magnetic and complex hull cleaning,” *IEEE Int. Conf. on Robotics and Automation (ICRA)*, Karlsruhe, pp. 3394–3399, 6–10 May, 2013.
- [10] A. Ya. Vernikov, *Magnetic and Electromagnetic Devices in Metalworking*. Moscow: Publisher Mashinostroenie, 1984. (in Russian)
- [11] B. K. Bul, O. B. Bul, V. A. Azanov, and V. N. Shoffa, *Electromechanical Apparatus of Automation: Proc. for universities on spec. “Electrical apparatus”*. Moscow: Publisher Vysshaya shkola, 1988.
- [12] Y. Kondratenko, O. Gerasin, and A. Topalov, “A simulation model for robot's slip displacement sensors,” in *International Journal of Computing*, vol. 15, Issue 4, pp. 224–236, 2016. <http://www.computingonline.net/computing/article/viewFile/854/768>
- [13] Y. P. Kondratenko, Y. M. Zaporozhets, J. Rudolph, O. S. Gerasin, A. M. Topalov, and O. V. Kozlov, “Features of clamping electromagnets using in wheel mobile robots and modeling of their interaction with ferromagnetic plate,” 9th *IEEE Int. Conf. on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications (IDAACS)*, Bucharest, Romania, Vol. 1, pp. 453–458, 21–23 September, 2017.
- [14] Y. P. Kondratenko, O. S. Gerasin, and A. M. Topalov, “Modern sensing systems of intelligent robots based on multi-component slip displacement sensors,” *IEEE 8th Int. Conf. on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications (IDAACS)*, Warsaw, Poland, vol. 2, pp. 902–907, September 24 – 26, 2015.
- [15] Y. Kondratenko, A. Topalov, and O. Gerasin, “Analysis and modeling of the slip signals’ registration processes based on sensors with multicomponent sensing elements”, 13th *Int. Conf. CADSM 2015*, Lviv: Publisher National university “Lviv politehnika”, pp. 109–112, 2015.
- [16] Y. M. Zaporozhets, Y. P. Kondratenko, and O. S. Shyshkin, “Mathematical model of slip displacement sensor with registration of transversal constituents of magnetic field of sensing element”, in *Technical Electrodynamics*, no. 4, pp. 67–72, 2012.
- [17] L. A. Neyman, and V. Yu. Neyman, “Conductivities method application for accounting asymmetrical electromagnet single-side magnetic attraction force,” in *Herald of IrGTU*, no. 2 (97), pp.214–217, 2015.
- [18] E. T. Markov, *Ship Electric Apparatus*, 2nd ed. Leningrad: Publisher Shipbuilding, 1981.
- [19] B. I. Ogorelkov, A. S. Tatevosyan, U. V. Pimonova, and D. A. Polyakov, “Experimental research and mathematical modeling of dynamics of the electromagnet of the direct current,” in *Electrical and Data Processing Facilities and Systems*, vol. 11, no. 1, pp. 5–14, 2015.
- [20] O. A. Cherkasova, *Research of the magnetic field of the permanent magnet by means of computer modeling*. [Online]. Available: <https://goo.gl/biytj1>.
- [21] K. M. Polivanov, *Theoretical Foundations of Electrical Engineering*. Moscow: Publisher Energiya, 1974.
- [22] I. E. Tamm, *Foundations of the Theory of Electricity*. Moscow: Publisher Nauka, 1976.
- [23] K. Bins, P. Laursen, *Analysis and Calculation of Electric and Magnetic Fields*, Trans. from English, Moscow: Publisher Energiya, 1970.
- [24] Y. P. Kondratenko, J. Rudolph, O. V. Kozlov, Y. M. Zaporozhets, and O. S. Gerasin, “Neuro-fuzzy observers of clamping force for magnetically operated movers of mobile robots,” in *Technical Electrodynamics*, no. 5, pp. 53–61, 2017.
- [25] Y. P. Kondratenko, O. V. Kozlov, O. S. Gerasin, and Y. M. Zaporozhets, “Synthesis and research of neuro-fuzzy observer of clamping force for mobile robot automatic control system,” *IEEE First International Conference on Data Stream Mining & Processing (DSMP)*, Lviv, Ukraine, pp. 90–95, August 23 – 27, 2016.
- [26] M. Pasiaka, N. Grzesik, and K. Kuźma, “Simulation modeling of fuzzy logic controller for aircraft engines,” in *International Journal of Computing*, vol. 16, Issue 1, pp. 27–33, 2017. <http://computingonline.net/computing/article/view/868>
- [27] S. K. Oh and W. Pedrycz, “The design of hybrid fuzzy controllers based on genetic algorithms and estimation techniques,” in *Journal Kybernetes*, vol. 31, no. 6, pp. 909–917, 2002.
- [28] Q. Suna, R. Li and P. Zhang, “Stable and optimal adaptive fuzzy control of complex systems using fuzzy dynamic model,” in *Journal Fuzzy Sets and Systems*, vol. 133, pp. 1–17, 2003.
- [29] Y. Kondratenko, O. Korobko, O. Kozlov, O. Gerasin, and A. Topalov, “PLC based system for remote liquids level control with radar sensor,” *IEEE 8th Int. Conf. on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications (IDAACS)*, Warsaw, Poland, vol. 1, pp. 47–52, Sept. 24 – 26, 2015.
- [30] S. V. Izmaylov, *The Course of Electrodynamics, Textbook for physical and mathematical faculties of teacher training institutes*. Moscow: Publisher of Gosudarstvennoe uchebno-pedagogicheskoe izdatelstvo ministerstva prosvescheniya RSFSR, 1962.
- [31] Y. M. Zaporozhets, I. I. Chudaykin, and E. G. Amirshadov, “Combination-generation algorithm triangulation of the ship's surface in deviation tasks,” *Proc. of the Jubilee Scientific and Technical Conf. dedicated to the 50th anniversary of the ship protection service*, St.-Petersburg, 1994.
- [32] A. N. Tihonov, and A. A. Samarskiy, *Equations of Mathematical Physics, Textbook for high schools*, 5th ed. Moscow: Publisher Nauka, 1977.
- [33] P. Benerdzhii, and R. Batterfield, *Boundary Element Methods in Applied Sciences*, Trans. from English. Moscow: Publisher Mir, 1984.
- [34] P. Silvester, R. Ferrari, *The Finite Element Method for Radio Engineers and Electrical Engineers: Trans. from English*, Moscow: Publisher Mir, 1986. (in Russian)
- [35] N. S. Piskunov, *Differential and Integral Calculus for Higher Technical Schools*, Vol.1: Textbook for technical universities, 13-th ed. Moscow: Publisher Nauka, Glavnaya redaktsiya fiziko-matematicheskoy literaturyi, 1985.

Models of IT Projects KPIs and Metrics

Volodymyr Ostakhov
Information Systems Development
Department
Alfa-Bank Ukraine
Kyiv, Ukraine
vladimir.ostakhov@gmail.com

Nadiia Artykulna
Information Systems Development
Department
Alfa-Bank Ukraine
Kyiv, Ukraine
nadiia.artikulna@gmail.com

Viktor Morozov
Management Technology Department
Taras Shevchenko National University
Kyiv, Ukraine
knumvv@gmail.com

Abstract — models of key performance indicators (KPI) and methods for their measurement based on metrics for IT projects are proposed for consideration. Advantages and disadvantages of different KPIs and their applicability for various projects, as well as project implementation methodologies, are analyzed. A comprehensive KPI and metrics model is proposed for IT project team.

Keywords — key performance indicator, metric, IT project, project team, effectiveness.

I. INTRODUCTION

In any area implementation of a project involves engagement of several parties, at least two - business as a Customer and IT as a performer, as a maximum - many Customers and many executors, not only IT but also other service units or external companies. The goal of any project is either development of a new product or service or process [1]. As a rule, at the stage of project initiation, not only the project objective is determined, but also its deliverables. Less frequently, financial KPIs are declared as a result of project implementation in the context of the payback period [2]. From the point of view of PMO or other monitoring unit that monitors and controls the projects' implementation in the organization, at the time of project approval the time, budget and scope of the project are generally understandable and determined, including obligations for implementing financial model of the project (ROI, IRR, CBR). These are inputs for the project team, which is responsible for the project implementation [3]. As the project team consists of representatives of different parties of the initiators of the project, different groups of participants responsible for different stages of the project have different KPIs. This state of things in the project provokes imbalance of efforts aimed at implementing the project and violation of its integrity.

In the context of achieving the goals and deliverables of the project, project manager, the Customer and the Sponsor are interested more, while the project team is interested much less. This is due to the fact that at a particular time the project participants are interested in the successful completion of their work package and transfer of results to the next stage. Due to the limited budget and/or timing and/or technology, each group at a certain stage implements the maximum in its area of work, not caring about the limitations of the subsequent stage [4].

Ultimately, connecting different stages of the project, each of the groups is subsequently forced to find a compromise solution that will allow them to implement their work area. For example, the Customer prepares requirements that most closely correspond to the objectives and deliverables of the project, usually not taking into

account the fact that only 20% of the functional result in 80% of the benefit. In turn, analysts based on the inputs formalize requirements, turning them into an ideal product or process, implementation of which can become inconsistent with financial metrics [5]. The next phase of development faces the problems on the level of resources, technologies and timing, which as a result leads to scope reduction or simplification of the product, service or process. To ensure the integrity of the product, service or process, project manager has to find a compromise between the results of all phases for the successful completion of the project - between "What?", "How?", "When?" and "How much money?". In any case, such a compromise is a flawed option because it represents only one of the variations in the course of the project. In this example, Customers and analysts wasted time on unnecessary details and formalization of the product, service or process, and developers spent resources on unnecessary functionality [6].

In the search for and implementation of such compromise options, the interests of the parties which will subsequently sell or use the product, service or process, are lost out of focus.

In the authors' opinion, this problematics can be solved by implementing project KPIs for the project team from initiation to completion, ensuring the integrity of the results at each stage of the project.

The subject of the article is the research and analysis of the development and implementation of KPIs to ensure a holistic approach to the project implementation, excluding implementation of a compromise option as the most unprofitable for the organization as a whole.

II. ANALYSIS OF RESEARCH AND PUBLICATIONS

The questions of applying different approaches to models of IT projects KPIs and metrics were considered in the works of Harold Kerzner [5], Rad, Parviz F., Ginger Levin [6], Daniel S. Vacanti [7] and other scientists.

The problems of using project metrics and KPIs for different projects were reflected in the works of following Ukrainian scientists: Yu.Teslia [8], I.Kononenko [9], V.Burkov, S.Bushuyev [10], V.Morozov [11], and others. In particular, the issues of the need to measure the key project indicators for various areas as well as the implementation of the KPIs of the project portfolio were highlighted [12]. However, using KPIs and metrics for the project team in order to achieve the maximum result has not been studied in depth to be practically solved.

Unsolved parts of the common problem. A large number of studies have been carried out in the field of innovative

projects management [13]. There are outstanding examples of successful completion of projects of scientific nature, but the peculiarity of using KPIs and the metric system for their regular measurement for the project team makes its adjustments to the classical project management. It is important to determine the main features of KPIs and metrics for their measurement in order to determine the most comprehensive and effective approach to project management to focus the project team's efforts on the project's outcome.

The purpose of the article is to study and analyze the various KPIs of the project team and metrics for regular measurement of the intermediate state of the project in order to achieve the maximum result due to effective use of resources and budget in the planned time frame.

III. MODELS OF IT PROJECTS KPIS AND METRICS

A KPI is a metric measuring how well the organization or an individual performs and operational, tactical or strategic activity that is critical for the current and future success of the organization. The value of the KPI must be well understood in order for it to be used correctly and for it to provide the necessary information for informed decision making. The project manager and the appropriate stakeholders must come to an agreement on which metrics to be used and how measurements will be made. There must also be agreement on which metrics will be part of the dashboard reporting system and how the metric measurement will be interpreted. When coming to projects, metrics represent the measurement in time of specific numerical indicators according to the approved KPIs. The value of metrics is the ability and indeed the need of measuring them throughout the implementation of the project but not at the completion stage or after the implementation of the project, to take timely corrective actions [14]. Values of metrics on the key date of measurement should be an integral part of the status report for informing both stakeholders and members of the project team. It is important to remember that metrics are measurements and, therefore, provide project managers with opportunities for continuous improvements to the project management process. Selecting metrics without considering a plan for future action is a waste of time and money. If a measurement indicates that the metric is significantly far away from the target, then the team must investigate the root cause of the deviation, determine what can be done to correct the deviation, get the plan to correct the deviation approved, and then implement the new plan. Metrics also allow project managers to create a database of historical information from which to analyze trends and improve future estimating.

The authors consider the following classes of metrics for IT projects.

Metrics with full project duration measurements - metrics, such as cost and schedule variances, that are used for the entire duration of the project and measured either weekly or monthly.

An example of such type of KPI is a quality of planning KPI, which is the basis for measuring the compliance of the planned project completion date or phase in comparison with the actual one. This KPI is a complex one and is designed for the straight-line and consistent execution of project work, which in turn minimizes the use of overtime,

uneven load on resources (both human and budget) in order to achieve planned deadlines when significant deviations are identified at the completion stages.

To measure quality of planning KPI, authors encourage to use the following metrics:

- measurement of the timely completion of the project's milestones and the timely commencement of the next block of work. Project milestones must be defined every 10-15 calendar days;

- measurement of manpower effort - the ratio of planned and actual labor costs to the key date of measurement:

$$KPI_{QoP} = \frac{\text{LabourCosts}_{planned}}{\text{LabourCosts}_{actual}} \quad (1)$$

-earned value management:

Another example of this type of KPI is the KPI of architecture control. The context of this KPI is the correspondence of business and technical implementation of the project to the development strategy of the organization [15]. To measure this KPI the authors propose to use the questionnaire approach:

- positive conclusion of enterprise-architect on the compliance with the target architecture of the IT landscape of the organization;

- positive conclusion of solution-architect on the compliance with the target architecture of the system (platform);

- positive conclusion of IT Architectural Committee on the implementation of a new system (platform) into IT landscape of the organization;

- positive conclusion of business owner of the system or process on the compliance with the business strategy of the organization.

The value of this KPI lies in the early detection of the inconsistency of the implemented project solution with the development strategy of the organization, as a result of which - exclusion of increased cost of solution due to the need of implementation of significant changes, in some cases - decommissioning of such a solution in short or medium term perspective. Serious consequences of this kind of inconsistencies can be significant investments in IT infrastructure to maintain the performance of the project solution.

It is important to take into account that the only exception can be those projects that were initially aimed at realizing short-term goals of the organization without long-term plans for using results of the project, which should be formally confirmed by the management of the organization.

Metrics with life cycle phase measurements - metrics that exist only during a particular life cycle phase. As an example, metrics that track the amount or percentage of direct labor costs used for project planning would probably be measured just in the project planning phase [5].

The basic KPI of this class for the project team can be an event KPI, reflecting the value expressed in money for each functionality of the product, service or process. The purpose of this KPI is to focus the project team on the implementation of the most significant parts of the

functional, providing the maximum value. Ideally, 20% of the functionality provide 80% of the value of the project solution. Essentially, this KPI is a guide for the project team, reflecting where significant part of the efforts of both the Customer and the project team should be spent.

The value KPI - event KPI, reflecting the value expressed in money for each functionality of the product, service or process - answers the question with “yes” or “no”. That is, is the project team provided with an assessment of the functionality of the solution being developed (product, service or process). It is important to note that in case of the appearance of new functionalities, they should be estimated and added to the primary evaluation of the scope. Otherwise, the KPI value is zero. The value of this KPI lies in focusing the project team and organization on the implementation of the core of the product, service or process, both in terms of business orientation and IT solutions. As a result of this balancing, the investment policy of both human resources and budgetary funds will be more effective for the organization as a whole, while respecting the principle of maximizing profits at minimum costs [16]. The next example is the KPI of development quality. The value of this KPI is manifested in the reduction of costs of the project solution due to the early identification of defects' sources of origin and their elimination at the initial stages of design and implementation of the solution. The authors suggest the following approach to measurement: the ratio of manpower effort of eliminating defects that occur at different stages of the project (analysis, development, testing and productive operation during the pilot operation) to the overall manpower effort of the project.

$$KPI_{DQ} = \frac{\text{ManpowerEffort}_{\text{defects}}}{\text{ManpowerEffort}_{\text{project}}} \quad (2)$$

It should be noted that this approach to monitoring and controlling of the project quality necessitates the following project works:

- testing the technical specification for compliance with the requirements of the Customer and the final beneficiary in terms of completeness and sufficiency;
- review of the program code before completion of the development of the functional for compliance with the standards of development in the organization;
- development of a test plan, test cases and creation of autotests based on requirements and technical specification before the start of development;
- development of a plan and test cases for user-acceptance testing before the start of the testing the technical specification.

Such an approach allows to ensure improvement of the quality of implemented solution by identifying defects in the "paper" project of solution before it is created on the level of program code, and accordingly allows to avoid increasing the cost of solution both from the point of view of investment and the timing of solution launch in pilot production, as well as negative client experience. Depending on the stage of defect detection, the cost of its elimination varies - the later the defect is identified, the more expensive its elimination is [6].

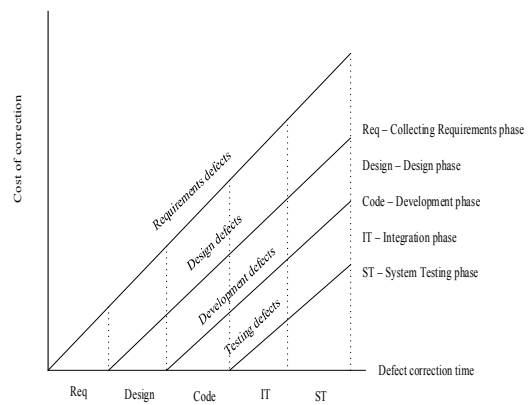


Fig. 1. Increase of cost of a defect in project timeline

A more specific representative of this KPI type is a KPI of quality of technical specification and is measured as % coverage of business requirements by technical specifications. For 20% of the functionality that ideally provides 80% of value of the product, service or process, the KPI value should be 100% before the development starts. Otherwise, eliminating the design defect of core of the product, service or process will be more expensive and longer with each subsequent stage on which the defect is detected. Another narrowly focused example of KPI of this class is the KPI of quality of testing and is measured as % coverage of the functionality by test cases. Similar to the previous one, the value of this KPI is in identifying and eliminating defects of a product, service or process at the stages preceding acceptance testing and pilot production. This approach ensures the effective use of investments aimed at development of an initially correct product, service or process, in contrast to investing budget funds in eliminating defects, up to rejecting the solution in consequence of financial inexpediency of its reworking. The coverage of test cases with a significant 20% of functionality should be provided at 100% while variation in coverage of test cases and/or autotests of 80% of functionality that generates only 20% of the product value may be allowed, sometimes leading to a technical debt.

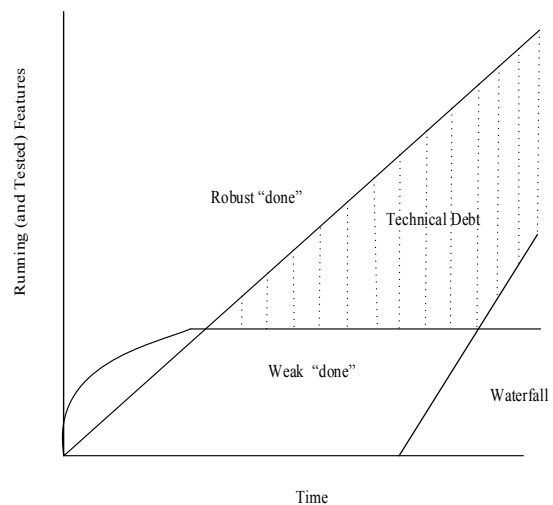


Fig. 2. Testing product features within time

It is important to note that all KPIs offered by the authors for consideration are based on the availability of the underlying value KPI, without the definition of which all the subsequent KPIs have no value for either the stakeholders or the project team.

Metrics with limited life measurements - metrics that exist for the life of an element of work or work package. As an example, we could track the manpower staffing rate for specific work packages or the number of deliverables produced in a specific month. This type of KPI, and accordingly metrics, is most applicable for projects using agile development methodologies. Examples of such KPIs can be the measurement of different parameters of agile projects - velocity of the project team, time to market of project deliverables, and the level of technology.

Basic Agile Metrics [7]:

- Agile methods are based on traditional measures
- Size, efforts, and velocity metrics are common
- Top-notch shops use complexity and testing metrics

TABLE I. AGILE METRICS

Type	Example
Size	Story, Story point, Task, Function Point, LOC, etc.
Effort	Ideal or Actual Hours, Days, Weeks, Months, Years, etc.
Velocity	Story, Story point, Function Point, or LOC per iteration/Sprint
Complexity	McCabe, Halstead, Object-Oriented, Relational Database, etc.
Quality	Defect Density, Defect Removal Efficiency, Rayleigh, etc.
Testing	Tests Passed/Failed/Broken, Running Tested Features, etc.
Reliability	Mean Time to Failure, Mean Time between Failure, etc.

Measurement of these indicators of the project team is mandatory, and the feature of this class of metrics is the continuous improvement of the values of these indicators. Special attention should be paid to the adaptation of value KPI, characteristic for projects with classical approaches to project management, to the project backlog estimated by product owner from the point of view of financial benefits for ensuring correct priorities for the implementation of the product, service or process [17]. It is impossible to create and control quality of planning KPI without a regular measurement of the project team's velocity and the product owner's estimated backlog of the project, as well as the initial top-level assessment of the project's backlog by the project team in terms of labor costs for implementation. In the Agile philosophy it is impossible to form a burn-down of a project or sprint without the values of these indicators, and accordingly to provide monitoring and controlling of any project KPI.

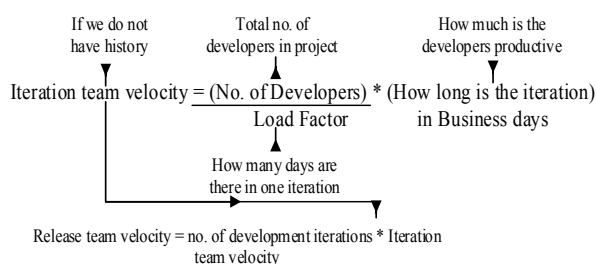


Fig. 3. Agile project indicators

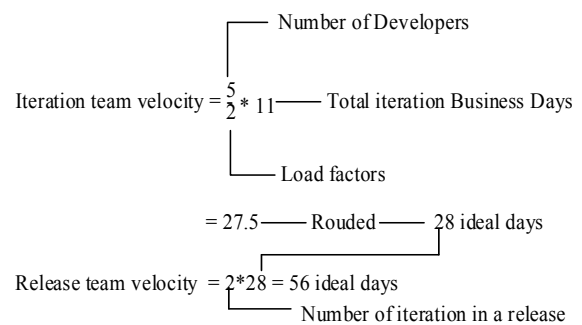


Fig. 4. Agile project measurements

The complex of described categories, and accordingly their regular measurement, provides the right product, its correct implementation and the correct process of implementation of the product, service or process, while ensuring the efficiency of investments and the reduction of effort to implement the project in a shorter period. This is possible only if the technological level of the project team is increased due to implementation of such engineering practices as auto-testing, auto-deployment, continuous integration and the construction of DevOps processes [19].

For projects using the Agile approach, most of the KPIs of quality described in the other KPI classes of this article are mandatory and initially focus both the Customer and the project team on their regular measurement and improvement, which significantly increases the quality of the product, service or process being developed, significantly speed up the start-up period for pilot production, and also force the project team to regularly improve the KPI process.

Metrics that use rolling-wave or moving-window measurements - metrics where the starting and finishing measurement dates can change as the project progresses. As an example, calculations for the cost performance index and schedule performance index are used to measure trends for forecasting. On long-term projects, a moving window of the most recent six data points (monthly measurements) may be used to obtain a linear curve for the trend line [5].

As an example of KPI of this class, authors propose to consider Customer's quality of requirements KPI. The essence and value of measurement of this KPI is shown in the sufficiency and readiness of the primary requirements to the product in terms of design and implementation to ensure the performance of financial indicators.

The measurement of this KPI is considered as the ratio of manpower effort for implementation of changes in requirements to overall manpower efforts for the project. Another indicator here is the shift in terms due to changes in requirements. Static deterioration of this KPI signals that the product is not ready for implementation and the need of taking immediate corrective actions, not excluding the suspense of the product, service or process development.

Preventive measures for this class of problems are the presence in the project plan of mandatory formal approvals with the Customer, Sponsor and representatives of end users of product requirements, functional and non-functional requirements, technical specifications of the product, service or process prior to the start of development, as well as early creation of plan and test cases in order to avoid detection of

critical inconsistencies at the user-acceptance testing stage. Otherwise, implementation of critical requirements for the product, service or process at the acceptance testing stage can lead not only to the need of significant additional financing, but also to the shutdown of the project due to financial inexpediency of launching such a product, service or process.

Another example of this class is the KPI of project risk identification - an event KPI reflecting the identification of all project risks at the planning stage. A mandatory condition for the positive value of this KPI is the availability of a response strategy for each identified risk. It should be noted that the qualitative aspect of this KPI, in addition, is the regular revision and actualization of project risks. However, it is important to take into account that the regular increase in the list of project risks with each subsequent stage, especially with the failure of Customer's quality of requirements KPI, may indicate serious problems on the project. As a consequence, the completion of the project is likely to require additional financial investment, which in turn will require an analysis of the financial feasibility of implementing the product, service or process [6].

An important KPI of this class, according to the authors, is the KPI for closing the project risks. This KPI is extremely important to monitor on a regular basis and in dynamics, since the value of KPI is in regular monitoring of the overall risk of the project.

From the point of view of measurement, KPI is proposed to be considered as % of risks taken place to the total number of identified risks, taking into account the impact on the financial component of the project in the form of additional financing and/or project losses.

The reverse side, a positive aspect, can be the identification of positive project risks that allow using them as a potential opportunity to reduce investment in the project or improve the quality or functionality of the product, service or process.

In general, working with project risks is a rather complex but important part of the implementation of any project. This topic is not the subject of this paper, but from the KPI perspective of the project team is considered to keep focus on regular work with project risks.

Alert metrics and measurements - metrics used to indicate that an out-of-tolerance condition exists. The metrics may exist just until the out-of-tolerance condition is corrected, but they may appear later on in the project if the situation appears again. Alert metrics could also be metrics that are used continuously but are highlighted differently when an out-of-tolerance condition exists.

From the authors' point of view, one of the essential KPIs for the project team is the KPI of the process, which shows how the project team follows the development process approved in the organization without deviations.

The value of measuring and tracking this KPI is the timely monitoring and controlling of the implementation processes of the organization in order to prevent skipping key development stages intended for the implementation of the right product, service or process and in the right way. In addition, it is important to note that following the process is one of the ways to ensure compliance with the principle of

maximizing profits while minimizing labor costs and financial investments of the organization.

KPI of the process is integral and demonstrates an integrated and holistic approach to measuring key project indicators that are necessary to properly focus the project team and maintain a balance of financial feasibility of implementing the product, service or process for the organization as a whole.

Considered classes of KPI with the corresponding metrics for their measurement, in the opinion of the authors, are sufficient to focus both the Customer and the project team on the successful implementation of the product, service or process.

Regular inclusion of these KPIs into the project's status report allows early identification of the inadequacy of requirements for a product, service or process, identification and elimination of product, service or process defects, and ensuring the required quality and functionality of the product, service or process within the project budget in target timelines.

IV. CONCLUSION

As a result of the study of the use of KPIs for the project team, various KPIs were analyzed, as well as metric classes for their regular measurement. For each stage of the project relevant KPIs are considered, the application of which maximizes the focus of the project team on achieving the final result of the project - implementation of the product, service or process.

It is important to consider the complexity and integrity of using KPIs for the project team, otherwise focusing only on one of them can lead to an emphasis on a particular stage or characteristic of the product, service or process to the detriment of others, which ultimately does not lead to the maximum effect.

Another important aspect of using KPIs is their regular measurement to track the dynamics of indicators, in other words - the state of the project at a particular point in time. The definition of the measurement system, metrics, for the project team is a key aspect for achieving KPI.

Determining KPIs for the project team will not be effective in measuring KPI values at the end of the project, since we will not be able to apply any corrective actions aimed at eliminating the terms of implementation, increase of the budget or improving the quality of the product, service or process. In order to timely influence and apply corrective actions to the project, metrics are used to periodically measure the project, and then develop and apply corrective actions, both from the side of the project team, the Customer and the project Sponsor.

The values of metrics that measure the intermediate state of achievement of KPIs on a regular basis should necessarily be included in the project status report to inform the project team, the Customer and the Sponsor, as well as corrective actions that will be applied to eliminate project problems.

Additional monitoring and controlling by using KPIs is also the implementation of a group of mandatory control works into the project's schedule. Examples may be the approval of requirements by the Customer, testing the

technical specification, development of a test plan including test cases for testing by the project team and end users.

The concept of an integrated approach to the use of KPI and a system of metrics for a project team is based on the principle of implementing the right product, service or process with the right architectural, technological solution in the planned timeframe and with defined project budget in favor of the client and/or end users.

REFERENCES

- [1] S. Berkun, *The art of managing IT projects*. SPB.: Piter, 2014.
- [2] G. Ellis, *Project Management in Product Development*, 1st Edition "Leadership Skills and Management Techniques to Deliver Great Products". Butterworth-Heinemann, 2015,
- [3] A. Ilarionov, and E. Klymenko, *Project portfolio: Instruments for company's strategic management*. M.: Alpina Publisher, 2013.
- [4] V. Morozov, O. Kalnichenko, and I. Liubyma, "Proactive Project Management for Development of Distributed Information Systems," 4th International Scientific and Practical Conference "Problems of Infocommunications. Science and Technology" (PIC S&T-2017), Kharkiv, Ukraine, pp.24-27, 10 - 13 October, 2017.
- [5] H. Kerzner, *Project Management Metrics, KPIs, and Dashboards: A Guide to Measuring and Monitoring Project Performance*, 3rd Edition. Wiley, 2017.
- [6] P. F. Rad, and Ginger Levin, *Metrics for project management : formalized approaches*. Vienna, VA : Management Concepts, 2006.
- [7] D. S. Vacanti, *Actionable agile metrics for predictability: an introduction*. Leanpub, 2015.
- [8] Yu. Teslia, A. Khlevnyi, and I. Khlevna, "Control of informational Impacts on project management", IEEE 1th International Conference on Data Stream Mining & Processing, Lviv, Ukraine, pp. 387-392, 23-27 August, 2016.
- [9] I. Kononenko, and S. Lucenko, "Method of choosing approach for managing project based on indistinct performance", *Visnyk NTU «HPI». Series: Strategic management, portfolio management, program management*, no. 2. pp. 8–17, 2014.
- [10] V. Burkov, S. Bushuyev, and A. Voznyj, *Managing resources of distributed projects and programs*. N: 2015.
- [11] V. Morozov, and A. Cherednichenko, *Managing projects: processes of planning project actions*. Kyiv: KROK, University of Economics and Law, 2014.
- [12] V. V. Ostakhov, V. V. Morozov, and N. F. Artikulnaya, *Method of project portfolio optimization based on metrics in conditions of IT transformation*. Series: Strategic management, portfolio management, program and project management, Kh.: NTU «KhPI», 2018.
- [13] S. Bushuyev, F. Yaroshenko, and Kh. Tanaka, *Management of innovative project and programs based on P2M system of knowledge*. K.: «Sammit-Book», 2012.
- [14] D. Cleland, *Global Project Management Handbook: Planning, Organizing, and Controlling International Projects*. McGraw-Hill Education, 2010.
- [15] F. Lefley and J. Sarkis, "Applying the FAP model to the evaluation of strategic information technology projects". *International Journal of Enterprise Information Systems*, 1, 2005, pp. 69–90.
- [16] I. Chumachenko, V. Morozov, *The Project Management: Process of Planning of Project Activities*, Handbook, K.: University of Economics and Law "Krok", 2014. – 673 p.
- [17] A. Biloshchytskyi, A. Kuchansky, Yu. Andrashko, S. Biloshchytska, "The Method of the Scientific Directions Potential Forecasting in Infocommunication Systems of an Assessment of the Research Activity Results". *Proceedings of the 4th International Scientific and Practical Conference "Problems of Infocommunications. Science and Technology" (PIC S&T-2017)*, 10 - 13 October, 2017, Kharkiv, Ukraine, pp. 70-73.
- [18] A. Alpatov, "Development of distributed technologies and systems," *Prospects for Science and Education*, 2015, vol. 2 (14).

Mobile Application for Decision Support in Multi-Criteria Problems

Yuliya Kozina
Applied of Mathematics Department
Odessa National Polytechnic University
Odessa, Ukraine
yuliyakc21@gmail.com

Natalya Volkova
Applied of Mathematics Department
Odessa National Polytechnic University
Odessa, Ukraine
volkovanp30@gmail.com

Daniil Horpenko
Applied of Mathematics Department
Odessa National Polytechnic University
Odessa, Ukraine
dieznote@gmail.com

Abstract—In this paper the issues of creating Decision Support System was considered. The classification of DSS at the user level is given. The importance of creating mobile DSS was shown. Known methods for evaluating and comparing multicriteria alternatives AHP, MAHP, Topsis was considered, also. The modified algorithm of the heuristic method Smart was proposed. A comparison of the proposed the Smart method modification with known methods, concluded that using the modified method Smart and method Topsis in mobile DSS is expediential. The architecture and the realization of mobile DSS were described. Mobile DSS was realized on the Android platform and it works on the smartphones and the tablets, which allows decision-makers to be mobile in off-line mode.

Keywords—Decision Support System; Topsis; Smart; AHP, Android; the alternative ; the criteria

I. INTRODUCTION

In Decision theory, the variants of solutions (alternatives) are characterized by the different indicators of their attractiveness for decision maker, which are the criteria for choosing a solution. In the most tasks, the solution must be evaluated from the different points of view, such as physical, economic, technical and others. The number of criteria also influences at the complexity of finding solutions to decision-making problems. If the number of criteria is increased, the task of comparing alternatives becomes not so obvious for decision maker, in that case used Decision Support System (DSS) [1].

DSS is an interactive automated system that helps decision makers to use the data and models to solve unstructured and ill-structured problems.

At the user level, DSS are divided into two groups: Enterprise-wide DSS and Desktop DSS. Enterprise-wide DSS is a multi-user DSS that runs on a server machine. Desktop DSS is a single-user DSS that runs on a personal computer. The development of web - technologies has led to the creation of web - based decision support systems (WB-DSS) [2], which are available to the Internet. But in the conditions of the mobile devices development and the growth of the information technology the creation of Mobile Decision Support System (mDSS) [3], which allows decision-makers to be mobile in off-line mode, becomes necessary.

There are many DSS that are based on known methods for evaluating and comparing multi-criteria alternatives, such as MAUT (Multi-Attribute Utility Theory), AHP (Analytic Hierarchy Process), Electre (Elimination Et Choix Traduisant

la Realite), Topsis. The most well-known and popular academic and commercial systems that implement methods for evaluating multi-criteria alternatives are listed and described in [4].

It is noted that a large part of the academic systems are used for the narrow class of tasks or not supported by the developers. Commercial systems are not always available to a wide range of users. For many desktop systems are being created web versions that, unlike desktop systems, are accessible to the global audience, easy to use, but there are security issues.

II. PROBLEM DESCRIPTION

WB-DSS allows: simultaneously use DSS by a large number of decision makers; to reduce the cost of installing the application; to reduce maintenance and system upgrades. But access to WB-DSS and the speed of processing information in such systems depends on access to the Internet. It limits the mobility of the decision maker. These limitations can be eliminated by creating mDSS.

Researches show that mobile technology is 22% of the total sales of e-commerce retailers and wholesalers in the U.S., and could rise to 27% in 2018. The main reason for the growth of mobile markets is the growth sales of tablets, smartphones and other mobile devices. The mobile technology for business information access comes instead the desktop computer solutions, thereby opening up the direction of Mobile Intelligence.

mDSS is easy to use, have a user-friendly interface and do not require additional training, which allows for any decision maker, to use such a system regardless of the qualifications level. mDSS allow decision maker to work offline that increased the speed of decision-making. The data is either on the mobile device, or decision maker enters them while working with the system and does not require additional download. In addition, using mDSS does not require monthly payment. The process of developing a mobile application is more longer and costly than the web application, but as a result, we obtain a convenient and functional system can be used by decision makers to help with decision-making at any time. During the creation of any mobile application, the main task is to design the interaction between the user and the system. The process of exchanging information between the user and the system should be convenient. The methods on which mobile applications are based should be not resource -intensive.

Thus, the aim of the work is to develop a mDSS on the Android platform, which based on the multi-criteria decision making method Topsis and the modified algorithm of heuristic method Smart, to simplify access to DSS for obtain the multi-criteria tasks solutions.

III. SOLVING PROBLEM

The unstructured and ill-structured problems, which are multicriteria, are solved by decision-making methods. Let us consider the problem of multicriteria optimization – the set of n alternatives and N criteria, which used to evaluate alternatives, is given. Each alternative is evaluated by each criterion by experts, or on the basis of the objective calculations. It is necessary to build a decisive rule based on the preferences of decision maker. This rule allows: to select the best alternative, to rate alternatives in quality, to assign alternatives to orderly classes solutions in quality.

Consider three well-known methods to evaluate and compare the multicriteria alternatives – AHP, MAHP, Topsis. In the AHP method for a given set of alternatives, the coefficients of the importance of the criteria, the goals (if there are many), and alternatives (for each criterion) are determine by pairwise comparisons. The overall score of alternative is as the sum of the product of the weights [5].

The Multiplicative AHP [6] is a modified method of the AHP proposed by F.A. Lootsma. It is based on two basic provisions. First, if the design maker determines the relationship (and not the absolute value) of two elements of the corresponding level of the hierarchy, then it is more logical to multiply such relationships than to addition of the values obtained from the comparisons. Secondly, the transition from the quality to quantity comparisons must happen on the basis of certain assumptions about human behavior when comparing measurements on the basis of certain assumptions.

The Topsis method consists in arranging alternatives by the degree of their proximity to the ideal positive and the distance from the ideal negative decision. It is based on the concept that the chosen alternative should have the smallest distance to the ideal solution and the largest one - to an ideally negative solution. The results of which are corrected on the basis of restrictions by criterion [7].

The heuristic method Smart (Simple Multi-Attribute Rating Technique) is based on a linear additive model. The design maker ranks criteria by importance and determines the criteria weights by using the 1 to 100 interval, and then they are normalized. The design maker evaluates each alternative for each criterion on by using the 0 to 100 interval. The overall score of each alternative is determined as the sum of the product of the normalized criteria weights by the measures each alternative for each criterion [8].

Let us consider the algorithm of the proposed Smart method modification:

Step 1: Rank the criteria by importance.

Step 2: Select optimality criterion - maximization or minimization.

Step 3: Determine the criteria weights of each criterion by using the 10-100 interval. The weight of the most important criterion is set to 100.

Step 4: Calculate the normalization of each criteria weights, using the formula

$$y_j = \frac{\omega_j}{\sum_{j=1}^n \omega_j} \quad j = \overline{1, n}, \quad (1)$$

where n – number of the criteria, ω_j – the value of criteria weight, y_j – normalized criteria weight, $\sum_{j=1}^n y_j = 1$.

Step 5: Construct the decision matrix $X (m \times n)$, where m – number of the alternatives, n – number of the criteria

$$X = \begin{matrix} & C_1 & C_2 & \dots & C_n \\ A_1 & x_{11} & x_{12} & \dots & x_{1n} \\ A_2 & x_{21} & x_{22} & \dots & x_{2n} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ A_m & x_{m1} & x_{m2} & \dots & x_{mn} \end{matrix}$$

where A_1, A_2, \dots, A_m are possible alternatives, C_1, C_2, \dots, C_n are criteria for which the alternative is measured.

Step 6: Calculate a normalized decision matrix, using the formula

- for criteria, which optimality criterion is maximization (Equation 2)

$$r_{ij} = \frac{x_{ij}}{\sqrt{\sum_{i=1}^m x_{ij}^2}} \quad (2)$$

- for criteria, which optimality criterion is minimization (Equation 3)

$$r_{ij} = 1 - \frac{x_{ij}}{\sqrt{\sum_{i=1}^m x_{ij}^2}} \quad (3)$$

Step 7: Determine the overall score of each alternative.

Overall score of each alternative is as the sum of the product of the according elements of each row of the normalized decision matrix on the normalized criterion weights (Equation 4).

$$V_i = \sum_{j=1}^n r_{ij} y_j \quad (4)$$

where V_i – overall score of the alternative i .

Step 8: Rank alternatives in the preference order.

IV. DESCRIPTION OF INFORMATION SYSTEM

The mobile information system was developed with Java language in Android Studio 3.0.1. The Topsis methods and the modified Smart method are implemented in this information system.

The class ChoseKolva.java – implements the user interface of the first desktop, namely:

- allows a user to choose a method by which the task will be solved;
- allows a user to enter the number of the criteria and the alternatives;
- allows a user to perform verification of the entered data.

The class Grid.java – implements the interface of the second user desktop, which allows a user to enter the names of the alternatives and the criteria and performs validation of the entered data. The class Thirdactivity.java – implements the interface of the third user desktop, namely:

- allows a user to enter the numerical estimates of the each alternative by the relevant criteria;
- allows a user to enter the criteria's evaluations;
- allows a user to enter optimality criterion (maximization/ minimization);
- implements Topsis method and modified Smart method.

The class FourthActivity.java – implements the interface of the fourth user desktop, which displays the result of the methods of decision making. The class help.java – implements the interface of the user desktop, which displays the reference information to user. Apply the proposed modified algorithm of the Smart method to several multicriterion optimization problems. Then, compare the results of the solution with the results obtained using the methods AHP, MAHP, Topsis.

Example 1. Consider the problem from [9], in which it is necessary to choose the best place for the construction of the airport. Four places are considered – A_1, A_2, A_3, A_4 . To evaluate the construction places, three main criteria were chosen – the cost of the constructing (C_1), the distance from the city (C_2), the minimal noise impact (C_3). The input data and the computation results are presented in the Table I.

TABLE I. THE INPUT DATA AND THE COMPUTATION RESULTS 1

$C_j \backslash A_i$	The input data			Selection of the alternatives		
	C_1	C_2	C_3	MAHP	Topsis	Modified Smart
A_1	180	70	10	0,005	0,4138	0,7520
A_2	170	40	15	0,072	0,2778	0,8371
A_3	160	55	20	0,155	0,6022	0,8151
A_4	150	50	25	0,432	0,6725	0,8691

Example 2. Consider the problem from [10], in which it is necessary to rank the consumers' preferences regarding the use of some typical appliances during peak hours when the price of electricity has increased. These appliances are dishwasher (A_1), home computer (A_2), hair dryer (A_3), iron (A_4), spa bath (A_5), television (A_6), vacuum cleaner (A_7). The criteria are: electrical cost (C_1), emergency (C_2), welfare

and comfort (C_3), joyness (C_4). The input data and the computation results are presented in the Table II.

TABLE II. THE INPUT DATA AND THE COMPUTATION RESULTS 2

$C_j \backslash A_i$	The input data				Selection of the alternatives		
	C_1	C_2	C_3	C_4	AHP	Topsis	Modified Smart
A_1	0,933	0,5	0,07	0,093	0,235	0,2261	1,281
A_2	0,2	1,0	0,03	0,4	0,183	0,8346	1,428
A_3	4,933	0,5	0,373	0,493	0,065	0,3937	0,657
A_4	0,933	0,05	0,07	0,093	0,049	0,1579	0,919
A_5	1,867	1,0	0,28	0,373	0,237	0,6123	1,336
A_6	1,467	0,3	0,066	0,088	0,112	0,2255	0,905
A_7	0,067	1,0	0,01	0,013	0,120	0,2766	1,072

Thus, the alternatives were ordered as follows:

AHP – $A_5 \succ A_1 \succ A_2 \succ A_7 \succ A_6 \succ A_3 \succ A_4$;

Topsis – $A_2 \succ A_5 \succ A_3 \succ A_7 \succ A_1 \succ A_6 \succ A_4$;

Modified Smart – $A_2 \succ A_5 \succ A_1 \succ A_7 \succ A_4 \succ A_6 \succ A_3$.

Thus, research has shown that all three methods gave two identical the consumers' preferences. The example user interface of the mobile application for solving the problem is shown in Fig. 1.

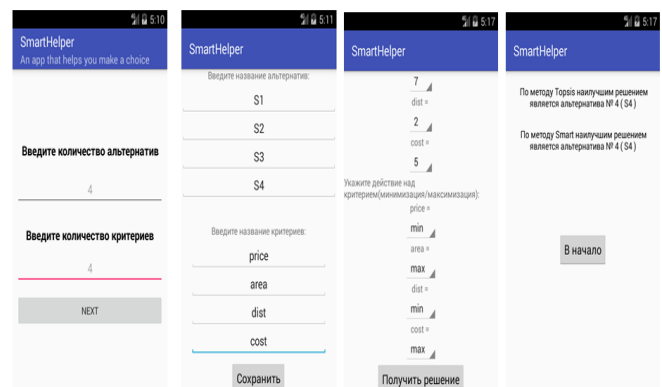


Fig. 1. The example user interface of the mobile application

Example 3. Consider the problem from [11], in which it is necessary to choose the best company for marketing research for a conditional enterprise. Eight companies were considered, such as: UMG, Marketing Lab, AIM, GFK Group, TNS, AC Nielsen, iVOX, MMG. We denote them by $A_i, (i = \overline{1,8})$. To evaluate the company the eight criteria were chosen: the customer satisfaction level (C_1), the quality and the complexity of techniques (C_2), the number of unique products that fit the company (C_3), work experience in Ukraine (C_4), the level of organizational support for marketing research (C_5), the reputation of the company (C_6), the efficiency of the research (C_7), the average cost of services in the company (C_8). The input data and the computation results are presented in the Table III.

TABLE III. THE INPUT DATA AND THE COMPUTATION RESULTS 3

$C_j \backslash A_i$	The input data								Selection of the alternatives		
	C_1	C_2	C_3	C_4	C_1	C_2	C_3	C_4	AHP	Topsis	Modified Smart
A_1	15	7	5	19	7	7	9	6300	0,0437	0,2988	0,5164
A_2	20	6	3	13	5	5	10	5300	0,1144	0,1935	0,4710
A_3	20	5	1	8	6	6	9	5100	0,1033	0,1056	0,4002
A_4	60	10	4	15	10	9	4	5100	0,1983	0,3592	0,9617
A_5	50	9	6	11	10	9	5	7100	0,1830	0,8383	0,9752
A_6	65	8	5	20	9	10	7	6900	0,1772	0,4002	1,0131
A_7	9	5	1	5	8	4	8	4900	0,0788	0,0917	0,2899
A_8	10	9	3	10	9	8	6	7000	0,1010	0,2100	0,4414

Thus, the alternatives were ordered as follows:

AHP – $A_4 \succ A_5 \succ A_6 \succ A_2 \succ A_3 \succ A_8 \succ A_7 \succ A_1$;

Topsis – $A_5 \succ A_6 \succ A_4 \succ A_1 \succ A_8 \succ A_2 \succ A_3 \succ A_7$;

Modified Smart – $A_6 \succ A_5 \succ A_4 \succ A_1 \succ A_2 \succ A_8 \succ A_3 \succ A_7$.

Thus, research has shown that the three main companies were the best: GFK Group, TNS, AC Nielsen. In [11], the authors also identified the same companies as leaders. The analysis of the comparison of the methods results allows to conclude, that the modified algorithm of the Smart method gives the same results as the known methods for solving multicriteria optimization problems gives. At the same time, the use method Topsis and the modified algorithm of heuristic method Smart in mDSS is expediency, because the calculations of these methods is not resource – intensive. So the hardware requirements for mDSS are low, that makes it possible to use it on the mobile devices. Any user can use the mDSS on his smartphone or tablet at any time and get support in making decisions independently of access to the Internet.

V. CONCLUSION

In the work, the mDSS was developed on the Android platform. It based on the multi-criteria decision making

method Topsis and the modified algorithm heuristic method Smart, to simplify access to DSS and provides technical support to the design maker for obtain the multi-criteria tasks solutions.

REFERENCES

- [1] M.A. Demydenko, “Systemy pidtrymky pryynyattya rishen,” Natsional’nyy hirnychnyy universytet, 2016.
- [2] A.N. Tselykh, L.A. Tsiluch, “Mobil’ni dodatky dlya informatsiynkh system pidtrymky pryynyattya rishen’ u maloho biznesu,” Yzvestyya YUFU. Tekhnichni nauky, Vol. 11 (148), pp. 232 – 237, 2013.
- [3] X. Guo, and A. Díaz López, “Mobile Decision Support System Usage in Organizations,” Proceedings of the Nineteenth Americas Conference on Information Systems, Chicago, Illinois, August 15–17, 2013.
- [4] B.I. Yatsalo, “Systema bahatokryterial’noho analizu rishen’ DecernsMCDA ta yiyi praktychne zastosuvannya,” Prohrammni produkty ta systemy, Vol. 2, pp. 73-84, 2014.
- [5] N.M. Ershova, “Pryynyattya rishen’ na osnovi metodu analizu kherurkhiy,” Visnyk Prydniprovs’koyi derzhavnoyi akademiyi budivnytstva ta arkhitektury, Vol. 9 (210), pp. 38-45, 2015.
- [6] V.M. Postnykov Metody pryynyattya rishen’ v systemakh orhanizatsiynoho upravlinnya: navchal’ni. posobye / V.M. Postnykov, V.M. Chernen’kyy. - M. : Yzd-vo MHTU ym. N.E. Baumana, 2014.
- [7] H. Ahmadi, M. S. Rad, M. Nilashi, O. Ibrahim, and A. Almaee, “Ranking the micro level critical factors of electronic medical records adoption using Topsis method,” Health Informatics In. Journal, vol. 2, no. 4, pp. 19–32. November 2013.
- [8] M. B. Barfod, and S. Leleur, Multi-criteria decision analysis for use in transport decision making. (2 ed.). DTU Lyngby: Technical University of Denmark, Transport. 2014.
- [9] I.S. Romanchenko, M.M. Pot’omkin, “Metod Topsis-yadro ta yoho vykorystannya dlya bahatokolimoho porivnyannya al’ternatyv,” Obrobka informatsiyi v skladnykh orhanizatsiynkh systemakh, Vol. 1 (138), pp. 103-106, 2016.
- [10] O. A. Sianaki, “Intelligent decision support for energy management in demand response programs and residential and industrial sectors of the smart grid,” PhD thesis, Curtin University, 2015.
- [11] T.V. Bilorus, I.M. Kornilova, S.H. Firsova, “Orhanizatsiya marketynhovykh doslidzhen’ iz vykorystannya metodiv bahatokryterial’noho analizu,” Ekonomika i suspil’stvo, Mukachevo, Vol. 7, pp. 206-215, 2016.

Fuzzy Reconstructions in Linguistics

Olena Basalkevych
Medical Informatics Department
Danylo Halytsky Lviv National Medical University
Lviv, Ukraine
olena.basalkevych@gmail.com

Olexandr Basalkevych
Product Engineering and Development
Globallogic Inc.
Lviv, Ukraine
basalkevych.alex@gmail.com

Abstract — Application of fuzzy modeling in historical linguistics constitutes the contribution of the paper. Synonymic sets and word semantic structures of an Older Scots adjective are interpreted as associative fields, and on the other hand, as fuzzy sets. The resulting fuzzy associative linguistic diachronic reconstructions, including fuzzy associative fields (FAF) and fuzzy associative word structures (FAWS), along with their combination into a fuzzy associative adjectival network (FAAN) with further fuzzy semantic differentiation and lexis stratification, represent a fragment of a fuzzy adjectival associative thesaurus of Older Scots. Generally saying, the work demonstrates the use of artificial intelligence tools in diachronic linguistic, which is rather unexperienced practice.

Keywords — fuzzy associative field, fuzzy associative word structure; fuzzy associative adjectival network, fuzzy diachronic semantic differential, fuzzy lexis stratification, fuzzy adjectival associative thesaurus of Older Scots.

I. INTRODUCTION

Vague limits of linguistic phenomena promise a good success to the adaptation of fuzzy logic to the area. Lotfi Zadeh, the honoured father of fuzzy logic, argued that probability lacks sufficient “expressiveness” to deal with uncertainty in the natural language [1]. The specially prolific ground for fuzzy upgrading seems to be a linguistic reconstruction, the main research tool of the historical linguistics. In the concrete case, Older Scots lexicography makes the language canvas for the modeling. Synonymic sets and word semantic structures of an Older Scots adjective find their associative interpretation in neurolinguistics, and then, fuzzy evaluation in fuzzy logic. The consequent fuzzy associative linguistic diachronic reconstructions do not have any precedents, their synchronic prototypes are G.Kiss’ probabilistic associative thesaurus [2], Osgood’s semantic differential and Zadeh’s fuzzy stratification [14]. Supported by the Medical Research Council of Great Britain, the associative experiment issued into a weighed graph of G.Kiss, recreating the associative habits of the speakers of the English language, their cognitive and mental dispositions carved in the language is a pharos of the research strategy. Osgood’s Semantic Differential represents a rating scale designed by Osgood to measure the semantics or meaning of words, particularly adjectives, and their referent concepts. The last lecture of Lotfi Zadeh represents a fuzzy categorization instrument – stratification of reference information [14]. All these works contributed greatly to the evolved mathematical model. The reconstructions of the sort in diachronic linguistics are not revealed, although the great computational steps were made since the early seventies by Eastlack, Burton-Hunter, Rimmel, Hewson, Kondrak and others [3] in identifying cognates, reconstructing proto-forms, deriving reflexes, generally saying, in equipping the

theory of comparative method with robust reconstructive phonology instruments all based on the manipulation with recurrent sound correspondences of phonemes in cognates.

II. PRE-RESEARCH

A. Diachronic synonymic sets of Older Scots

The initial point of the study was the construction of Older Scots adjectival synonymic strings on the basis of the Dictionary of the Older Scottish Tongue (DOST), a part of the Dictionary of the Scots Language, available online [13]. The method involved the words from the English synonymic sets as subjects of the Dictionary surfing. The DOST answers formed unexpectedly large strings of semantically close words of very different connotations. Any word of the string was weighted following the formula:

$$W_i^{(y)} = \frac{y_i - y_{\min}}{y_{\max} - y_{\min}} \left(\frac{1}{n+1} - 1 \right) + 1 \quad (1)$$

where $(y_{\max} - y_{\min})$ is a diachronic range of a set, $(y_i - y_{\min})$ is a distance to the appearance of the concrete word in a set, and n – the number of words in a set [4]. The words were ordered in the way of ascending their advent dates so that the oldest ones were the first in the string. Any word was equipped with advent year, weight coefficient, orthography forms, etymology, meaning, diachronic text prototype, its literature source and author. The Access database IsetIntro was configured for the purpose [5].

B. Diachronic semantic word structure

On the basis of the mentioned adjectival sets, the database query for registering word entries into different sets was organized in order to detect all available meanings of a word, together with a text prototype, its author and date that is a criterion for the structure expanding. The entities were called diachronic semantic word structures [5].

III. ASSOCIATIVE INTERPRETATION

The multi-coloration of connotative values of set words and the presence of precedent texts makes it possible to reflect the study into associative linguistics’ plane. The mentioned diachronic synonymic sets of Older Scots could be interpreted as associative fields whilst DOST would be treated as a collective Older Scots brain representing its answers to the stimuli within the framework of an associative experiment (AE) with “historical reaction”. The principle of AE says that the first replicas are the strongest associations based on the most frequent usage.[6] Meanwhile the Zipf’s law states that the earliest words are the most frequently used ones [7]. Since the first positions in the gained synonymic sets belong to the oldest words with the biggest weight

coefficients, according to the Zipf's law we have kept the norm of AE and by this received the measure of associativity – the weight coefficient W_i . Consequently, the diachronic semantic word structure could also gain its associative interpretation. Having turned the diachronic component W_i into an associativity one, the main quantitative formant of a semantic structure becomes the measure of similarity or associativity of a certain word with different semantic features. Therefore it could be called an associative word structure.

IV. FUZZIFICATION

Associativity and vagueness seem to have a feature in common. One thing could be associated with others due to different factors (similarity, relevance or opposition) and just to some certain extend. In our case, the words of an associative field are associated with a forming semantic feature due to their synonymic nature or similarity that is rarely going to be absolute. The level of appropriateness or significance of set's components and the degree of belief or confidence (trueness) level of the made statements in fuzzy logic is introduced by a membership function (MF) [8]. The coefficient W_i favorably fits the nature of MF, introducing the level of word associativity with a semantic feature within the range [0;1]. So the words with higher MFs are treated to be more characteristic in the set and better associated with the semantic feature regarded. Therefore the fuzzy version of a mathematical model whose role is to regulate and prognosticate the studied process is expected to be the most profitable one.

A. Fuzzy associative fields

The mathematical model of research considers an associative field as a fuzzy subset of the set of all language adjectives. The subset is named by a semantic feature (a dominant) grouped under with elements allotted by MFs calculated according to formula 1. The fuzzy associative field (FAF) could be illustrated by e.g. Lucky={happy/1, hapin/0.99, sely/0.93, wel/0.9, wele/0.9, fortunit/0.76, fortunate/0.63, mervailous/0.63, fortunable/0.6, chancy/0.59, ewrous/0.53, lukkie/0.51, sonsy/0.4, canny/0.07}

A semantic feature or a dominant plays here the role of a linguistic variable, and all indicated subset names constitute its term-set: Dominant={Flexible, Inflexible, Brave, Cowardly, Intelligent, Stupid, Large, Small, Lucky, Unlucky,... }[9].

B. Fuzzy associative word structure.

Taking through the same fuzzification of the coefficient W_i , the associative word structure becomes a fuzzy one. It is clear that a word could be associated with this or that meaning to different extend, being closer to its direct value, and farther from a lateral one. The example for the word *Happy* is derived from the represented above FAF *Lucky*, together with these of *Successful*, and *Competent*:

$$\text{Happy} = \{\text{lucky}/1, \text{successful}/0.57, \text{competent}/0.38\}$$

First of all the structure is extended for the element *competent* as compared to [8], since the database has been enriched with more entries so far. Analyzing the MFs, we should state that *Happy* is the most associated with the

semantic feature *lucky* (or belongs mainly to the fuzzy subset *lucky*, in terms of fuzzy logic), less associated with *successful* and the least associated with *competent*.

Hence, we will call the structure of this sort a fuzzy associative word structure (FAWS) [8].

C. Fuzzy associative adjectival network. N-association trees. Proper application software.

The articulated two categories of words: meanings associated with a word, and words rendering the same value are the main relations in G.Kiss' Associative Thesaurus [3]. So, combining FAFs through their elements' FAWS results into a fuzzy associative adjectival network (FAAN). The example of it, connecting fields *Good*, *Strong*, *Inflexible*, *Lucky* and *Successful* through numerous FAWS including *Happy* is represented in fig. 1 [9].

The mentioned lexicon can be easily animated by the brain function of the speaker. To simulate the process, we are activating the created fuzzy network, where formally we can discern the set of places P, the set of transitions T, the input function I and the output function O. The input function I reflects the transition t_j into the set of positions $I(t_j)$ called the initial positions of a transition release. The output function

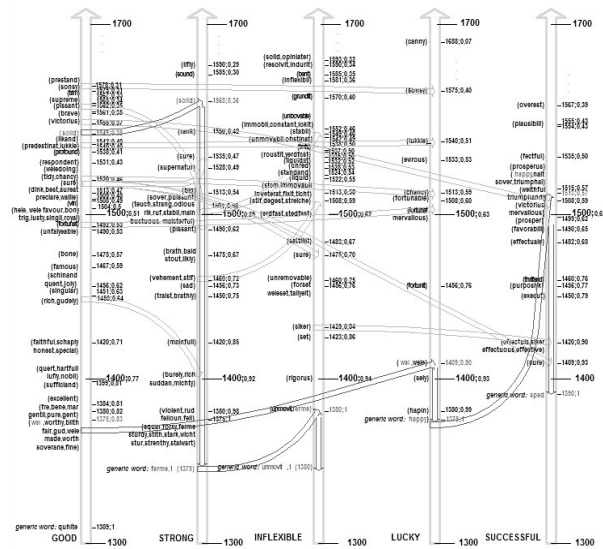


Fig. 1. Fragment of FAAN.

O reflects t_j into the set of output transition positions $O(t_j)$. Fuzziness is introduced into the set of places P, taken by words with appropriate MFs. The release of a transition rearranges the fuzziness distribution of the route, and the output place-word according to the rule $\mu(x_1 \cap x_2) = \min[\mu(x_1), \mu(x_2)]$ gains the minimum value of MFs of the input positions [9]. The fruit of this animation is in the probing of the n - step environment of any word within the net, the technique introduced by Kiss in his weighed graph. According to the FAWS of the regarded word, the transition to the most significant (with a maximum MF) output position which belongs to another associative field will be released. The following transition will be made to a dominant of the reached field. For the dominant we repeat the same algorithm. At the 2n-th step we will reach n-

association to the studied stimulus since as was already mentioned two transitions normally lead to any new association gain: between-field one and inter-field one. The performance of the constructed fuzzy associative adjectival network is oriented to the maximum belief (maximum MF) but is still sensitive to the weakest link of the route [9].

The mentioned algorithm of association-chain building is implemented with Java. The library Swing is involved for data visualization. The architecture is based on MVP design template (*Model-View-Presenter*) that delivers the visual reflection and event-procission behaviour into different classes, namely *View* and *Presenter*. The input data are given into format XML. The model DOM (*Document Object Model*) is employed for the work with XML-format. First of all the file of the model should be chosen and downloaded into the program. Then a user is pressing any word with the help of a mouse. The program visualizes the connections among words and calculates the general chain fuzziness. For example, association chain for *Victorious* is realized in Fig.2a. It is *Victorious-Sped* with MF or a belief degree 0.37. Associative reaction to *Wel* studied manually in [9] is *Happy-Sped* with belief 0.57 (Fig.2b). The association chain *Solid-Firm* is reached with the belief degree 0.36 (Fig.2c).

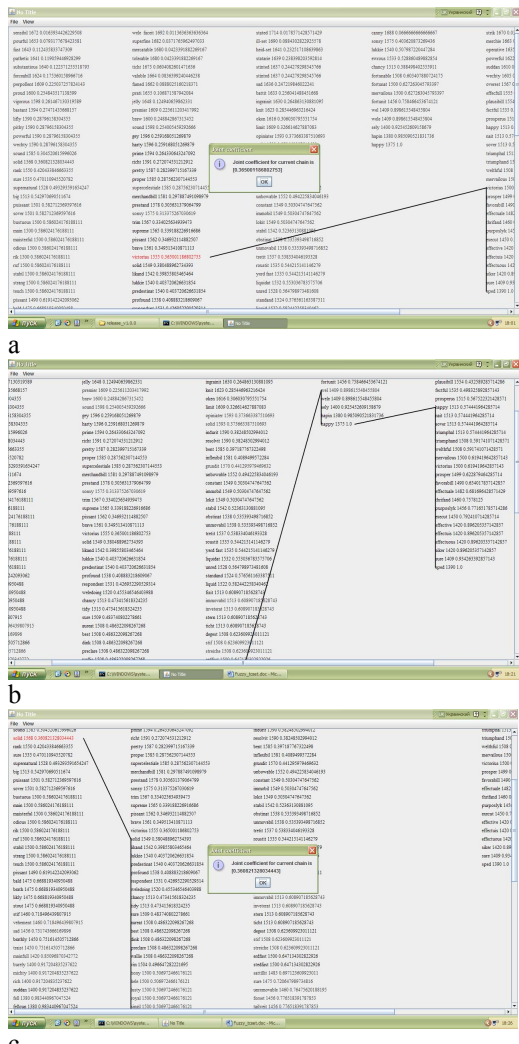


Fig. 2. Program realisation of association chains: a) Victorious-Sped; b) Happy-Sped; c) Solid-Firm.

D. Two way FAAN

The following research step is an addition of an antonymous flank to the received structure. It is known from AE that the largest associative power belongs to the antonym of the stimulus. So we connect the antonymic fuzzy associative field to the original one, creating the base of a triangular plane. The basic opposition is the pair of a dominant and its antonym: Successful-Unsuccessful, Lucky-Unlucky, Flexible-Inflexible, Strong-Weak, Good-Bad with two way connection (Fig.3). In the case, an association chain is going to contain two branches: positive reactions from Successful-Lucky-Inflexible-Strong-Good and negative from Unsuccessful-Unlucky-Flexible-Weak-Bad, thus creating an association tree. For instance, we can choose for the stimulus the regarded in *IVc* word *Solid*. Evolving the algorithm of network simulation, the other 2n steps are applied to the time-antonym of the studied stimulus with a little bit shuffling order: the transition to a dominant is realized immediately since the gravity to the main bearer of the meaning triggers first [8].

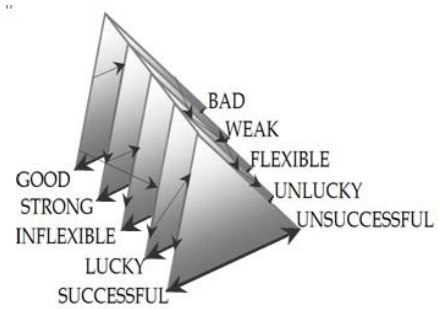


Fig. 3. The model of two-way FAAN.

Then we get the enriched chain reaction $Solid = (+)Firme/0.36 \text{ And } (-)Lidder-Wikit-Perilous/0.38$. The general association will get $0.36 (= \min(0.36; 0.38))$ belief degree, which is not too strong sureness that the reaction of a Scottish man of 1300-1700 to the word *Solid* would be like this.

E. Fuzzy diachronic semantic differential

The special interest is attracted to the triangular cut of the structure. If to unbend the wings of the figure, the chronomodel of Osgood's semantic differential could be clearly observed. Following the same Zipf's law background along with fuzzy interpretation of the weight coefficient W_i , we proceed to fuzzy diachronic semantic differential (FDSD). Its normalized variant is shown in Fig.4.

All the meaning subtleties for concepts' dichotomy Strong-Weak are introduced along the triangle sides. The chaotic layout of the meanings is normalized by the introduction of time intervals [11]. Now with the help of this historic evaluative device we can estimate the main characteristics of certain words throughout the history of Older Scots. Following the represented in DOST diachronic text prototypes of the reporting words of our differential, we can state that:

Company could be **buryly** with the degree of belief **0.917**.

Women could be **odious** with the degree of belief **0.586**

Watteris, fluidis could be **proud** with the degree of belief **0.255**

Leggs could be **wankle** with the degree of belief **0.019**

Me could be **bauche** with the degree of belief **0.290**

Complexioun could be **selie** with the degree of belief **0.606**

Woman could be **brukil** with the degree of belief **0.984**

The most prominent representatives of syntagmatic associations are taken from more numerous selection of [12]. The short research issue reveals that according to Older Scots pattern of cross-cultural universal Strong-Weak the negative personality descriptor for **woman** is more often **brukil** than **odious**: Woman = {brukil/0.984, odious/0.586}.

By this study we are extending our reconstruction with syntagmatic associations, the very important component of Kiss's Associative Thesaurus and AE in general.

analysis could open possibilities of ascertaining the lexis stratum gravity poles, time antonymy and significance status of the strata, all enabled by the manipulation with border MFs. The fuzzy variable Gravity = (strong/ $\mu_1(x)$; weak/ $\mu_2(x)$) is introduced in any stratum, where $\mu_1(x)$ and $\mu_2(x)$ are the MFs of the sectional antonymic pair. It makes possible to evaluate the preferring concept for the time-stratum, in other words, the intensity of positive and negative concepts development during some period. Time antonymy could be analyzed not only for sectional pairs but at any moment within any stratum by straight bridging two flanks and their immediate representatives. The notion of a core and a periphery of a fuzzy set will help us to analyze the significance of a concrete stratum establishing the specific fuzziness that divides the core strata from peripheral ones, by this restricting the research concentration [11].

In our case the contrasted pairs are

1400, (Burely/0.917) - 1380, (Wery/0.984)

1500, (Rik/0.586) - 1500, (Frail/0.606)

1600, (Proud/0.255) - 1600, (Bauch/0.290)

For the taken error of 50 years, in 1700 the time opposition is absent (Fig.3).

Now adapting the fuzzy variable Gravity = (strong/ $\mu_1(x)$; weak/ $\mu_2(x)$) for any of strata, again considering the sectional MFs, we could characterize them:

1300-1400 stratum: Gravity = (strong/0.917; weak/0.984);

1400-1500 stratum: Gravity = (strong/0.586; weak/0.606);

1500-1600 stratum: Gravity = (strong/0.255; weak/0.290).

The Gravity gives arguments in favour of the Weak concept for all strata. So negative, a more developed since more exciting and alarming concept, becomes the gravity pole for the whole stratification figure [12].

For the core strata let us adopt MF 0.25 as a fuzzy measure. Assimilating this level for both flanks we cut the last 1600-1700 stratum and analyze the preceding three only.

V. CONCLUSIONS

New notions of fuzzy associative fields and fuzzy associative word structure benefit from the realized symbiotic approach. Being combined together, they constitute the fuzzy associative adjectival network, a singular vertical cut of which is an extended fuzzy modification of a semantic differential or a fuzzy diachronic semantic differential. Another evaluation attitude, more volumetric one, considers the gained plane as a fuzzy lexis stratification structure consisting of the strata of lexical oppositions. The described metamorphoses are realized with the help of the database in Access and implemented with Java using Eclipse IDE. The prospect of the research is the evolution of fuzzy adjectival associative thesaurus of Older Scots. The practical implementation of the proposed fuzzy verbal network opens possibility of simulating brain activity of the speakers of more or less antique languages revealing their cognitive and mental peculiarities, by this recreating and contrasting the world verbal pictures of different older languages, which is of interest for historical cognitivists.

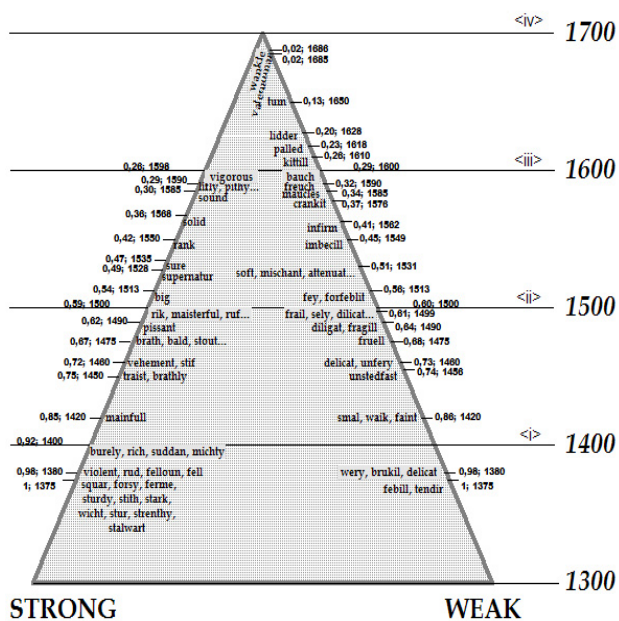


Fig. 4. Model of the normalized FDS.

The complete picture claims to be the fragmental fuzzy adjectival associative thesaurus of Older Scots, and the monads of words collocated with a DOST reaction-words should enrich the previous research practices. The propagation of fuzziness through the fuzzy network is playing here its activating role.

The other known function of a semantic differential is to evaluate the psycho-linguistic portrait of a person. In a concrete case we could speculate about characteristics of a collective Older Scottish speaker represented by the myriad of Scottish penmen. For instance, the mentioned fuzzy subset Woman reveals the expressions of a negative attitude of an average Older Scot towards females. The aforesaid makes the sketch of prospect semantic research following all the conventional cross-cultural universals (evaluation, potency and activity): Good-Bad, Strong-Weak, Active-Passive.

F. Fuzzy lexis stratification

The scheme in Fig.2 implies another research-strategy – the stratification of lexis. The time intervals, common for both flanks, comprise the time lexis oppositions, combining them into four strata. The idea of fuzzy stratification belongs to late Lotfi Zadeh, and is widely involved into processing encyclopedias and notebooks. In the concrete case, the fuzzy

ACKNOWLEDGMENT

The special gratitude should be expressed to Vitaliy Arseniyovych Lishchynsky (1940-2015), the supervisor of the author's diploma project "Fuzzy network quality-research and modeling" performed in 1997 in "Lviv Polytechnic" State University at the Department of Applied Mathematics, whose advanced ideas and benevolent assistance made the premises for evolving the proposed mathematical model.

REFERENCES

- [1] M. Laviolette, J. W. Seaman, Jr., J. Douglas Barrett, and W. H. Woodall, "A Probabilistic and Statistical View of Fuzzy Methods," *Technometrix*, Taylor & Francis, Ltd., Vol. 37, No. 3, pp. 249-261, Aug., 1995.
- [2] G. R. Kiss., C. Armstrong, R. Milroy, and J. Piper, "An associative thesaurus of English and its computer analysis," In Aitken, A.J., Bailey, R.W. and Hamilton-Smith, N. (Eds.), *The Computer and Literary Studies*. Edinburgh: University Press., 1973.
- [3] G. Kondrak "Algorithms for Language Reconstruction," University of Toronto, pp.169, 2002.
- [4] M. Bilynskyi, "Synonymous string as a diachronic reconstruction: The OED Middle English Evidence for verbs and deverbal coinages," *Inozemna Philologia*, Lviv, Vol.121, pp. 20-43, 2009.
- [5] O. Basalkevych, "Diachronic model of adjective synonymic set evolution in Older Scots," IX International scientific practical conference: "Language-Culture-Personality", Scientific records of "Ostroh Academy" National University, Ostroh, Ukraine, vol. 54, pp. 91-97. 2015.
- [6] M. Arapov, "Mathematical methods in historical linguistics," Moscow, "Nauka", 1974.
- [7] D. Nelson, C. McEvoy, and S. Dennis, "What is and what does free association measure?" *Free Association*, pp.3-46, 2002.
- [8] M. Siavavko, "Informational system "Fuzzy expert"," Lviv, Ivan Franko Lviv National University Press, 2007.
- [9] O. Basalkevych, "Fuzzy modelling of associative maps," IV International scientific practical conference "Linguocognitive and sociocultural aspects of communication," Scientific records of "Ostroh Academy" National University, Ostroh, Ukraine, vol. 56, pp. 33-38. 2015.
- [10] V. Lishchynsky, and O. Basalkevych, "Analysis of Petri Nets with one fuzzy element," the Proceedings of the Seventh Pan-Ukrainian Scientific Conference, Lviv, Ukraine, pp.61-62, 2000.
- [11] O. Basalkevych, "Fuzzy diachronic differential," Proceedings of the IV Ukrainian-German conference "Informatics. Culture. Technics.", Odesa, Ukraine, 30 June-2 July, pp.106-108 2016.
- [12] O. Basalkevych, "A model of the diachronic semantic differential: stratification of lexis on fuzzy principles (the case of Older Scots adjectives)," *Inozemna Philologia*, LIFNU Press, vol. 130, pp. 16-27, 2017.
- [13] "The Dictionary of the Scots Language," Edinburgh: Scottish Language Dictionaries, 2007. www.dsl.ac.uk
- [14] L. Zadeh, "Stratification, target set reachability and incremental enlargement principle," The Design of Robotics and Embedded systems, Analysis and Modeling Seminar, 250 Saturdja Dai Hall, UC Berkeley, 2016. www.youtube.com/watch?v=Ok0vJ0uykCA&list=PL8092C965854FA5BB&index=30

Fuzzy Mathematical Modeling Financial Risks

Oleksey Voloshyn

*Department of complex systems modelling
Taras Shevchenko National University*

Kyiv, Ukraine
olvoloshyn@ukr.net

Marianna Sharkadi

*Department of cybernetics and applied mathematics
Uzhgorod National University*

Uzhgorod, Ukraine
marianna.sharkadi@uzhnu.edu.ua

Mykola Malyar

*Department of cybernetics and applied mathematics
Uzhgorod National University*

Uzhgorod, Ukraine
malyarimm@gmail.com

Volodymyr Polishchuk

*Department of Software Systems
Uzhgorod National University*

Uzhgorod, Ukraine
v.polishchuk87@gmail.com

Abstract — The research of the urgent task of developing a fuzzy mathematical model of financial risks for evaluating projects regarding the level of security of their financing has been carried out. The development of such technology will provide an opportunity to adequately approach the consideration of projects, increase the degree of validity of investment decisions and increase economic security.

Keywords — projects, risks, linguistic evaluation, information technology, multicriteria, security financing, venture capital funds.

I. INTRODUCTION

Fuzzy mathematical modeling is one of the most active and promising areas of applied research in the field of management and decision making in weakly structured systems. The range of fuzzy methods is expanding every year, embracing various new areas. Fuzzy mathematical modeling is this when the elements of the study are not numbers, and some fuzzy sets or their combination. The basis of this approach lies not in traditional logic, but in logic with fuzzy truth, fuzzy ties and fuzzy rules of output. The main characteristics of this approach are the use of linguistic variables instead of numerical variables, the relationship between variables is described by fuzzy statements, and complex relationships are described by fuzzy algorithms.

When designing and managing a complex socio-economic system, a problem arises when a person is not able to give accurate and, at the same time, practical meanings of judgments about their behavior.

The paper proposes research of the actual problem of developing a mathematical model of information technology for risk assessment of projects regarding the level of security of their financing, using fuzzy mathematics, for various investment subjects. The development of such technology will provide an opportunity to adequately approach the consideration of projects, increase the degree of validity of investment decisions and, in general, increase economic and managerial security.

Financing projects of any nature (project startup or classical investment) is a risky activity. Depending on the origin of the project, there are various options for its financing, such as business angels, venture and investment funds, banks. Each of these institutions has its own risk management policy. But all of them combine one thing: to

find and finance a successful project with minimal risks.

Risk is closely linked to the concept of economic security of the project, both as the security of the entity representing the project, and the security of the investor. The subject's security is that a risky and unsuccessful project will lead to damage to the enterprise. The investor's security is directly dependent on an adequate assessment of the project and the entity presenting the project. Increasing the security of investment projects provides stability of the regional economy [1].

Recent scientific studies indicate the need to systematize risk minimization tools and develop a algorithm for selecting a model for evaluating projects of different origins. The issue of quantitative risk assessment and risk management during investing is disclosed in many papers [1-3], but a holistic concept for determining the level of risk, reducing it and taking into account the subjective aspects of the assessment has not yet been developed.

II. FORMULATION OF THE PROBLEM

Depending on the origin, commercial projects are dealt with in two types: classical - investment projects under which a well-formulated business plan emerges in a company operating on the market and requires a partial involvement of funds from the outside; startup projects - the "idea" that arises in companies whose business is based on innovative technologies, such companies have not entered the market or have just started to come out of it and need to attract external resources.

We formulate the task of evaluation as follows. Suppose we have some projects S_1, S_2, \dots, S_n , for which an evaluation of the risk with regard to the level of security of their financing should be done. Projects can have different perspectives, nature and security of implementation. Without diminishing the universality, we will continue to consider one project. In case of a plurality of projects, they can be ordered according to the initial estimates received. The model of the problem is represented in the following form:

$$SPF = O(O_S, O_G, O_R), \quad (1)$$

where O_S – evaluation of the project under consideration, depending on its origin (classic investment project [4] or

startup project [5]), O_G – assessment of the economy in which a commercial project will be implemented [6], O_R – aggregated risk assessment for project implementation. SPF – initial assessment and linguistic treatment of risk in relation to the level of project financing security. O – operator that matches the output variable SPF , with input estimates O_S, O_G, O_R .

III. MATERIALS AND METHODS

Let us offer the next set of start-up risk assessment criteria on which the platform can assess risks of start-ups. There are four groups of criteria: K_O – "operational risk"; K_I – "investment risks"; K_F – "financial risks"; K_S – "risks of innovation".

A person who decides (Decision Maker) from each group chooses risk criteria are criteria that can assess the proposed project. Let us represent each group of criteria in the form of a set of indicators. Then to the group of criteria K_O – "operational risk" the following indicators can be considered: K_{O1} – the risk of loss of the client base; K_{O2} – the risk of loss of the supplier; K_{O3} – the risk of losing market share; K_{O4} – the risk of lowering the level of management; K_{O5} – the risk of industrial conflict and ineffective motivation; K_{O6} – the risk of lowering the quality of the processes; K_{O7} – the risk of lowering labor productivity; K_{O8} – personnel risks; K_{O9} – the risk of unsecured resources.

The "investment risks" group – K_I we will express through the following indicators: K_{I1} – the risk of inefficiency of investments; K_{I2} – risk of disruption of the terms of creation of production assets; K_{I3} – the risk of failure to achieve the return on investment capital; K_{I4} – the risk of exceeding the amount of start-up investment; K_{I5} – the risk of a lack of investment capital.

The "financial risks" group – K_F we will express through the following indicators: K_{F1} – the risk of inefficient use of capital; K_{F2} – risk of loss (arises due to price changes, when sudden expenses cover revenue); K_{F3} – the risk of investor loss; K_{F4} – the risk of loss of solvency; K_{F5} – the risk of a suboptimal capital price.

The criteria of the "risks of innovation" group – K_S we will express with the help of such indicators: K_{S1} – the risk of ineffective innovation investments; K_{S2} – the risk of ineffective promotion of innovations; K_{S3} – risks of breaking the terms of innovation development; K_{S4} – risks of technology innovation; K_{S5} – risks of resource insufficiency when designing innovations.

This set of risk criteria can not reveal all aspects of any startup of the project in various areas of implementation, so

it is open and any expert can add some criteria depending on the scope of investment.

Each risk criteria is evaluated by experts with one of the terms of the following term-set of linguistic variables $L = \{H; HC; C; BC; B\}$, where: H – «low risk level»; HC – «risk level below average»; C – «average risk level»; BC – «risk level above average»; B – «high risk level». Also, an expert puts the number of «authenticity» for each assessment of risk level $\mu(L)$ of his consideration concerning the interval [0; 1].

We describe a two-tier scheme of a project risk assessment model based on input linguistic variables. The inputs are presented in the form of linguistic variables and the reliability of the expert's consideration of their assignment. Therefore, at the first level, it is necessary to build membership rules and knowledge base in order to obtain the resulting term-assessment L^α for each group of risk criteria. On the basis of obtained resultant term evaluation L^α to determine the aggregated estimation of reliability $\mu(L^\alpha)$. At the second level, estimates are obtained L^α and $\mu(L^\alpha)$ we will design a "risk axis" to determine one project risk assessment for each group of criteria α .

Consider the first level - the construction of the rules of ownership of the resulting term evaluation by the groups of risk criteria.

Level H: «low risk level». The minimum amount of criteria with low risk level term should not be less than 60% and the remaining 40% of the criteria should not have terms lower than «risk level below average».

Level HC: «risk level below average». The project should have the minimal amount of criteria with the term «risk level below average» not less than 60%, and the other 40% of criteria should have terms not lower than the «average risk level».

Level C: «average risk level». The minimal amount of criteria with the term «average risk level» not less than 60%, and the other 40% of criteria should have terms not lower than the «risk level above average».

Level BC: «risk level above average». The minimal amount of criteria with the term «risk level above average» not less than 60%, and the other 40% of criteria should have terms not lower than the «high risk level».

Level B: «high risk level». The project gets the resulting term-evaluation «B» in case the amount of criteria with the term «high risk level» compiles 60% and more.

Then, based on the established rules of ownership of the resulting term evaluation for the groups of risk criteria, we can give a fragment of the knowledge base, for example, according to five criteria, Table 1.

Because the expert puts each variable L_i^α the reliability of their reasoning – $\mu(L_i^\alpha)$ from the interval [0; 1], $\alpha = \{O; I; F; S\}$ then linguistic variables can be represented,

for example, in the form of triangular membership functions.

TABLE I. A FRAGMENT OF KNOWLEDGE BASE

No	$K_{\alpha 1}$	$K_{\alpha 2}$	$K_{\alpha 3}$	$K_{\alpha 4}$	$K_{\alpha 5}$	Resulting term evaluation
1	H	H	H	HC	HC	H
2	H	H	HC	HC	HC	HC
3	HC	HC	HC	C	C	
4	HC	HC	HC	H	C	
5	C	C	C	HC	HC	C
6	C	C	C	BC	BC	
7	C	C	C	HC	BC	
8	C	C	C	H	HC	BC
9	BC	BC	BC	C	C	
10	BC	BC	BC	C	B	
11	BC	BC	BC	B	B	B
12	BC	BC	BC	HC	C	
13	B	B	B	BC	BC	
14	B	B	B	BC	C	

The aggregated authenticity estimation $\mu(L^\alpha)$ $\alpha = \{O; I; F; S\}$ is calculated with the following formula:

$$\mu(L^\alpha) = \frac{1}{n} \sum_{i=1}^m \mu(L_i^\alpha), \quad \alpha = \{O; I; F; S\}, \quad (2)$$

Where $\mu(L_i^\alpha)$ – is the authenticity estimation of the linguistic variables which match the resulting term-evaluation for i -criterion of α risk criteria group.

At the second level, we will design the data on the risk criteria groups into a "axis of risk" to determine a generalized risk assessment of the project for each group of criteria α and obtaining an aggregated risk assessment, as well as its linguistic interpretation.

Next for each group of criteria α , let's express it x^α :

$$x^\alpha = \begin{cases} \sqrt{\frac{\mu(L^\alpha)}{2}}(b-a) + a, & 0 \leq \mu(L^\alpha) \leq 0,5; \\ b - \sqrt{\frac{1-\mu(L^\alpha)}{2}}(b-a), & 0,5 < \mu(L^\alpha) \leq 1. \end{cases} \quad (3)$$

To obtain a generalized project risk assessment for groups of criteria α , use the following formula:

$$z^\alpha = \frac{x^\alpha}{100}. \quad (4)$$

Three variables $(x(L^\alpha); \mu(L^\alpha); z^\alpha)$ we interpret the three-dimensional coordinate system, where $x = x(L^\alpha)$ – the value of a function equal to the numerical interpretation of the resulting term-estimates $L = \{H; HC; C; BC; B\}$ $y = \mu(L^\alpha)$ – aggregated assessment of the reliability of the expert's thoughts, and the axis $z = z^\alpha$ – project risk assessment for each group of criteria α , which project is on the "axis of risk".

Aggregated risk assessment for all groups of criteria α we calculate as follows:

$$O_R = \frac{1}{4} \sum_{\alpha} (1 - z^\alpha). \quad (5)$$

Since the evaluation is received O_R normalized, then to compare it with the output variable R the following scale is proposed: $r_1 = \langle \text{negligible risk level of project} \rangle$; $r_2 = \langle \text{low risk of project} \rangle$; $r_3 = \langle \text{average risk of a project} \rangle$; $r_4 = \langle \text{high risk of project} \rangle$; $r_5 = \langle \text{extreme risk level of project} \rangle$. Linguistic treatment of aggregated risk assessment $R = \{r_1, r_2, r_3, r_4, r_5\}$ define, for example, the following scale: $O_R \in (0,87; 1] - r_1$; $O_R \in (0,67; 0,87] - r_2$; $O_R \in (0,36; 0,67] - r_3$; $O_R \in (0,21; 0,36] - r_4$; $O_R \in [0; 0,21] - r_5$.

IV. RESULTS AND DISCUSSION

Formulate a generic algorithm to obtain aggregated safety assessment project financing.

1 step. Determine the resulting term-evaluation.

Based on expert input imposed on the project and built the knowledge base determines the resulting term-evaluation criteria for groups: K_O ; K_I ; K_F ; K_S .

2 step. Determination of the aggregated estimation of reliability of expert considerations.

Aggregate validation calculates $\mu(L^\alpha)$, $\alpha = \{O; I; F; S\}$ according to the formula (2).

3 step. Obtaining a single generalized project risk assessment for groups of criteria α .

For each group of criteria we calculate the relative percentage scale $[a; b]$ and resultant term evaluation L^α , (which has the level of risk content) by the formula (3). A generalized project risk assessment for each group of criteria α gets for (4).

4 step. Calculation of aggregated risk assessment for all groups of criteria.

Aggregate risk assessment is determined by (5).

5 step. Output level of project financing security.

Match the evaluation O_R with output variable R to obtain a linguistic interpretation of the level of security of project financing.

For an example, consider the following problem. You need to build an initial estimate SPF and a linguistic interpretation of risk regarding the level of security of project financing. At this stage, we have an assessment of the project, depending on its origin – O_S , an assessment of the economy in which a commercial project will be implemented – O_G , aggregated risk assessment for project implementation – O_R .

Let the Decision Maker for each assessment may specify weight ratios $\{p_S, p_G, p_R\}$ from some interval. Then we will determine the normalized weight coefficients accordingly:

$$\alpha_\delta = \frac{p_\delta}{\sum_\delta p_\delta}, \quad \delta = \{S, G, R\}, \quad \sum_\delta p_\delta = 1. \quad (6)$$

Since all the estimates obtained are normalized from the interval $[0; 1]$, we use the following approach to obtain a final assessment of the security level of the project financing. Depending on the psychological perception of the situation Decision Maker can choose one of the convolutions [6]:

$$\text{Pessimistic } M_1(S) = \frac{1}{\sum_\delta \frac{\alpha_\delta}{O_\delta}}; \quad (7)$$

$$\text{Cautious } M_2(S) = \prod_\delta (O_\delta)^{\alpha_\delta}; \quad (8)$$

$$\text{Average } M_3(S) = \sum_\delta \alpha_\delta O_\delta; \quad (9)$$

$$\text{Optimistic } M_4(S) = \sqrt{\sum_\delta \alpha_\delta (O_\delta)^2}. \quad (10)$$

Thus, we obtain an initial estimate from the interval $[0; 1]$. For the linguistic interpretation of risk, the value obtained by formulas (7) - (10) is comparable to one of the term sets $SPF = \{SPF_1, SPF_2, \dots, SPF_5\}$. The scale of estimates can be determined as follows: $M(S) \in (0,77; 1]$ – SPF_5 («high level of security of project financing»); $M(S) \in (0,57; 0,77]$ – SPF_4 («the level of security of project financing is above average»); $M(S) \in (0,36; 0,57]$ – SPF_3 («average level of project financing security»); $M(S) \in (0,21; 0,36]$ – SPF_2 («low level of project financing security»); $M(S) \in [0; 0,21]$ – SPF_1 («very low level of

project financing security»).

Depending on the different periods of the project implementation, we can review the initial assessment and aggregate risk assessment of the project implementation.

V. EXPERIMENTS

Let some investment project undergo an expert evaluation. The values of the linguistic variables and the authenticity of their assignment are as follows:

1. K^O – "operational risk": K_1^O (H; 0,8); K_2^O (H; 0,7); K_3^O (HC; 0,9); K_4^O (H; 0,6); K_5^O (HC; 0,7); K_6^O (C; 0,5); K_7^O (H; 0,7); K_8^O (H; 0,8); K_9^O (H; 0,9).

2. K^I – "investment risks": K_1^I (HC; 0,7); K_2^I (H; 0,5); K_3^I (C; 0,6); K_4^I (HC; 0,8); K_5^I (HC; 0,9).

3. K^F – "financial risks": K_1^F (HC; 0,3); K_2^F (HC; 0,6); K_3^F (HC; 0,2); K_4^F (H; 0,7); K_5^F (H; 0,6).

4. K^S – "risks of innovation": K_1^S (H; 0,8); K_2^S (H; 0,9); K_3^S (HC; 0,1); K_4^S (HC; 0,7); K_5^S (HC; 0,6).

To obtain an aggregated safety assessment of a project financing, use the following algorithm:

1 step. Determination of the resultant term evaluation.

Based on the knowledge base for each group of risk criteria, we define the resulting term evaluation: "operational risk" – H; "investment risks" – HC; "financial risks" – HC; "risks of innovation" – HC.

2 step. Calculation of the aggregated estimation, of the reliability, of the experts reasoning. Aggregate validation $\mu(L^\alpha)$, $\alpha = \{O; I; F; S\}$ calculate according to the formula (2):

$$\mu(L^O) = 1/6 * (0,8 + 0,7 + 0,6 + 0,7 + 0,8 + 0,9) = 0,8;$$

$$\mu(L^I) = 1/3 * (0,7 + 0,8 + 0,9) = 0,8;$$

$$\mu(L^F) = 1/3 * (0,3 + 0,6 + 0,2) = 0,6;$$

$$\mu(L^S) = 1/3 * (0,1 + 0,7 + 0,6) = 0,5.$$

3 step. Obtaining a single generalized project risk assessment for groups of criteria α .

For each group of criteria α we calculate by the formula (3). A generalized project risk assessment for each groups of criteria α get for (4).

$$x^0 = 20 - \sqrt{\frac{1-0,8}{2}}(20-0) = 13;$$

$$z^0 = \frac{13}{100} \approx 0,13;$$

$$x^I = 40 - \sqrt{\frac{1-0,8}{2}}(40-20) = 33,7; \quad z^I = \frac{33,7}{100} \approx 0,34;$$

$$x^F = 40 - \sqrt{\frac{1-0,6}{2}}(40-20) = 29; \quad z^F = \frac{29}{100} \approx 0,29;$$

$$x^S = \sqrt{\frac{1-0,5}{2}} \cdot 20 + 20 = 30; \quad z^S = \frac{30}{100} \approx 0,3.$$

4 step. Calculation of aggregated risk assessment for all groups of criteria α . Aggregate risk assessment is determined by (5):

$$O_R = \frac{1}{4}((1-0,13) + (1-0,34) + (1-0,29) + (1-0,3)) = 0,74.$$

5 step. Determine the level of security of project financing.

Match the evaluation O_R with output variable R to obtain a linguistic interpretation of the level of security of project financing.

Because $O_R \in (0,67; 0,87] - r_2$, then the project under consideration will receive "a low level of project risk or a level of security of project financing above average".

Depending on the different periods of the project implementation, we can review the initial assessment and aggregate risk assessment of the project implementation.

Built in such a way two-level fuzzy mathematical model,

obtaining an aggregated risk assessment of the project, has a number of advantages, namely: uses the expert's reasoning for assessing the various risk criteria; the reliability of his reasoning and, based on this; the aggregation of opinions according to the groups of criteria in the final assessment. The disadvantages of this approach include the use of different models of membership functions, which can lead to ambiguity of end results.

REFERENCES

- [1] M. Kelemen, "Problems of Protected Interests in the Security Sectors," Warsaw: Wydawnictwo Wyższej szkoły menedżerskiej w Warszawie im. Prof. Leszka J. Krzyżanowskiego, 2015.
- [2] I. I. Verbitska, "Ryzik-menedzhment yak suchasna systema upravlinnya ryzykamy pidpryyemnytskykh struktur," Stalyy rozvytok ekonomiky, Vol. 5, pp. 282-291, 2013.
- [3] M. Kelemen, S. Krizovsky, S. Kocan, "Vyuzitie technologie LVA (vrstvena analiza hlasu) v bezpecnostnej praxi, na prevenciu proti podvodom u poisovacich a financnych spolocnosti", 1. vyd. Kosice: Vysoka skola bezpecnostneho manazerstva v Kosiciach. 100, 2012.
- [4] M. Malyar, V. Polishchuk, "Choice and evaluation methodics of investment projects," Kosicka bezpecnostna revue, Kosice. 1, pp. 117-126, 2013.
- [5] M. Malyar, V. Polishchuk, M. Sharkadi, and I. Liakh, "Model of start-ups assessment under conditions of information uncertainty," Eastern European Journal of Enterprise Technologies, Mathematics and cybernetics – applied aspects, vol. 3/4(81), pp.43-49, 2016.
- [6] M. Malyar, V. Polishchuk. Ranking method of alternative options of inhomogeneous nature. Kosicka bezpecnostna revue, Kosice, 1, 60-67, 2016

The Methods Bayesian Analysis of the Threshold Stochastic Volatility Model

Peter Bidyuk
National Technical University of Ukraine
"Igor Sikorsky Kyiv Polytechnic Institute"
Kiev, Ukraine
pbidyke@gmail.com

Aleksandr Gozhyj
Petro Mohyla Black Sea National
University
Nikolaev, Ukraine
alex.gozhyj@gmail.com

Zdislaw Szymanski
Spoleczna Akademia Nauk
Warsaw, Poland
zszymanski@san.edu.pl

Irina Kalinina
Petro Mohyla Black Sea National
University
Nikolaev, Ukraine
irina.kalinina1612@gmail.com

Volodymyr Beglytsia
Petro Mohyla Black Sea National
University
Nikolaev, Ukraine
science@chmnu.edu.ua

Abstract—The paper considers the Bayesian analysis of the threshold stochastic volatility models. Studies of methods for analyzing stochastic volatility and improving models of stochastic volatility significantly improve the quality of forecast models and their estimates. Bayesian inference is performed by tailoring Markov chain Monte Carlo (MCMC) or sequential Monte Carlo (SMC) schemes that take into account the specific characteristics of models. The results of applying the method demonstrated in models heteroscedastic non-stationary processes.

Keywords—Bayesian analysis; stochastic volatility; Threshold stochastic volatility model

I. INTRODUCTION

With the current economic instability, trading on stock markets carries a high risk. Therefore, the study of volatility, a statistical measure of stock prices, becomes instrumental and often indispensable. Currently, the study of volatility has become the basis of the financial economics, and one of the main tools of financial analysis and modeling, in particular. The ground of these studies lies in various probabilistic and statistical volatility models. Statistical volatility models are widely used in various financial tasks, such as the estimation of the standard deviation of the market returns, risk assessment, evaluation of the financial instruments, etc. There are dozens of volatility measurement methods, ranging from technical indicators such as the average true range (ATR), historical volatility, stochastic volatility of various types, standard deviation, etc. In addition to the financial analysis, conditional variance models are widely used in medical and technical diagnostic systems, risk assessment and management, social studies, etc. The study of stochastic volatility analysis methods and the improvement of the models' structure can substantially improve the quality of their forecasting and estimates. Therefore, this research is aimed at the investigation of the method of Bayesian analysis of the threshold stochastic volatility model.

II. VOLATILITY MODELING WAYS

The stochastic volatility model is based on the conditional heteroscedastic model. The conditional heteroscedasticity model (ARCH) was developed by Robert F. Engle [1] to create a model of inflation in the UK. This

model was later used for stock prices and exchange rates modeling [2]. The further development of ARCH is the generalized autoregressive conditional heteroscedasticity (GARCH) described in the investigation [3,4], which is still actively used for volatility forecasting [5, 6]. Models like GARCH allow recreating the phenomenon of volatility clustering (GARCH-effect). Parameters of ARCH / GARCH models are most often evaluated by the maximum likelihood estimation. One of the main disadvantages of the GARCH model is that the model memory is "not long enough" since its autocorrelation function (ACF) is characterized by the exponential decay. When the sum of coefficients of the model $\alpha + \beta$ is close to 1, the GARCH model degenerates into a non-stationary process, named an integrated generalized autoregressive conditionally heteroscedastic (IGARCH) process [2]. However, the latter model implies the dependence of volatility on the initial conditions which does not disappear within an infinite planning horizon, and therefore cannot be claimed to be an adequate reflection of the reality. An alternative approach is to use stochastic processes or models whose theoretical properties imply the presence of the so-called "long" memory.

III. MODELS OF STOCHASTIC VOLATILITY

The main idea behind the stochastic volatility models is to increase the number of randomness sources. In conditional heteroscedasticity models, there is only one source of randomness, and the variation of the process is assumed to be dependent on its previous realizations in one form or another. An alternative way of modeling is to provide the price dynamics in the form of a simple model, like a differential equation. However, the volatility σ in this equation is rather a separate stochastic process, than a parameter. Consequently, there are two independent sources of randomness. The first stochastic volatility model was suggested by [11]. In particular, it assumed that logarithmic volatility is a process AR (1):

$$\begin{aligned}r_t &= \mu\sigma_t\varepsilon_t \\ \ln \sigma_t^2 &= \varphi \ln \sigma_{t-1}^2 + \nu_t\end{aligned}\tag{1}$$

where μ is a positive constant, the inclusion of which allows to remove the free term from the equation for the sake of volatility, and ϕ is an autoregressive parameter that determines the memory in its relation to volatility. The main properties of autoregressive stochastic volatility models (ARSV) are investigated in [12,13].

The stochastic volatility model can be represented as follows:

$$x(k) = \psi_0 + \psi_1 x(k-1) + y(k) \quad (2)$$

$$y(k) = \sqrt{h(k)u(k)}, u(k) \approx N(0,1) \quad (3)$$

$$\log h(k+1) = \alpha + \phi \log h(k) + \eta(k), \eta(k) \approx N(0, \sigma^2) \quad (4)$$

where $x(k)$ is the statistical time series on which the model is based; $u(k)$ and $\eta(k)$ are independent of white noise stochastic processes. In some formulations, it is assumed that ψ_0 and ψ_1 are zero. The AR process (1) with the innovation $y(k)$ in the time series, as determined by the equation (2), explains the possibility of autoregression in the process $x(k)$.

Threshold stochastic volatility model. Dispersion of the incomes tends to increase with the decrease of share prices. Such dispersion behavior can be described using a constant correlation coefficient ρ between $u(k)$ and $\eta(k)$, keeping all other assumptions unchanged. In the initial model defined by the equations (2) – (4), it is zero. Numerous empirical studies have shown that the coefficients ρ are negative in the assumption that negative income is associated with a positive dispersion.

This investigation suggests a new approach to fixing the time series dispersion asymmetry. Since it has been established that the dispersion tends to grow under the influence of bad news (disappointing global forecasts), then it is likely that the dynamics of autoregression in the equation (4) is determined by the income in a previous period of time. There is a hypothesis that the amount of income is dependent on the prior income (income sign). This kind of income asymmetry can also be taken into account, summing up the equation (2) to the piece-linear structure. Thus, it will be more natural to consider the threshold nonlinear structures than the linear autoregressive processes represented by the equations (2) – (4).

Let's define a set of Bernoulli random variables as follows:

$$s(k) = \begin{cases} 0, & \text{if } x(k-1) < 0, \\ 1, & \text{if } x(k-2) \geq 0. \end{cases} \quad (5)$$

The threshold stochastic volatility model takes the following form:

$$x(k) = \psi_{0s(k)} + \psi_{1s(k)}x(k-1) + y(k) \quad (6)$$

$$y(k) = \sqrt{h(k)u(k)}, \quad u(k) \sim N(0,1) \quad (7)$$

$$\log h(k+1) = \alpha_{s_k+1} + \phi_{s_k+1} \log h(k) + \eta(k), \quad \eta(k) \sim N(0, \sigma^2) \quad (8)$$

Like in the initial formulation, $u(k)$ and $\eta(k)$ are stochastically independent. At a time $k-1$, when there is an unexpected fall in prices due to the disappointing news, $x(k-1) < 0$ and $s(k) = 0$. On the contrary, if there is good news at the time $k-1$, then $x(k-1) > 0$ and $s(k) = 1$. Therefore, the value $s(k)$ is determined by a magnitude sign $x(k-1)$. In the threshold stochastic volatility model, the values of the parameters ψ_0, ψ_1, α and ϕ switch between these two modes, which corresponds to the asset prices increases and falls.

In the symmetric case, the two sets of parameters are identical. In particular, if $\phi_0 = \phi_1$, taking into account that $\alpha_0 \geq \alpha_1$, it follows that the dispersion will be higher when the prior income is negative, than when it is positive.

The generalized model ϕ_0 may differ from ϕ_1 . Indeed, the coefficient ϕ_{s_t} is describing the magnitude of the impact of prior income on the current dispersion. If ϕ_0 is more than ϕ_1 , then the dispersion in previous periods will have a greater impact on the current dispersion after the price fall than after its rise. It is expected that in a similar hypothetical market situation, more time will be needed to “handle” the negative information that is contained in the previous dispersion data. This kind of asymmetry has not yet been sufficiently described in the researches related to the stochastic volatility analysis.

IV. BAYESIAN ANALYSIS OF THE THRESHOLD STOCHASTIC VOLATILITY MODEL

In the standard Bayesian conclusion, marginal posterior distributions of unknown parameters are used. However, in many cases, the common posterior distribution or even marginal posterior distribution do not have closed-form solutions. It is also quite difficult to obtain the model values from the desirable posterior distribution.

Monte Carlo method-based approaches for Markov chains are Markov update algorithms aimed at obtaining a sample from the common posterior distribution. A separate case is Gibbs sampling. This method based on the Monte Carlo procedure is close to the approach based on the generation and reproduction of the data samples.

Let's make a sample selection of the unified distribution $F(\omega_1, \dots, \omega_m)$ where $\omega = (\omega_1, \dots, \omega_m)^T$ is a vector of unknown parameters or hidden variables. With the known initial values $[\omega_1^{(0)}, \dots, \omega_m^{(0)}]$, the algorithm gives an estimate of the value $\omega_1^{(i)}$ with $F(\omega_1 | \omega_2^{(i-1)}, \dots, \omega_m^{(i-1)})$, value $\omega_2^{(i)}$ with $F(\omega_2 | \omega_1^{(i)}, \omega_3^{(i-1)}, \dots, \omega_m^{(i-1)})$, and value $\omega_m^{(i)}$ with $F(\omega_m | \omega_1^{(i-1)}, \dots, \omega_{m-1}^{(i-1)})$, considering that $i = 1, \dots, M + N$. Under non-rigid conditions, the vector of parameters

$[\omega_1^{(i)}, \dots, \omega_m^{(i)}]$ coincides with the distribution to the shared distribution $F(\omega_1, \dots, \omega_m)$ with $i \rightarrow \infty$.

Usually, the first M of the transitive iterations are skipped, and the last N of the iterations are taken for an approximate sample selection, dependent on $F(\omega_1, \dots, \omega_m)$. The density of the probability distribution can be performed in two different ways.

The first approach is the traditional assessment of core density distribution. The second approach is shown by the formula:

$$f(\omega_j) = \frac{1}{N} \sum_{i=M+1}^{M+N} f(\omega_j | \omega_{-j}^{(i)}) \quad (9)$$

where ω_{-j} means all the parameters except ω_j . The point estimates of any function ω , e.g. $g(\omega)$, can also be found using the Gibbs variable selection. One of the commonly used approaches is to use the posterior average, i.e.

$$\frac{1}{N} \sum_{i=M+1}^{M+N} g(\omega^{(i)}) \quad (10)$$

The Gibbs sampler (sampling) can be used if it is possible to make a model with all conditional distributions. Gibbs specification and other models based on the Markov chain Monte Carlo (MCMC) methods are investigated in [15,16].

Let's make a sample selection and calculate the distribution parameters following the given MCMC algorithm:

- (1) Calculate $\rho(k)$, $k = 1, \dots, n$.
- (2) Calculate σ^2 following the Metropolis step model (random samples).
- (3) Calculate $h(k)$, $k = 1, \dots, n$, using multi-stage discretization.
- (4) Calculate $(\psi_0, \delta, \psi_1, c)^T$ using its multidimensional normal complete conditional distribution.
- (5) Calculate x_0 using its normal full conditional distribution.
- (6) Calculate $(\alpha, \gamma, \phi, d)^T$ using its multidimensional normal complete conditional distribution.

The completion of the algorithm iteration using the MCMC method.

Dispersion smoothing and forecasting. To implement the procedure of the dispersion smoothing and forecasting, the Gibbs sample selection is used. After performing the iteration required for Gibbs sampling, one can get the approximation (approximate sample) from the common posterior distribution $f(x_0, H_n, \theta | X_n)$, marked as $(x_0^{(i)}, h_1^{(i)}, \dots, h_n^{(i)}, \theta^{(i)})$, $i = M + 1, \dots, M + N$. Smoothed estimates $h(k)$, ($k = 1, \dots, n$) are the h_i estimates calculated

from the marginal posterior distribution $f(h(k) | X_n)$. The natural choice is the marginal posterior expectation, $E(h(k) | X_n)$ which can be estimated as a sample mean:

$$\frac{1}{N} \sum_{i=M+1}^{M+N} h(k)^{(i)} \quad (11)$$

To predict the future dispersion, using the currently available information, it is necessary to generate the samples from $f(h_{n+j} | X_n)$ with $j > 0$. This can be effectively done using the composition method. Thus, when $j = 1$ it can be written:

$$\begin{aligned} f(h_{n+1} | X_n) &= \\ &= \int f(h_{n+1} | h_n, \theta, X_n) f(h_n, \theta | X_n) d(h_n, \theta) = \\ &= \int f(h_{n+1} | h_n, \theta, s_{n+1}) f(h_n, \theta | X_n) d(h_n, \theta). \end{aligned} \quad (12)$$

Therefore, $h_{n+1}^{(2)}$, $i = M + 1, \dots, M + N$ as an approximate sample of $f(h_{n+1} | X_n)$, is modeled using the log-normal distribution density $f(h_{n+1} | h_n^{(i)}, \theta^{(i)}, s_{n+1})$. Using this sample, the estimates of the distribution density as the h_{n+1} point estimates can be formed. This approach is generalized to calculate a multi-year forecast.

It should be taken into account, that $x(n+1)^{(i)}$, $i = M + 1, \dots, M + N$ calculated during the step 1, are model values with $f(x(n+1) | X_n)$. If it is necessary to have an extreme and percentile p -th forecast, let's say, with $p = 1$ for estimating the Value-at-Risk (VaR) value on financial markets, then the sample $x(n+1)^{(i)}$ will provide the choice from the p -th empirical percentile. Obviously, as soon as the $h_{n+j}^{(i)}$ from the distribution $f(h_{n+j} | X_n)$ is known, the value of the multi-year VaR forecast can be also calculated.

V. EXAMPLE OF STOCHASTIC VOLATILITY MODELS USAGE

As an example of the stochastic volatility usage, the following model can be presented. A model of the stochastic volatility can be used to make a formal description of the mental dispersion on the market in case of financial risk estimation. That can be done in the following way.

$$\begin{cases} y(k) | h(k) = e^{\frac{1}{2}h(k)} \varepsilon(k), & \varepsilon(k) \sim N(0, 1) \\ h(k) | h(k-1), \mu, \phi, \tau^2 = \mu + \phi(h(k-1) - \mu) + \\ + v(k), & v(k) \sim N(0, \tau^2) \end{cases} \quad (13)$$

To calculate potential losses, the Value-at-Risk (VaR) method can be used. It is based on the estimates of the exchange rate volatility, calculated on the basis of the reciprocal stochastic volatility model. In general, a different kind of model, which complies with the suitability and adequacy of the process, can be used. To investigate the volatility estimate methodology and possible VaR losses,

other selections, generated with the MCMC procedure, can be used. In this case, the posterior VaR mean value can be calculated by the equation:

$$E[VaR_{\pi}(N+1)|\mathbf{r}] = \frac{1}{M} \sum_{l=1}^M VaR_{\pi}^{(l)}(k) \quad (14)$$

where N is a number of values, which were actually used from the amount M generated through the MCMC procedure; \mathbf{r} represents the available measurements of the key variable of the investigated process; $VaR_{\pi}^{(l)}(N+1)$ is the value of $VaR_{\pi}(k)$ with l -th iteration of the MCMC procedure, which is calculated by the equation:

$$VaR_{\pi}^{(l)}(k) = \left(h^{(l)}(k)\right)^{1/2} \xi_{\pi}^{(l)}(k) \quad (15)$$

where $\xi_{\pi}^{(l)}(k)$ is the quantile of the generated distribution

$$\hat{f}^{(l)}[\xi(k)] = \sum_{s=1}^{s^*(l)} \frac{n_s^{(l)}}{N} \Phi[\xi(k) | \mu_s^{(l)}, \sigma_s^{2(l)}]$$

The calculated sample of values $VaR_{\pi}^{(l)}(k)$, $l=1, 2, \dots, M$ gives an opportunity to find the median value and the Bayesian tolerance intervals using the quantile of the distribution.

This example shows how a nonexistent (heteroscedastic) process is generated with the following model:

$$\begin{aligned} h(k) &= \alpha_0 + \alpha_1 r^2(k-1) + \alpha_2 r^2(k-2) + \\ &+ \beta_1 h(k-1) + \beta_2 h(k-2) + \xi(k) = \\ &= 0,032 + 0,23 r^2(k-1) - 0,095 r^2(k-2) + \\ &+ 0,67 h(k-1) - 0,15 h(k-2) + \xi(k) \end{aligned}$$

To describe the key variable, the following equation is used:

$$y(k)|h(k) = \exp(-0,5h(k)) \varepsilon(k), \quad \varepsilon(k) \sim N(\bar{\varepsilon}, 1).$$

The generation of an innovative process (random variables) was performed using this combination of normal distributions:

$$\xi(k) \sim 0,75 N(0,11; 0,55) + 0,25 N(-0,95; 2,25)$$

According to the MCMC algorithm, the sequences of the pseudorandomized numbers in the overall amount of values (in 20000 algorithm iterations) was generated to estimate the parameters of the model. The first 10,000 values were not further examined because they refer to the transitional stage of the estimation process, during which the chosen data generating method is configured. It means that 10000 values were actually used. Considering that

$\bar{\varepsilon} = E[\varepsilon(k)] = \alpha_0^{-1/2} E[\xi(k)]$, the posterior distribution for $\bar{\varepsilon}$ can be calculated as follows:

$$\bar{\varepsilon}^{(l)} = \frac{1}{(\hat{\alpha}^{(l)})^{1/2}} \sum_{s=1}^p \mu_s^{(l)}, \quad l=1, \dots, M$$

where p is the number of distributions used to generate the mixture.

The results of imitational modeling of the non-stationary heteroscedastic process and the estimation of the mathematical model parameters are given in the Table I. It contains posterior average estimates of the model parameters calculated using the proposed method. For purposes of comparison, it also includes the values of estimates calculated using a simple Gaussian distribution of the innovative random process.

TABLE I. RESULTS OF IMITATIONAL MODELING OF THE NON-STATIONARY HETEROSCEDASTIC PROCESS

№ n/n	Parameter and its value		Suggested method	Gaussian distribution
	Parameter	True value	Posterior average	Posterior average
1	α_0	0,032	0,0297 (7,2%)	0,0355 (10,94%)
2	α_1	0,230	0,245 (6,5%)	0,258 (12,21%)
3	α_2	-0,095	-0,0877 (7,7%)	-0,0998 (5,05%)
4	β_1	0,670	0,6581 (1,8%)	0,749 (11,79%)
5	β_2	-0,150	-0,163 (8,7%)	-0,132 (12,01%)
6	$\bar{\varepsilon}$	0,0095	0,0078 (17,9%)	0,0114 (20,0%)
7	Average error of estimate %		8,30%	12,01%

The percentage in parentheses indicates the average estimate errors related to the exact values of the given model. It is evident that the parameter estimates, calculated according to the suggested method, are much closer to the true values. Thus, the average estimate error in percentages is 8.3% and 12.01% accordingly. I.e. the estimate errors decreased 1.5 times approximately. Consequently, the alternative method, chosen for the comparison, allows getting the estimates, close to the true values of the used model.

VI. CONCLUSION

Investigations related to the probabilistic-statistical volatility modeling are highly important due to the necessity of the forecasts estimates quality improvement and the decisions taken on their basis. Therefore, a particular consideration is given to the method of Bayesian analysis of non-stationary (heteroscedastic) processes, which are widely distributed in various spheres of human life. The development of the volatility estimation methods based on the Bayesian analysis allows to significantly improve the quality of forecasts and their estimation.

The usage of various modifications of Monte Carlo method-based approaches for Markov chains makes it

possible to correctly solve the issues of mathematical model parameters estimation within the complex structures, provided that there are random influences with arbitrary distributions. Further investigation of the Bayesian analysis may be aimed at improving the methods of estimating the parameters of various probabilistic and statistical volatility models with the help of adaptive estimation schemes; expanding the criteria basis for analyzing the quality of intermediate and final results; building specialized decision support systems for the analysis of nonlinear non-stationary processes in order to take substantiated financial and economic decisions. In particular, this applies to the current systems of risk management, analysis of the price formation processes on stock exchanges, investments and economical diagnostic.

REFERENCES

- [1] R. Engle, "Autoregressive conditional heteroscedasticity with estimates of variance of united kingdom inflation," *Econometrica*. vol. 50. pp. 987–1008. 1982.
- [2] R. Engle, and T. Bollerslev, "Modelling the persistence of conditional variances," *Econometric Reviews*. vol. 5, no. 1, pp. 1–50. 1986.
- [3] M. Asai, M. McAleer, and Jun. Yu, "Multivariate stochastic volatility: areview". *Econometric Reviews*, vol. 25, pp.145-75, 2006.
- [4] M. Asai, and M. McAleer, "The structure of dynamic correlations in multi variate stochastic volatility models," *Journal of Econometrics*, vol.150, pp.182-192, 2009.
- [5] T. Bollerslev, "Generalized autoregressive conditional heteroskedasticity". *Journal of Econometrics*. vol. 31, no. 3. pp. 307–327, 1986.
- [6] T. Bollerslev, "A conditionally heteroskedastic time series model for speculative prices and rates of return," *The Review of Economics and Statistics*. vol. 69, no. 3. pp. 542–547, 1987.
- [7] E. Jacquier, N. G. Polson, and P. E. Rossi. "Bayesian analysis of stochastic volatility models," *Journal of Business and Economic Statistics*, 20, pp.69-87, 1994.
- [8] E. Jacquier, N. G. Polson, and P. E. Rossi. "Bayesian analysis of stochastic volatility models with fat-tails and correlated errors," *Journal of Econometrics*, vol.122, pp.185-212, 2004.
- [9] N. Shephard, *Stochastic Volatility: Selected Readings*. Oxford: University Press. 2005.
- [10] N. Shephard, and T. A. Andersen, *Stochastic Volatility: Origins and Overview*. Handbook of Financial Time Series. New York: Springer. 2009.
- [11] T. Bollerslev, R. Chou, K. Kroner, "Arch modeling in finance : A review of the theory and empirical evidence," *Journal of Econometrics*. vol. 52, no. 1-2, pp. 5–59, 1992.
- [12] T. Andersen, "Stochastic autoregressive volatility: a framework for volatility modeling," *Mathematical Finance*. vol. 4, pp. 75– 102, 1994.
- [13] S. Taylor, "Modeling stochastic volatility: A review and comparative study," *Mathematical Finance*. Vol. 4, no. 2, pp. 183–204, 1994.
- [14] Capobianco, E. State-spaces to chastic volatility models: areview of estimation algorithms. *Applied Stochastic Models and Data Analysis*. vol. 12, pp. 265–279, 1996.
- [15] T. Andersen, T. Bollerslev, S. Lange, "Forecasting financial market volatility: Sample frequency visa-vis forecast horizon," *Journal of Empirical Finance*. vol. 6, no. 5, pp. 457–477, 1999.
- [16] S. Babichev, V. Lytvynenko, M. Korobchynskiy, and M.A. Taiff, "Objective clustering inductive technology of gene expression sequences features," *Communications in Computer and Information Science*, 716, pp. 359-372, 2017

Choosing a Programming Language for a New Project from a Code Quality Perspective

Max Garkavtsev
QArea
Switzerland
president@qarea.com

Nataliya Lamonova
Diligences Inc.
Kharkiv, Ukraine
lamonova@diligences.com

Alexander Gostev
Diligences Inc.
Kharkiv, Ukraine
gostev@diligences.com

Abstract—Those just starting a new project, let alone setting up a business of their own, often find themselves facing quite a few problems. The first one, more often than not, is choosing the best technology to do it. With this article, we will try to help those who have taken the first decisive step to make the right choice. We are going to analyze various aspects of programming languages and draw conclusions on their applicability, which will be of use to small teams and startups creating their software business from scratch. Our defectiveness forecast can help to adjust timeline of a product (project) delivery. As well as helping to estimate costs for fixing non-functional issues in the code. This information helps to address product (project) delivery risks at different levels like financial, project, software development. With the article we are only giving an additional angle for analysis, not a full-scale framework for choosing a programming language.

Keywords—*Technical debt, software quality, source code analysis, defectiveness, automatic static analysis*

I. INTRODUCTION

At the early stages of developing a business, one is to construct a stable basis. For any project involving software solutions, this basis boils down to the quality of code as, technically, this is one aspect on which the viability of an entire business depends. We have deliberately focused on analyzing a small user repository. Having collected and analyzed open source code, we have been able to add a new dimension to the decision making process – code defectiveness in the programming language.

When choosing a language for the project, people are mostly guided by subjective criteria without taking into account certain lesser known facts. We are going to try, through concrete figures and analysis, to enhance your ability to choose the best tool to suit your project.

Undoubtedly, there are no dominant criteria that should serve as major guides for the choice, such as project objectives, its size and type, complexity, availability of specialists, the possibility of development and further support of the developed product, to name just a few.

Having all relevant information at one's disposal, one can make an objective choice, which in the future will help save the most important thing – one's time.

We are not discussing the classification of programmers. It is quite an extensive and interesting subject and, since it borders on psychology, deserves special attention [1,2], which goes beyond the scope of our current investigation.

Consequently, the subject of discussion in this article is the programming language as a tool, while the mastery of it is a secondary issue and rests solely on the budget the company or the project can afford in choosing between experienced and junior developers.

II. GOALS

Language should be regarded as a tool for increasing the productivity of a team.

When choosing a programming language, or a multiple languages for your project, it is necessary to be guided by objective criteria with an eye to the tasks of the project.

The larger the project, the larger the stack of technologies that it uses. A number of programming languages are often used in large projects because each individual language was designed to solve a specific range of problems.

In order to evaluate the scope of developing a software product with a particular programming language, an effective indicator is usually applied - the number of lines of code (Source Lines of Code — SLOC [3,4,5]).

Since we want to have a look at what is most common and popular, we have selected the following languages for analysis: PHP, Python, Java, C#, JavaScript.

Metrics, naturally, represent statistical data and describe the quality of the program code written according to the guidelines of best practices in programming and productivity, which in turn gives us some idea as to the potential development of the project. Such metrics can be of service to project managers and team leads in assessing the complexity of an already completed project or the amount of work for a future project, as well as provide an understanding of operational disruption risks associated with the stylistic features of source code.

III. CASE STUDY

To collect information, GitHub was used - the largest web-service for hosting IT projects and their cooperative development. Among the large number of static code analyzers, open source platform SonarQube [6] was chosen, which was designed for continuous analyzing and measuring code quality. The scanner was improved by a few modifications and the final version of the code analyzer allows us to identify technical debt in source code, namely modularity violations, code smells (errors), and automatic static analysis for each language with a different scanner for each language.

A. Data Collection

Data set consists of about 191400 GitHub profiles: C# - 17900 profiles; Java – 64500 profiles; JavaScript – 50000 profiles; PHP - 31000 profiles; Python – 28000 profiles. While we were collecting data, our intention was to focus on programmers who have one project, one committer and who use only one programming language. Repositories for scan were selected in a way to exclude 3-rd party code that can dilute personal output. Repository size is understood as the mass of code (in kilobytes) which is found in the user’s repository.

B. Metrics

Risks linked to the quality of source code could harm software business viability that's in concern of code maintainability and ability to be updated at a reasonable cost. A negative scenario could make the business lose control over own software, causing it to be unable to keep the software up to date or even to go bankrupt for technical reasons. Thus we treat this risk as a high severity hazard. As an example, the number of non-functional defects found influences the risk of a late project delivery. We are talking of “technical debt” here. During a project’s life cycle, situations inevitably arise where non-optimal technical or strategic decisions have to be made which will later be paid for in technical debt. Alongside that, there exists the notion of unclean code which results in technical debt that increases over time if not refactored into clean code.

The metric which ensures the proper assessment of potential risks to do with technical debt is one revealing language cleanliness. It shows which of the languages is the “cleanest” in relation to the technical debt created as per the lines of code.

Fig.1 reflects probabilistic risks which each of the languages analyzed is subject to. Significantly standing out are Python and C#. We can conclude that choosing either of those two is less likely to cause problems later.

Another metric - code defectiveness (all violations/SLOC)—is another really important characteristic of code quality. It is measured as the ratio of all defects to the number of lines of code. This information can serve as a reference for both developers and managers trying to assess productivity.

What we are looking into here is the number of issues present in code depending on the language of coding as well as on the impact of the code issues on system operation (severity). A code issue is an error, flaw, failure or fault in code or whatever causes it to produce an incorrect or unexpected result, or to behave in unintended ways [7]. Different levels of severity are distinguished: blocker, critical, major, minor, trivial.

The average values for defectiveness in each of the languages look as follows:

- C#: Average = 0.004,
- Java: Average =0.106,
- JavaScript: Average = 0.077,
- PHP: Average = 0.049,
- Python: Average = 0.066.

Visualizing a different approach to the analysis of language defectiveness reveals how defectiveness is spread in the range of each repository size (Fig. 2).

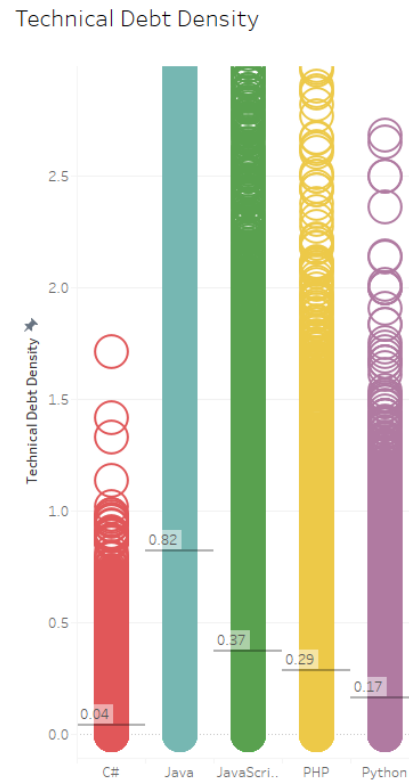


Fig. 1. Technical Debt Density among all languages

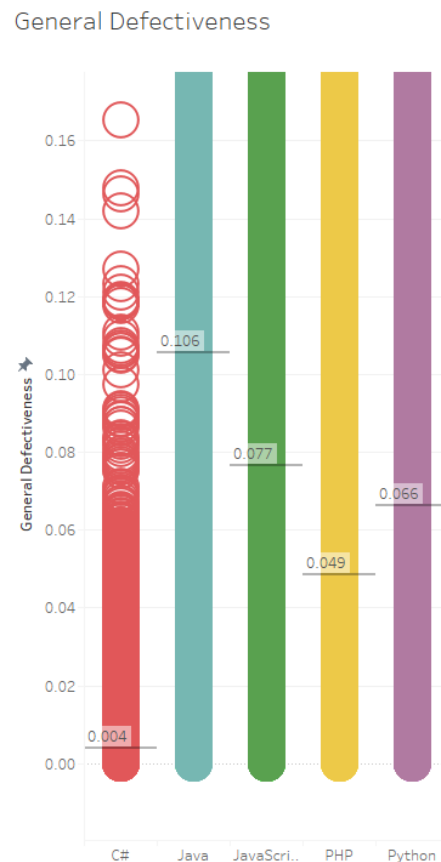


Fig.2. Average of general Defectiveness among all languages

C. Clustering results

For more detailed analysis and uncovering data with similar characteristics in our data we used clustering techniques and particular - the k-means algorithm. On the basis of the analysis the following can be said on each particular language:

- C# shows defectiveness within 0.2 in all profiles spread equally in repositories of up to 140 000 SLOC (Fig.3 a)) and forms 3 trend clusters. Each cluster shows its own tendency depending on the ratio between defectiveness and lines of code.
- Java and JavaScript possess defectiveness up to 5 (Fig.3 b), 3 c)) and form 2 clusters each.
- PHP defectiveness, spreads over the entire range of repository sizes and doesn't exceed 2 (Fig.3 d)) with 2 clusters of the same tendency.
- Python demonstrates 5 different tendencies for profiles up to 400 000 SLOC (Fig.3 c)).

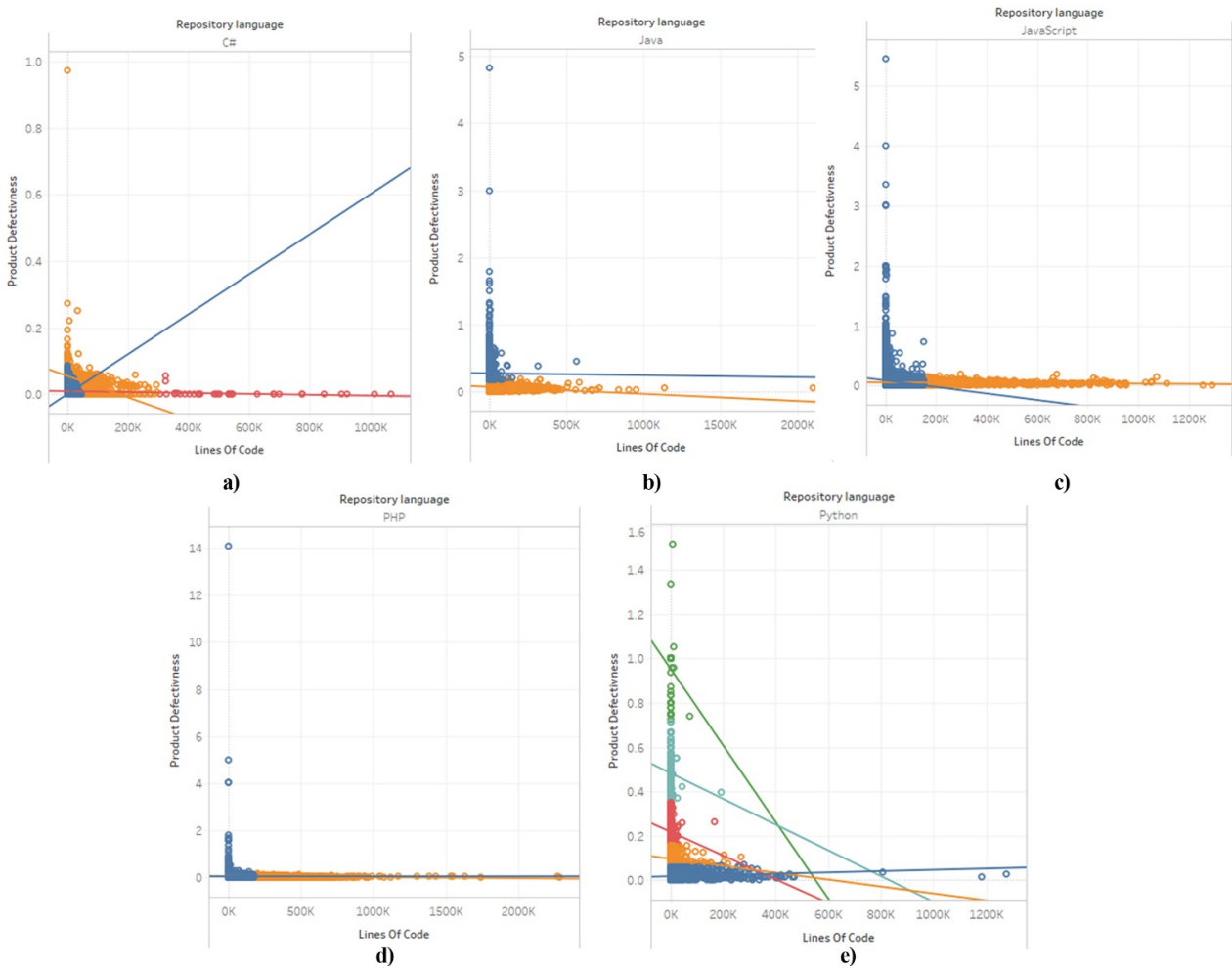


Fig. 3: Spread of defects in a) C#, b) Java, c) JavaScript, d) PHP and e) Python as per the number of lines in code

Each language has shown a different number of clusters and different tendencies within the clusters. For example, for C# one can identify 3 clusters with their own trend line for different number of SLOC. A language that prominently stands out against the general background is Python. It has 5 clusters.

D. Business application.

The next step in applying our findings to a business context is to add numerical value of technical debt [8] in minutes to the defects that were found. Counting average Technical Debt Density makes us able to predict amount of time that's necessary for fixing the non-functional defects. Adding hourly pay rate of a software developer in the U.S. [9] enables us to predict budget overrun (Table I).

TABLE I. PER SLOC ESTIMATE FOR PROJECT BUDGET AND PROJECT DELIVERY OVERRUN

Expenses	Repository language				
	C#	Java	JavaScript	PHP	Python
Delay in product delivery (min. per SLOC)	0.04	0.82	0.37	0.29	0.17
Additional costs for fixing code issues (\$ per SLOC)	0.03	0.032	0.03	0.027	0.033

Applying above-mentioned numbers to a potential software project that is of a comparable size to Windows Vista, MS Visual Studio 2012 or Large Hadron Collider systems [10] commands actual numbers that are reflected in the Table II.

TABLE II. ESTIMATE FOR PROJECT BUDGET AND PROJECT DELIVERY OVERRUN FOR 50 MLN. LOC CODEBASE

Expenses	Repository language				
	<i>C#</i>	<i>Java</i>	<i>Java Script</i>	<i>PHP</i>	<i>Python</i>
Delay in product delivery (man/hours.)	33,333	683,333	308,333	241,667	141,667
Additional costs for fixing code issues (\$)	1,500,000	32,800,000	13,875,000	9,666,666	7,012,500

IV. CONCLUSION

Since we set out with this article to try and help anyone choose a programming language, we have reviewed five most popular languages.

Analysis shows that we cannot escape the fact that each language is in some degree subject to defectiveness. Consequently, expectations will vary as to any further steps necessary to rectify defects stemming from that.

Having results on hand for a 50 000 000 SLOC project written in Java we suggest to book ahead \$32,800,000 and 683,333 man/hour on fixing non-functional defects. If the project is written on C# the amount of resources that should be allocated is \$1,500 000 and 33,333 man/hour.

Whenever complex projects are concerned, the use of stack technologies can be recommended. It must be made clear that they are done on frameworks and clean programming languages, which were used to create the

world's biggest sites, for instance: PHP - Facebook; Python - Instagram; Java -Ebay, Amazon; C# -StackOverflow; JavaScript - LinkedIn.

Finally, we would like to voice a note of caution in that the article's major metric, defectiveness, is but an additional factor to be looked at in deciding on the programming language. Please, bear in mind that before you fully commit to any one technology, you will have to carefully consider the objective criteria and aims of the project itself.

The scanners that were used for finding non-functional defects in code carry differences in implementation as each particular language has a unique structure, semantics, underlying logic etc. Thus we don't recommend directly comparing the results obtained between languages. The point of this article is enable the reader to book possible unforeseen expenses and schedule overrun beforehand.

The next step of the research is to get from the project level analysis into analyzing the personal commitment of a single software developer so as to be able to make conclusions with a higher accuracy and controlled variance.

REFERENCES

- [1] Frederick P. Brooks Jr. , The Mythical Man-Month. Addison-Wesley 1995 [1975].
- [2] Jump up^ Maggie Fox NBC News, October 21, 2013, Better use the phone: Why Obamacare website is such a fail. Accessed Oct 21, 2013
- [3] Overview of COCOMO model - <http://www.softstarsystems.com/overview.htm>
- [4] Lines Of Code on C2 wiki - <http://wiki.c2.com/?LinesOfCode>
- [5] Source lines of code - https://en.wikipedia.org/wiki/Source_lines_of_code
- [6] SonarQube: Continuous Code Quality - <https://www.sonarqube.org/>
- [7] Software bug: https://en.wikipedia.org/wiki/Software_bug
- [8] https://resources.sei.cmu.edu/asset_files/WhitePaper/2012_019_001_58818.pdf
- [9] <https://insights.stackoverflow.com/survey/2017#technology>
- [10] <https://informationisbeautiful.net/visualizations/million-lines-of-code/>

Data Mining of Network Events with Space-Time Cube Application

Viktor Putrenko
World Data Center for Geoinformatics
and Sustainable Development
Igor Sikorsky KPI
Kyiv, Ukraine
putrenko@wdc.org.ua

Nataliia Pashynska
Department of Intellectual and
Information Systems
Taras Shevchenko National University of
Kyiv
Kyiv, Ukraine
n_pashynska@ukr.net

Sergiy Nazarenko
World Data Center for Geoinformatics
and Sustainable Development
Igor Sikorsky KPI
Kyiv, Ukraine
nazarenko@wdc.org.ua

Abstract— The scientific and practical principles of data analysis are proposed with the help of methodology of space-time cube construction as one of the types of data mining with spatial-temporal distribution. The use of this method on the example of information analysis from subscribers of one of the major mobile operator networks allows to carry out statistical analysis and to detect statistically significant spatio-temporal clusters in the data that can be used during data structuring in order to provide safety and react quickly to hazardous situations.

Keywords— data mining, spatial-temporal cube, national security, GIS, data visualization

I. INTRODUCTION

Starting from 2013, Ukraine's national security becomes more and more important. Every day, citizens of Ukraine meet with the most diverse threats of natural, technological, social and military character. Dangerous processes, extreme events, catastrophes, virtual and real terrorism, and so on - these are exactly the things to which public authorities need to react practically every day. In addition, rapid response to an emergency significantly increases the chances of reducing the number of victims or other negative consequences, ranging from a specific person to the size of the state.

The modern world is extremely rich in the most diverse information, the vast arrays of which people collect, store, analyze, and on the basis of it tries to make forecasts and predictions. This is especially true for information that directly and indirectly affects human security. Equally important information for government bodies, whether public or private institutions, the correct analysis of which allows you to take a step in the right (progressive) direction. But operating big data, which are dispersed not only in a large area, but also in a large time interval, usually make it difficult to make the right decision, or significantly increase the time for its adoption.

One of the main issues of our time is "security". Often, for which you need to make a quick and balanced decision. Especially when it comes to eliminating the consequences of a certain disaster, where the bill goes, sometimes, for a minute. For example, international statistics show that the number of rescued after the earthquake directly depends on the beginning of rescue operations. If the saviors arrive in the earthquake zone in the first three hours, they can save up to 90% of the survivors, in six hours - 50%. In the future, the chances of salvation are greatly reduced. Only by means of

rapid response can reduce the number of victims by 20-30 percent [1-2]. That is why it is important to get information and to give an answer immediately after an emergency, than after a while, and moreover, to prevent an emergency in advance, which can lead to a significant number of victims.

II. BACKGROUND

Recently, security issues are becoming increasingly significant, due to the increasing number of threats to ordinary people and the region or the country as a whole. One of the options for solving this issue is to study, analyze and forecast the event by building a spatial-temporal cube. For the first time, the use of the space-time cube was proposed by T. Hägerstrand in the early 70's, [3] whose possibilities he described in his work "What about people in regional science?". But the active development of geographic information systems (further GIS), its use was limited. Only in the 2000s there are work on the use of the spatial-temporal cube in GIS. In the works of this period, new possibilities for using the spatial-temporal cube were presented using GIS, including earthquake surveys [4-7].

The next steps in using the spatial-temporal cube method were its application in the intelligent analysis of data of a variety of nature: crime analysis, infrastructure studies, animal behavior analysis, human motion visualization, dependence studies on weather conditions changes over time, and much more [8-13]. In the field of data mining in Ukraine widely known of the Institute for Applied System Analysis NTUU "KPI", World Data Center for Geoinformatics and Sustainable Development [14-18].

III. GOAL AND TASKS

The goal of the work is to analyze the spatial-temporal regularities in the distribution of events in the Vodafone network based on the use of the methodology of space-time cube construction.

The tasks are:

- to study the methodology of using the space-time cube for the data mining of spatial-temporal data;
- study of the application peculiarities of the space-time cube construction method for the analysis of space-time series of data generated by users of Vodafone telecommunication network;

- the use of building space-time cube for distribution analysis of spatial and temporal patterns of mobile data for the purpose of emergency response to natural and social emergencies.

IV. SPATIAL-TIME CUBE

Spatial-time cube is a 3D visualization technology designed to simultaneously represent spatial and temporal characteristics of motion. Accordingly, trajectory points are displayed in three-dimensional space, where the vertical axis usually expresses time [19].

In the early 70's T. Hägerstrand [3] proposed the use of a graphical approach to reflecting time as an addition to spatial measurements. He developed a three-dimensional diagram as a spatio-temporal cube, which allows you to visually explore space-time events and processes interactions. The cube's base reflects a flat geographic dimension, and the cube's height is time. Initially, the tool was designed to manually reproduce graphics. In our time, there are several approaches to the automated construction of such models using the tools of modern GIS.

The use of the space-time cube requires spatial and temporal data, for the purpose of analyzing certain events. Examples of such events include earthquakes, road accidents, cases of disease or the observation of rare animals [6].

T. Hägerstrand proposed to apply the space-time cube to the data on the motion of objects on the changes of spatial sites with an anchor to time. In this paper, the authors propose to apply the concept of T. Hägerstrand to another type of data, namely, to analyze network events.

In addition, the use of the spatial-temporal cube makes it possible to answer 3 questions of Puke [5], supplied to spatial-temporal data [20]:

- when plus where → what: description of objects or a set of objects that are present in a certain place or a set of locations for a certain time or time interval;
- when plus what → where: a description of the location or set of locations occupied by a particular object or set of objects at a specific time or time interval;
- where plus what → when: a description of a specific time or interval of time when a particular object or set of objects occupied a particular place or set of locations.

In addition to the space-time cube, a number of other methods are also used to display the dynamics of events in time.

The Time Slices model is one of the first spatio-temporal data models. Its main features:

1. Storage of data at a regular interval of time.
2. Separate data sets for each specific time interval.
3. Time-dependent (cross-time) classification of data storage objects.

Such a model is convenient at the stage of transition from the spatial to the space-time model.

The model of time series. This is a model with a base state and subsequent changes. Unlike the model of temporary layers, only the basic state of objects and their changes are

stored, through irregular, in general, intervals of time. Thus, the time series model contains much less redundant data than the temporal layer model.

The research uses the data provided by Vodafone, which has spatial and temporal bindings, as well as some attribute information. The processed database has 1,5 million calls, messages and exits to the Internet, from the most diverse devices and from different subscribers. All "events" are concentrated practically in the western regions of Ukraine and has geographic coordinate system WGS84.

In order to achieve the goals, the authors use a set of tools for in-depth analysis of spatial and temporal regularities in the software ArcMap 10.5. This toolkit contains statistical tools for analyzing data distribution and identifying patterns in the context of space-time. The set includes tools: Create a Space-Time Cube and Analyze the emergence of hot spots.

The dataset structure has a combined set of attributes that characterize the nature of the communication event, location, feature of calls and devices, as well as subscriber preferences. The description of network events is the event type, which is divided into incoming-outgoing calls, SMS and Internet usage. The location is described by the direction and coordinates of the signal receiving station. The peculiarities of network events include the tariff plan, the category of numbers, the amount of Internet traffic, the cost of use and the type of device that is distributed to ordinary mobile phones and different types of smartphones. Personal preferences of the client are presented in the form of three attributes describing the interests located in the first, second and third place for the subscriber. As examples of such preferences are the categories of science, culture, tourism, travel, football, etc.

The creation of a spatio-temporal cube takes place by arranging point data of events in space and time in the form of a cubic structure, which is formed in a special netCDF format. The hot spot analysis tool uses a cubic structure to detect statistically significant trends over time. This type of analysis is well suited for studying offenses, outbreaks of infections, events in social networks.

The base unit of the cube is the bin of space-time (Fig.1), which counts the number of points in time and each location using the Mann-Kendall statistics. [20].

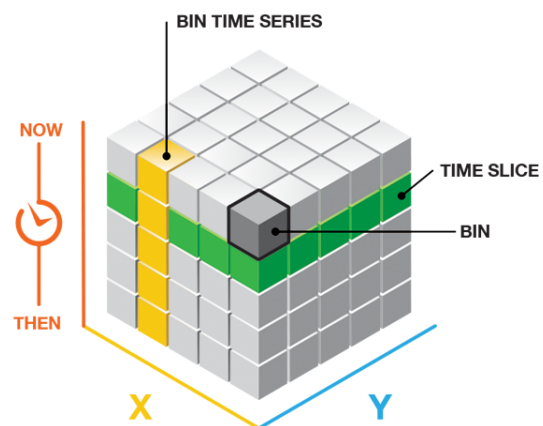


Fig. 1. Bin display in the space-time cube (Source: ESRI ArcMap [20])

The spatial-temporal cube consists of rows, columns, and time steps, which together form the total number of bins in the cube. The rows and columns correspond to the placement of objects in the latitude and longitude plane, and the cube height corresponds to the time period. If an event occurred for a certain period of time, it will be fixed in a certain bin with spatial-temporal characteristics.

In the presence of at least one spatial-temporal event creates a new bin of data. Bin without data gets a zero number of events, but can be saved in a common structure to maintain data continuity. Information on the number of such bins is given as the amount of discharge.

V. MANN-KENDALL TEST

As input objects there can be only point classes that describe the events that have taken place. Such events may include network events, emergencies, trade operations, and other events that are time-consuming and space-based. Time binding is done using the attribute in the Date format. The toolkit works in a range of events from 60 to 2 billion, which allows for sufficient flexibility in data processing. In order to obtain valid data of distances calculations, rectangular coordinate systems with corresponding projections are used.

An important part of the tool's operation is the analytical operations over the data bins used during the simulation. The basic set of operations is the definition of the general trend of data, which is calculated on the basis of time series. Using trend analysis allows you to determine the positive or negative trends in the number of events. The trend analysis is based on Mann-Kendall's statistics.

The non-parametric Mann-Kendall test is commonly employed to detect monotonic trends in series of data. The null hypothesis, H_0 , is that the data come from a population with independent realizations and are identically distributed. The alternative hypothesis, H_A , is that the data follow a monotonic trend. The Mann-Kendall test statistic is calculated according to:

$$S = \sum_{k=1}^{n-1} \sum_{j=k+1}^n \text{sgn}(X_j - X_k) \quad (1)$$

with

$$\text{sgn}(x) = \begin{cases} 1, & \text{if } x > 0, \\ 0, & \text{if } x = 0, \\ -1, & \text{if } x < 0. \end{cases} \quad (2)$$

The mean of S is $E[S] = 0$ and the variance σ^2 is

$$\sigma^2 = \frac{\left\{ n(n-1)(2n+5) - \sum_{j=1}^p t_j(t_j-1)(2t_j+5) \right\}}{18} \quad (3)$$

where p is the number of the tied groups in the data set and t_j is the number of data points in the j th tied group. The statistic S is approximately normal distributed provided that the following Z-transformation is employed:

$$Z = \begin{cases} \frac{S-1}{\sigma}, & \text{if } S > 0, \\ 0, & \text{if } S = 0, \\ \frac{S+1}{\sigma}, & \text{if } S < 0. \end{cases} \quad (4)$$

The statistic S is closely related to Kendall's τ as given by:

$$\tau = \frac{S}{D} \quad (5)$$

where

$$D = \left[\frac{1}{2}n(n-1) - \frac{1}{2} \sum_{j=1}^p t_j(t_j-1) \right]^{\frac{1}{2}} \left[\frac{1}{2}n(n-1) \right]^{\frac{1}{2}}. \quad (6)$$

The resulting Vodafone data set is in the time period from June 1, 2017 to August 31, 2017. For the convenience of analysis, the authors used a 5-day time step. As a result, the tool built a cube with a height of 19 bins (Fig.2).

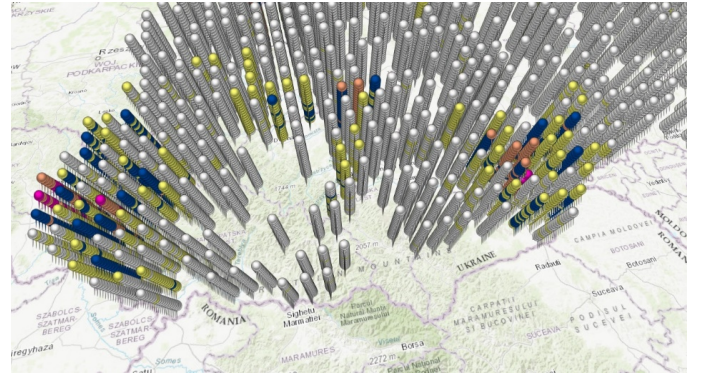


Fig. 2. 3D visualization of space-time cube for western regions of Ukraine

VI. HOT SPOTS ANALYSIS

Tool Analysis of hot spots determines trends in the cluster of density of points (calculations) or fields of sums in a cube. The categories of cold and hot spots include the following characteristics [21]: new, consistent, growing, constant, declining, sporadic, and fluctuating historical (Fig.3).

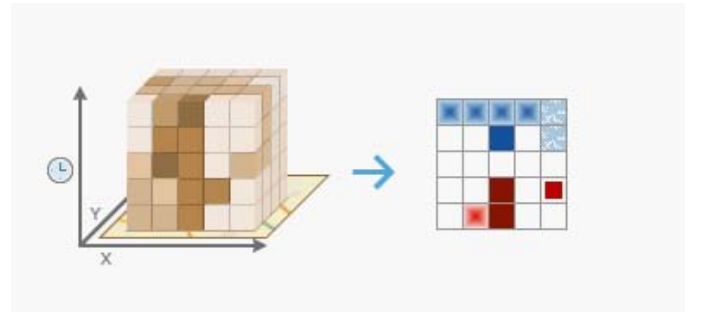


Fig. 3. Transformation of space-time cube for analysis through hot spots

Bin in the cube has the properties of position and time recording in a three-dimensional structure, which is written in the attributes LOCATION_ID, time_step_ID, COUNT. The same values of the spatial and temporal identifiers of the bins can be associated with the corresponding rows and

columns of the cube, which allows to support calculations by the individual segments of the cube. For each bin sum up the value of the number of events.

The tool for analysis of hot spots determines the variability in the input cube based on the application of the mathematical calculator of Getis-Ord G_i^* statistics. This calculation is made for each bin in the cube in relation to its neighbors.

The Hot Spot method calculates a statistic for each event in the data set. The final values of p (probability) and z -estimates (standard deviations) indicate that in what region of the space clustered events with high or low values [5]. The method works by analyzing each event in the context of the neighboring geography of events. To be a statistically significant hot spot, the event must have a high value and be surrounded by other approaches with also high values. The local amount for the event and its neighbors is proportional to the sum of all events; when the local amount is very different from the expected local amount, and if this discrepancy is too large to be the result of a random process, a statistically significant z -score is obtained. Hot dots statistics uses the formula:

$$G_i^* = \frac{\sum_{j=1}^n w_{i,j} x_j - \bar{X} \sum_{j=1}^n w_{i,j}}{S \sqrt{\frac{n \sum_{j=1}^n w_{i,j}^2 - \left(\sum_{j=1}^n w_{i,j} \right)^2}{n-1}}}, \quad (7)$$

$$\bar{X} = \frac{\sum_{j=1}^n x_j}{n}, \quad (8)$$

$$S = \sqrt{\frac{\sum_{j=1}^n x_j^2}{n} - (\bar{x})^2} \quad (9)$$

where x_j is the attributive value for the event j , $w_{i,j}$ is the spatial weight between the events i and j , n is the total number of events, \bar{X} is the mean of the arithmetic values of the course, S is the dispersion.

The statistical value G_i gives each object in the set its own z -score. If the z -score has a positive value, then the probability of the intensity of the clustering of hot spots increases, which is proportional to the size of the positive estimate. Negative z -values are directly proportional to the intensity of clustering of low values and correspond to cold points.

Output objects are added to the Table of Contents and represent a summary of the spatial-temporal analysis for all the analyzed locations. In addition to creating a class of Output objects, the summary analysis results are recorded in the Results window (Fig.4).

VII. LOCAL OUTLIER ANALYSIS

The analysis tools group includes the Local Outlier Analysis tool, which allows you to identify significant statistical data in both space and time. To determine statistically significant data outliers, the Anselin Local Moran I statistical is used statistic option, which calculates the value of each bin relative to its neighbors.

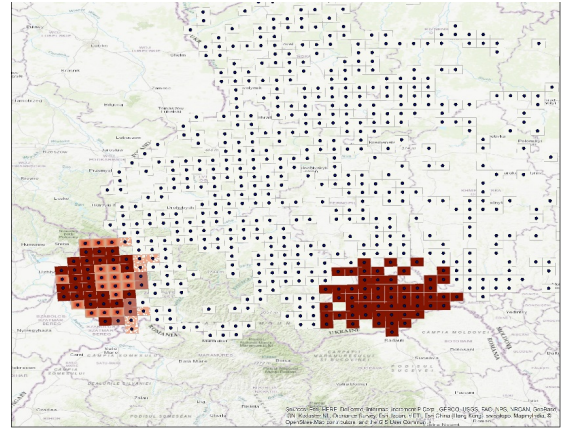


Fig. 4. Map occurrence of hot spots, resulting space-time analysis cube

The cluster analysis tool divides bins and sets of objects in the category of cluster allocation with high and low trend values. In this process, statistical outliers in spatial data are also determined. Based on the calculation of z and p values of Anselin Local Moran I statistics, each time series receives the coded value of belonging to a particular cluster with the corresponding statistical value. The local Moran's I statistic of spatial association is given as:

$$I_i = \frac{x_i - \bar{X}}{S_i^2} \sum_{j=1, j \neq i}^n w_{i,j} (x_j - \bar{X}) \quad (10)$$

where x_i is an attribute for feature i , \bar{X} is the mean of the corresponding attribute $w_{i,j}$ is the spatial weight between feature i and j , and:

$$S_i^2 = \frac{\sum_{j=1, j \neq i}^n (x_j - \bar{X})^2}{n-1} \quad (11)$$

with n equating to the total number of features.

The z_{I_i} -score for the statistics are computed as:

$$z_{I_i} = \frac{I_i - E[I_i]}{\sqrt{V[I_i]}} \quad (12)$$

where:

$$E[I_i] = - \frac{\sum_{j=1, j \neq i}^n w_{i,j}}{n-1} \quad (13)$$

$$V[I_i] = E[I_i^2] - E[I_i]^2 \quad (14)$$

The presence of positive evaluations for I is a certificate that is adjacent to objects with similar values that may be part of a cluster. Negative values indicate the difference between the estimates of the object and its neighbors. In all cases, the value of p for the object must be small so that the cluster is determined to be statistically significant.

To determine the belonging of the bin to the clusters, the rules of the conceptualization of spatial relationships are first defined, which determine the belonging of the bin to one of the clusters. Further, the values of bins are estimated in proximity to the center of the cluster.

Bins with high values of local emissions contain abnormal changes in the behavior of users, which may have a different nature both positive and negative. Together with the use of classifiers and social news dissemination channels, they can be identified and transmitted to relevant government agencies and services.

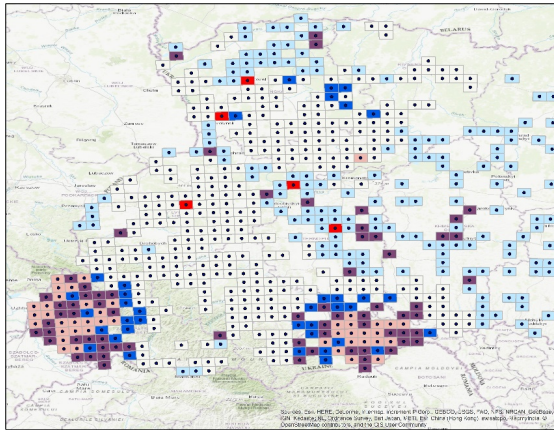


Fig. 5. The map of local outlier, created as a result of the analysis of the space-time cube

VIII. CONCLUSIONS

The toolkit for building a spatial-temporal cube provides a convenient visual interface for data mining of big data. The use of the spatial-temporal cube is practically possible in virtually all areas where it is necessary to analyze the behavior of objects and events occurring with the change of location in space and time.

An example of the use of spatial-temporal analysis of data for events in mobile networks, for example, of the Vodafone network, makes it possible to use the data more effectively, primarily for security purposes, which will be useful to governmental organizations for the rapid detection or prevention of dangerous situations (such as terrorism, extraordinary events, catastrophes, etc.). In the future, using the spatial-temporal cube based on the data of mobile operators, it is possible to analyze the statistical emissions in the activity of subscribers in calls or connecting people to the Internet with an anchorage of a certain territory, which will allow to identify certain anomalies and respond accordingly.

REFERENCES

[1] J. Rak, J. Bay, I. Kotenko, L. Popyack, V. Skormin, and K. Szczypiorski, "Computer Network Security," 7th International Conference on Mathematical Methods, Models, and Architectures for Computer Network Security, MMM-ACNS 2017, Warsaw, Poland, August 28-30, 2017.

[2] ESRI, ArcGIS Analysis Workflows for Public Safety, 25.10.2017, Available: <https://community.esri.com/docs/DOC-10345-arcgis-analysis-workflows-for-public-safety>

[3] T. Hägerstrand, "What about people in regional science?," Papers, Regional Science Association, vol. 24, pp. 7-21, 1970.

[4] M. Kraak, "The Space Time Cube Revisited from a Geovisualization Perspective," 21st Int'l Cartographic Conf., pp. 1988-1996, 2003.

[5] N. Andrienko, G. Andrienko, and P. Gatalisky, "Exploratory SpatioTemporal Visualization: An Analytical Review," J. Visual Languages and Computing, vol. 14, no. 6, pp. 503-541, 2003.

[6] P. Gatalisky, N. Andrienko, and G. Andrienko, "Interaction Analysis of Event Data Using Space-Time Cube," IEEE Eighth Int'l Conf. Information Visualization (IV '04), pp. 145-152, 2004.

[7] G. Andrienko, and N. Andrienko, "Interactive maps for visual data exploration," International Journal Geographical Information Science, vol. 13, pp. 355-374, 1999.

[8] T. Cheng, and M. Adepeju, "Detecting emerging space-time crime patterns by prospective STSS," 12th International Conference on GeoComputation, pp. 281-285, 2014.

[9] B. R. Shapiro, and R. P. Hall, "Interaction Geography in a Museum," CHI Conference Extended Abstracts on Human Factors in Computing Systems, Denver, Colorado, USA, pp. 2076-2083, May 06-11, 2017

[10] M. Baas, Space-time cube analysis of animal behaviour, Thesis report GIRS-2013-14, 2013, Available: <http://library.wur.nl/WebQuery/theses/2037155>

[11] T. Gonçalves, A. Afonso, and B. Martins, "Visualizing Human Trajectories: Comparing Space-Time Cubes and Static Maps," 28th International BCS Human Computer Interaction Conference on HCI 2014, Southport, UK, pp. 207-212, September 09-12, 2014,

[12] N. Yusof, R. Zurita-Milla, M. Kraak, and V. Retsios, "Mining frequent spatio-temporal patterns in wind speed and direction," Connecting a digital Europe through location and space, pp. 143-161, 2014.

[13] K. Khurtsilava, "On the question of space-time GIS and some of their applications," Decision support systems. Theory and practice, Kyiv, IMMSPU NASU, pp. 52-55, 2013.

[14] M. Zgurovsky, A. Boldak, and K. Yefremov, "Intellectual Analysis and System Coordination of Scientific Data in Interdisciplinary Studies," Cybernetics and Systems Analysis, vol. 4, pp. 62-75, 2013.

[15] M. Zgurovsky, and N. Pankratova, System analysis: Problems, methodology, applications. K: Naukova dumka, 2005. (in Ukrainian).

[16] A. Petrenko, "Grid and Data Mining," System research and information technology, vol. 4, pp. 97-110, 2008. (in Ukrainian).

[17] V. Putrenko, "Data mining of relationship in crowdsourcing projects and social activities of citizens," IEEE 1st Ukraine Conference on Electrical and Computer Engineering, UKRCON 2017, pp. 1060-1065, 2017.

[18] V. Putrenko, and N. Pashynska, "Application of geoinformation modeling tools for intelligent data analysis of fire hazards," Bulletin of National Technical University "KhPI", Ser. New solutions in modern technologies, vol. 7 (1229), pp. 156-163, 2017.

[19] D. J. Peuquet, "It's about time: a conceptual framework for the representation of temporal dynamics in geographic information systems," Annals of the Association of American Geographers, 84 (3), pp. 441-461, 1994.

[20] ESRI, An overview of the Space Time Pattern Mining toolbox, Available: <http://desktop.arcgis.com/ru/arcmap/10.3/tools/>

[21] L. Anselin, "Local Indicators of Spatial Association – LISA," Geographical Analysis, 27(2), pp. 93–115, 1995.

Bipartite Graph Analysis as an Alternative to Reveal Clusterization in Complex Systems

Vasyl Palchykov

*Laboratory for Statistical Physics of Complex Systems
Institute for Condensed Matter Physics, NAS of Ukraine*

Lviv, Ukraine

*L⁴ Collaboration & Doctoral College
for the Statistical Physics of Complex Systems,
Leipzig-Lorraine-Lviv-Coventry, Europe
palchykov@icmp.lviv.ua*

Yurij Holovatch

*Laboratory for Statistical Physics of Complex Systems
Institute for Condensed Matter Physics, NAS of Ukraine*

Lviv, Ukraine

*L⁴ Collaboration & Doctoral College
for the Statistical Physics of Complex Systems,
Leipzig-Lorraine-Lviv-Coventry, Europe
hol@icmp.lviv.ua*

Abstract—We demonstrate how analysis of co-clustering in bipartite networks may be used as a bridge to connect, compare and complement clustering results about community structure in two different spaces: single-mode bipartite network projections. As a case study we consider scientific knowledge, which is represented as a complex bipartite network of articles and related concepts. Connecting clusters of articles and clusters of concepts via article-to-concept bipartite co-clustering, we demonstrate how concept features (e.g. subject classes) may be inferred from the article ones.

Index Terms—bipartite network, knowledge graph, clustering, modularity, one-mode projections

I. INTRODUCTION

Among different types of systems one may distinguish a special class of complex systems [1]. Such systems consist of many interacting parts and these interaction patterns may result in a new level of organization in the entire system – emergent phenomena, which appear as a result of bottom-up local interactions rather than a centralized top-down control, see e.g. [2]. Consequently, such decentralized system may behave as a new organism; its behaviour differs significantly from the behaviour of its constituents and is rather governed by connectivity patterns between these parts, which often exhibit quite complex properties and form a complex network topology [3], [4].

The systems that belong to a class of complex systems may be observed in different environments and include both natural and social phenomena. Examples are formations in flock of birds or school of fishes, evacuation of crowds, opinion formation in society, financial markets and formation of financial bubbles. In order to understand, to model or to investigate the scenarios of behaviour for such systems one may need to analyze the underlying network topology. This include two-step procedure: i) to represent the system as a network or a graph, i.e. a collection of nodes connected by links, and ii) to perform analysis of the corresponding network topology.

The structure of the underlying network may be often expressed as a bipartite graph (or even more complicated),

see e.g. [4]. Such networks has two types of nodes. The links connect the nodes of different types and mean the existence of some kind of relations between them. One may think of scientific publications and authors [5], public transport routes and serviced stations [6], foods and their ingredients [7], etc. However, these networks are often simplified to one mode projections, e.g. scientific publications connected to each other if they have an author in common, or co-authorship network of researchers. Then these projections are investigated using some tools. For instance community detection or clustering approaches [8] may be used to extract groups or modules in such systems.

One of the goals of this paper is to demonstrate how one-mode projections of the bipartite network may give us different insight into the whole system and how the three representations (two one-mode projections and the bipartite one) may organically supplement one another via clustering analysis. As a case study we will use scientific publication records of manuscripts in a physics domain and extracted concepts [9]. Besides an obvious demonstration of how bipartite graphs and their projection may be used to extract clustering structure within a system, our article answers a general question concerning organization of scientific knowledge. The rest of the paper is organized as follows: in Section II we describe two alternative ways to define scientific knowledge. Section III describes the way the data have been collected, followed by a brief overview of the dataset. In Section IV we show different ways how to represent data set as a network. In Section V we present clustering analysis of these networks and summarize the results in Section VI.

II. KNOWLEDGE: ALTERNATIVE APPROACHES

There is no strict and unique mathematical formalization of scientific knowledge. Let us assume that each original scientific publication produces a new piece of knowledge. These pieces of knowledge do not “live” in isolation, but are connected to the other pieces. They may be connected to existing pieces of knowledge by citation links. Thus, one

arrives at a simplified view of the knowledge as a citation graph [10]. The nodes of such graphs correspond to scientific publications and chronologically directed links indicate citation links between papers. Besides citation graph one may consider other types of relationships between scientific publications, e.g. content or author-based similarity metrics, see e.g. [11]. Alternative view of the knowledge (and its constituents) assumes that scientific publications consist of the pieces of knowledge rather than represent such pieces themselves. Here scientific ideas may be considered as constituents of knowledge, which may be represented by concepts in scientific publications. Each publication may contain a number of concepts. Examples of scientific concepts in physics domain of science include `Center of mass`, `Momenta of inertia`, `Conservation of energy`, etc. Once a new concept appears in a scientific communication, it is assumed that the (scope of) existing knowledge has been extended, see e.g. [12]. To use this approach, scientific concepts have to be extracted from the bodies of scientific publications. In the next chapter we describe the process of such concept extraction.

III. DATA AND EXPLORATORY ANALYSIS

It is possible to obtain basic or generic set of concepts in scientific domain, but it is much more complicated to get a comprehensive vocabulary of concepts even in a single domain of science due to the following reasons: i) constant evolution of concepts due to the development of scientific knowledge, and ii) fragmentation of science. It is natural to assume that a comprehensive vocabulary of scientific concepts may be possessed only by active scientist, and only in the field of his or her research. In order to create up-to-date vocabulary (ontology) of scientific concepts it may be required to combine *computational capabilities* with the efforts of researchers from different domains and regularly (if needed) upgrade it. With this purpose in mind `ScienceWISE (SW)` platform has been deployed [13], [14]. This platform was designed to allow scientists to navigate over scientific literature, using scientific concepts as “labels” (or “tags”) for navigation. Initial vocabulary of concepts has been obtained using a semi-automated import from science-oriented ontologies and online encyclopedia [13]. Afterwards the users of SW were allowed to edit ontology to make navigation more reliable. For instance if a user observes that some concept (e.g. `bipartite network`) is missed, and the concept will help to navigate over the literature, he or she may add the concept to the vocabulary. Once a new concept is added to the vocabulary, all the articles have to be re-scanned in order to upgrade concept list for each paper.

Below we investigate the structure of concept-related networks of scientific knowledge. A primary source of literature for SW platform is `arXiv` e-print repository of manuscript [15], which allows for the full-text access to all manuscripts even before they are officially published.

Here we use scientific concepts extracted from research publications (`arXiv` preprints) using the SW platform. These concepts have been previously investigated in [9], [14], [16].

Here we are interested in investigation of the static picture of concept (knowledge) network, thus we restrict ourselves to a single year, namely 2013. Another reason is that some structural properties of the corresponding network of publications (on the same data sets) have been investigated in [9]. Here we again restrict our analysis to the manuscripts to which a single category has been attached by the authors. It is worth to mention that during manuscript submission process to `arXiv` the authors are required to classify their manuscript, i.e. assign it to at least one subject class in `arXiv` classification scheme. This classification scheme consists of thirteen subject classes, which include in particular `astro-ph`, `cond-mat`, four classes of high energy physics (`hep-ph`, `hep-ex`, `hep-lat` and `hep-th`), two classes of nuclear physics (`nucl-th` and `nucl-ex`), `general relativity and quantum cosmology (gr-qc)`, `physics`, `quantum physics (quant-ph)`, `nonlinear sciences (nlin)` and `mathematical physics (math-ph)`.

The subject of our analysis was a collection of manuscripts submitted during the year 2013 that accounts for 36386 articles. These articles contain 12200 unique concepts in common. 347 of these concepts have an expert given generic label, which means that the concept has a generic meaning in physics. Examples include `Energy` or `Field`. Each manuscript contains on average 37 unique non-generic concepts with significant differences among the articles: the number of identified concept within an article varies between 1 and ≈ 400 . A brief summary of the data set properties is shown in Table I.

TABLE I
BASIC CHARACTERISTICS OF THE DATASET: TOTAL NUMBER OF MANUSCRIPTS (N), TOTAL NUMBER OF IDENTIFIED CONCEPTS (V) AND THE NUMBER OF GENERIC ONES (V_{gen}) AMONG THEM; $\langle k \rangle$ STANDS FOR THE AVERAGE NUMBER OF NON-GENERIC CONCEPTS IN AN ARTICLE.

	N	V	V_{gen}	$\langle k \rangle$
<code>arxivPhys2013</code>	36386	12200	347	37

IV. NETWORK REPRESENTATIONS

As a basic network representation of the publication system we consider bipartite network. Here manuscripts and the identified concepts are mapped into two types of network nodes. A link connects an article-node and a concept-node if the corresponding concept has been identified within the text of the article. The corresponding representation is shown in Fig. 1a.

There are two possible one-mode projections of this bipartite network. Considering only concept-nodes and linking each pair of these nodes that co-occurred in (at least once) the same publication, we arrive at the projection to the `concept space`, see Fig. 1b. For simplicity all links are assumed to have the same (unit) weight.

Alternatively, one may build a supplementing projection to the `article space`. Here two article-nodes are connected to each other if they use at least one concept in common.

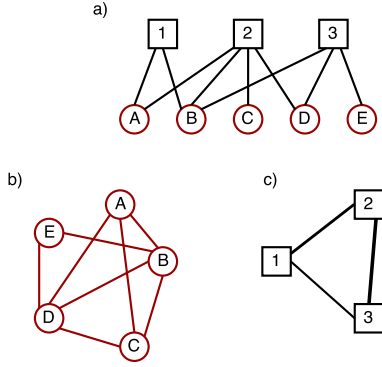


Fig. 1. Examples of network representations for scientific publication system. Panel **a)**: unweighted bipartite representation that contains two types of nodes: articles (squares) and concepts (circles) and the links connecting the nodes of distinct types. Two other panels represent one-mode projections of the bipartite network. Panel **b)**: concept network that consists only of concept-nodes; two nodes are connected in the corresponding concepts co-occurred together at least in one publication. Panel **c)**: the nodes of the network represent scientific articles and two article-nodes are connected if they used the same terminology (they have at least one concept in common). This network may be also called `content (concept) coupling network` in analogy to the `bibliographic coupling network`.

Since such networks are extremely dense (there are over 50% of all possible links in a network), a weighted networks configuration is considered here. Firstly we represent each article i as a vector a_i in a multi-dimensional space, where each dimension corresponds to each of $V - V_{\text{gen}}$ unique concept. A weight w_{ij} of a link between articles i and j is defined as a cosine between the corresponding article vectors:

$$w_{ij} = \cos(\vec{a}_i, \vec{a}_j). \quad (1)$$

To account for the heterogeneity in the widespread of concepts, each concept is weighted according to its `idf` factor:

$$\text{idf}(c) = \log \frac{N}{N_c}, \quad (2)$$

where N_c is the number of articles in collection that contain concept c . Thus, defining \vec{a}_i such that

$$a_{i,c} = \begin{cases} \text{idf}(c), & \text{if } c \in i \\ 0, & \text{otherwise} \end{cases} \quad (3)$$

we take into account that the usage of a common widespread concept affects the similarity score between two articles less than the usage of a more specific one.

V. MODULAR STRUCTURE

Once the networks are generated, let us investigate their community structure. In order to identify clusters in a network, modularity [17] optimization approach has been applied, with the Barber's modification [18] for bipartite networks. To maximize modularity, greedy optimization Louvain algorithm [19] has been applied. Due to its stochastic nature, which results in a local rather than global maximum at each run, we performed 100 runs on unipartite networks and 1000 runs on a bipartite network. Then a single partition for each network

with the highest value of modularity has been chosen. Note that the clusters of one-mode projections consist of the nodes of a single type only, while the cluster of a bipartite network, in general, consists of both article nodes and concept-nodes. The scores of the highest modularity M (M_B for bipartite network) obtained:

- Bipartite network: $M_B = 0.453262$
- Concept network: $M = 0.245219$
- Article network: $M = 0.329519$

In what follows below we consider only the clusters that have more than one node (unipartite networks) or more than two nodes (bipartite network).

Previously [9] the clusters of scientific manuscripts have been compared to their expert made classification, both for bipartite network and its projection to article space. Since each cluster of a bipartite partition consists of a set of articles and concepts, a subset from each cluster, namely the articles that fall into this cluster, have been considered. The results demonstrated: i) similarity between the obtained clusters and author made classification, and ii) some discrepancies between the two. The detailed analysis showed us that some of these discrepancies have underlying reasons behind, indicating both historical classification reasons and methodological similarity between rather unrelated domains. It appeared [9] that the bipartite network contained six clusters and the article-to-article one contained four clusters, which provided rather similar results.

In the present research we extend the analysis of Ref. [9] by adding the concept dimension: instead of ignoring the concepts that fall within each bipartite cluster we take them into account and use the bipartite combined clusters as a bridge between pure partitions of article and concept unipartite projection partitions. To illustrate this let us mention that unlike articles, scientific concepts lack expert made classification. So, we cannot directly assign arXiv subject classes to the clusters of concepts. This, however, may be done using combined (consisting of concepts and articles) partitions of bipartite clusters, as illustrated in Fig. 2.

Fig. 2 displays a brief description of the optimal partitions in bipartite network and its both projections together with some relations. Beside six clusters of the bipartite network optimal partition, and four ones of the article-to-article network, the optimal partition of the concept network consists of three clusters. Let us first consider the first two clusters of bipartite networks (shown at the top of Fig. 2). The articles that belong to these clusters are dominated by high energy physics (`hep-`) categories. The main difference between these two clusters is that the first one is rather focussed on experimental observations, while the second cluster is more about theoretical approaches to the problem, for details see [9]. In the unipartite projection to the article space these two clusters rather form a single one (at the top), which we may call `article:hep`. Moreover there is quite good correspondence between these clusters: 97% of articles of `article:hep` cluster belong to either of the two clusters of the bipartite partition. On the other hand, there is a good correspondence between these clusters

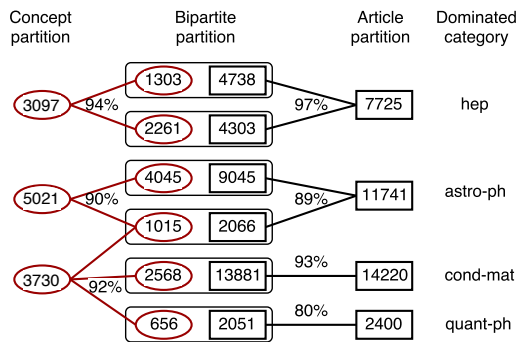


Fig. 2. Clusters of concept network (left hand side, ovals), bipartite network (center) and a article network (right hand side, rectangles). The number in an ovals gives the number of concepts in the cluster, the number in a rectangle gives the number of articles in the cluster. A pair of numbers in a bipartite network gives the number of concepts and the number of articles in each cluster of a bipartite network. A category code next to each article cluster denotes the category to which most of the articles of the cluster belong. Percentage near each cluster of an article or a concept network show the fraction of nodes in the cluster that belongs to a linked cluster(s) of the bipartite network.

and the top cluster (Fig. 2) of a unipartite projection to concept space: 94% of all concepts that fall into this cluster belong to the considered clusters of bipartite partition. Thus, we call this cluster as `concept:hep`. These are the concepts that has dominant usage in high energy physics research.

Similar picture may be observed for `astro-ph` subject class. 89% of nodes of the second cluster in article space belong either to the third or to the fourth cluster in bipartite space, and again the concepts of the latter clusters form 90% of all concepts belonging to the second cluster in concept space. Thus, as a good approximation, the concepts belonging to this cluster may be labeled as `concept:astro-ph` ones.

The situation with `cond-mat` and `quant-ph` subject classes is different. While they form rather separated clusters in article space, the concepts used in these subject classes fall into a single cluster. These observations hint that besides considering different objects and on different scales the methodology behind these subject classes may overlap significantly.

VI. CONCLUSIONS

The structure of many complex systems can be expressed in terms of the underlying bipartite network, which connects different types of components. Investigation of such systems is often done considering one mode projections of these separately. Examples include co-authorship network of paper-to-author bipartite graph [20], etc. To find patterns in such networks one may apply a variety of tools including clustering algorithm to identify groups of tightly related items in a system. Here we show how co-clustering of bipartite network may be used as a bridge to connect and complement clustering results in two different projections. Considering scientific publications in physics domain and a set of extracted concepts

from their texts, we build bipartite article-to-concept networks and made its both one mode projections. We show how the information about one part of the system may add value to the other one. In the considered case publications possess the author made classification according to arXiv subject classes, however scientific concepts lack such classification. By comparing such concepts we were able to assign such classes to the concepts. Moreover such approach allows us to compare groups of completely different objects: articles and concepts. We see a hint that two different subject classes, `cond-mat` and `quant-ph`, even though being well distinguishable in terms of scientific publications, use terminology (concepts), which is quite similar.

ACKNOWLEDGMENT

This work was supported in part by FP7 EU IRSES project No. 612707 “Dynamics of and in Complex Systems” and by the Ukrainian DFFD by the project 76/105-2017 (Yu.H.).

REFERENCES

- [1] S. Thurner, 43 Visions for Complexity. Singapore: World Scientific, 2017.
- [2] Yu. Holovatch, R. Kenna, S. Thurner, “Complex systems: physics beyond physics,” *Eur. J. Phys.*, vol. 38, p. 023002, 2017.
- [3] M. Newman, *Networks: an introduction*. Oxford: Oxford university press, 2010.
- [4] A.-L. Barabási, *Network science*. Cambridge: Cambridge university press, 2016.
- [5] M. E. Newman, “The structure of scientific collaboration networks,” *Proc. Natl. Acad. Sci. USA*, vol. 98, pp.404–409, 2001.
- [6] C. von Ferber, T. Holovatch, Yu. Holovatch, and V. Palchykov, “Public transport networks: empirical analysis and modeling,” *Eur. Phys. J. B*, vol. 68, pp. 261–275, 2009.
- [7] Y. Y. Ahn, S. E. Ahnert, J. P. Bagrow, and A.-L. Barabási, “Flavor network and the principles of food pairing,” *Sci. Rep.*, vol. 1, p. 196, 2011.
- [8] S. Fortunato, “Community detection in graphs,” *Phys. Rep.*, vol. 486, pp. 75–174, 2010.
- [9] V. Palchykov, V. Gemmetto, A. Boyarsky, and D. Garlaschelli, “Ground truth? Concept-based communities versus the external classification of physics manuscripts,” *EPJ Data Science*, vol. 5, p. 28, 2016.
- [10] D. J. D. S. Price, “Networks of scientific papers,” *Science*, pp. 510-515, 1965.
- [11] L. Waltman, N. J. van Eck, and E. C. Noyons, “A unified approach to mapping and clustering of bibliometric networks,” *J. Informetr.*, vol. 4, pp. 629–635, 2010.
- [12] I. Iacopini, S. Milojević, and V. Latora, “Network Dynamics of Innovation Processes,” *Phys. Rev. Lett.*, vol. 120, 048301 (2018).
- [13] A. Astafiev, R. Prokofyev, C. Guéret, A. Boyarsky, and O. Ruchayskiy, “ScienceWISE: a Web-based interactive semantic platform for paper annotation and ontology editing”. In: Simperl E. et al., Eds. Berlin, Heidelberg: Springer, 2015.
- [14] A. Martini, A. Lutov, V. Gemmetto, A. Magalich, A. Cardillo, et al., “ScienceWISE: Topic Modeling over Scientific Literature Networks,” arXiv preprint arXiv:1612.07636, 2016.
- [15] P. Ginsparg, “ArXiv at 20,” *Nature*, vol. 476, pp. 145–147, 2011.
- [16] A. Martini, A. Cardillo, and P. De Los Rios, “Entropic selection of concepts in networks of similarity between documents,” arXiv preprint arXiv:1705.06510, 2017.
- [17] M. E. Newman, M. Girvan, “Finding and evaluating community structure in networks,” *Phys. Rev. E*, vol. 69, p. 026113, 2004.
- [18] M. J. Barber, “Modularity and community detection in bipartite networks,” *Phys. Rev. E*, vol. 76, p. 066102, 2007.
- [19] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, E. Lefebvre, “Fast unfolding of communities in large networks,” *J. Stat. Mech. Theory Exp.*, vol. 10, P10008, 2008.
- [20] M. E. Newman, and J. Park, “Why social networks are different from other types of networks,” *Phys. Rev. E*, vol. 68, p. 036122, 2003.

Comparative Study of Massively Parallel GPU Realizations of Wavelet Transform Computation with Lattice Structure and Matrix-Based Approach

Dariusz Puchala
Institute of Information Technology
Lodz University of Technology
Łódź, Poland
dariusz.puchala@p.lodz.pl

Kamil Stokfiszewski
Institute of Information Technology
Lodz University of Technology
Łódź, Poland
kamil.stokfiszewski@p.lodz.pl

Kamil Wieloch
Institute of Information Technology
Lodz University of Technology
Łódź, Poland
kamil.wieloch@dokt.p.lodz.pl

Mykhaylo Yatsymirskyy
Institute of Information Technology
Lodz University of Technology
Łódź, Poland
mykhaylo.yatsymirskyy@p.lodz.pl

Abstract—In this paper the authors analyze the effectiveness of parallel graphics processing unit (GPU) realizations of discrete wavelet transform (DWT) using lattice structure and matrix-based approach. Experimental verification shows that, in general, for smaller input vector sizes along with the larger filter lengths DWT computation based on the direct approach with the use of the direct matrix multiplication significantly faster than the application of the lattice structure factorization while for large vector sizes the lattice structure becomes more effective. The detailed results define boundaries of performance for both implementations and determine the most advantageous situations in which one might use a given approach. The results also include comparative analysis of time efficiency of the presented methods for two different GPU architectures. The presented effectiveness characteristics of the considered realizations of the two selected DWT computation methods allows for making the proper choice of the particular method in future applications depending on input data sizes, filter lengths and underlying GPU architecture.

Index Terms—discrete wavelet transform, graphics processing units, lattice structure, massively parallel computations, time effectiveness, data mining and processing

I. INTRODUCTION

Discrete wavelet transforms (DWTs) are important computational tools used in many areas of data processing such as image and video processing [1], [2], image watermarking, e.g. [3], analysis and clustering of high dimensional data [4], [5], data mining [6], and many other important data processing and analysis tasks, e.g [7], [8]. This makes the research on improvement of the algorithms of the wavelet transform calculation very intense in recent years, see e.g. [9], [10], especially with the growing popularity of graphics processing units (GPUs) computations, for which massively parallel algorithms have been constructed, c.f. [11]. Although the subject of GPU realizations of various computational tasks has become common in recent years, a large part of the

analyzes focus on comparing the performance of CPU and GPU processors' implementations of the considered methods (see e.g. [12]) while at least as much an important issue is an analysis of the performance of different algorithms suited for massively parallel realizations solving the same computational problem with the use of GPUs. Optimal task defragmentation aiming for the maximum utilization of computational and resource potential of given GPU architecture seems an interesting direction of research, c.f. [13], [14] and [15]. As shown by earlier studies, e.g. [14], various parallel GPU implementations the same computational task involving discrete linear transforms can have very diverse time efficiencies.

In this article the authors compare time efficiency of two DWT calculation algorithms aimed at GPU realizations, namely, the DWT computation algorithm utilizing the lattice structure, (see [9]), and a direct matrix-based approach DWT computation, (c.f. [16]). Both algorithms have been refined for attaining maximum time efficiency when executed on GPUs. Our goal was to verify whether the fast algorithms that reduce computational complexity of DWT computation turn out to be actually more time-efficient on massively parallel systems in comparison to direct modulation matrix approach, and, possibly, determine the division boundary between both approaches with respect to various DWT input data sizes and different filter lengths.

II. GPUS' ARCHITECTURES AND EXECUTION MODEL

In this section we'll briefly discuss modern GPUs' architectures and kernel execution models on the example of CUDA (*Compute Unified Device Architecture*) proposed by NVIDIA Corporation, see [15]. The GPU architecture is built around a scalable array of *Streaming Multiprocessors*, so called SMs. Fig. 1 shows a simplified scheme of GPUs' architecture. Each SM consists of a large number (at least 32 starting

with Fermi microarchitecture) of *cores* (small rectangles inside SMs in Fig. 1). A set of 32 CUDA cores (on most of the GPU architectures) share a single instruction pipeline capable of issuing and executing a single *thread* instruction on 32 arithmetic-logic units (ALUs) performing integer, floating-point and logic operations in a fully concurrent fashion, each on its own data. Such execution organization causes the threads to be arranged into groups called *warps* containing 32 threads each, which constitute the smallest portions of concurrently executing computation processes at a given moment of time.

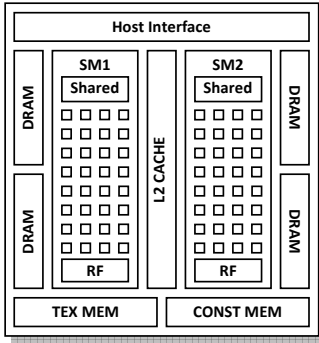


Fig. 1. Simplified scheme of modern GPUs' architecture

Although all threads in a warp execute the same instruction at the given moment of time, each thread has its own instruction counter and register state and can be treated individually. If some of the threads of a warp branch, the warp would serially execute each branch path, disabling the threads that do not take that path. Such execution model is referred to as a *Single Instruction Multiple Thread* (SIMT) architecture, see [15].

When a *kernel*, i.e. the set of all threads to be executed on the GPU, is launched the user provides the GPU's host interface with the compiled thread code, a number of *blocks* into which threads are partitioned and a number of threads present in a single block. Blocks are then scheduled evenly between GPU's SMs for execution. Once a block is scheduled to a particular SM it resides there until its execution completes. Fig. 2 shows a logical structure of the kernel's execution process.

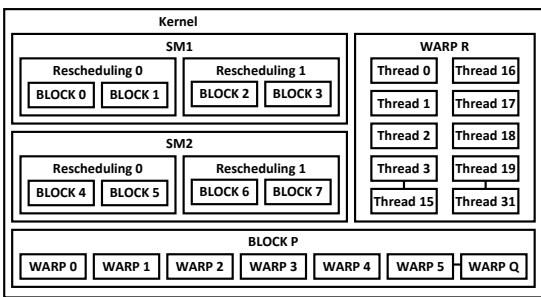


Fig. 2. Simplified logical structure of the kernel's execution process

The number of *active* warps, i.e. the warps which are concurrently executing on a given SM at a given time instance, is limited by the SM resources. Resources include registers and shared memory (thread's local variables and shared variables) which are used by a single thread and, consequently, by a

single warp (c.f. register file (RF) and Shared symbols present in Figure 1). E.g. the SM's register file size for CUDA devices with *compute capability* 2.1 is equal to 32768 32-bit floating-point registers and it determines the maximum number of warps which can be concurrently executed on a single SM. Maximum number of concurrent SM warps is also limited by the SM's hardware limits, e.g. for CUDA architecture compute capability 2.1 maximum number of active blocks is 8 and maximum number of active warps equals to 48, which amounts to maximum possible number of concurrent threads running on a single SM to be equal to 1536 concurrent threads. When the number of warps scheduled to a single SM exceeds those limits the remaining warps are rescheduled in even portions for serial execution (c.f. Rescheduling boxes in Fig. 2).

In light of the above considerations the overall execution time for a selected computational method depends in particular on the number of sequential instructions present in the threads' code and on the chosen threads' partitioning to blocks what determines the level of SM occupancy. This, on the other hand, determines the ability of a computational process to hide latency of memory operations and to minimize the number of serial warp reschedulings. One has also take into account the number of serial kernel executions needed for data synchronization since this also might take the considerable amount of the overall execution time. All those factors considerably influence the execution time of a considered computational method and have to be jointly taken into consideration when comparing selected approaches to GPU implementations of the chosen computational task, c.f [15] or [18].

III. MATRIX-BASED REALIZATION OF DWT

This section describes the practical implementation of a discrete wavelet transform in a matrix-based approach. Such implementation in the confrontation with the approach based on lattice structures (described in Section IV) will allow for experimental verification of the thesis that algorithms that have lower computational complexity in the sequential approach do not have to be characterized by higher computational efficiency in the case of GPU accelerated mass-parallel calculations.

The elementary way to calculate any linear transformation, in this case a DWT, is an algebraic approach based on matrix multiplication by the vector of the input signal samples. We can then describe such an operation in matrix form as:

$$y = Ax,$$

where x is a N -element vector of the input signal, y is N -element vector representing the input signal in the dilation and translation domain of DWT, while A is a square N on N element matrix that performs the filtration operation of the input signal. It is well known that a discrete wavelet transform can be implemented as an analysis stage of two-channel bank of filters with finite impulse responses (see [8]). Then, both filters of an analysis stage, i.e. filters h and g with a number of K coefficients each, which are additionally bound by a perfect reconstruction condition (PR) (see [7]), are at the same time

the elements of DWT filters used in the process of convolution based filtration of input signal. It should be noted that the results of such filtration are subject to a process of decimation by a factor of 2, what results in lack of redundant data at the output of an analysis stage. Further on by assuming the cyclic rotation of input signal as a boundary condition, the filtration matrix A takes the following form:

$$A = \begin{bmatrix} h_{K-1} & \dots & h_1 & h_0 & 0 & 0 & \dots & 0 & 0 \\ g_{K-1} & \dots & g_1 & g_0 & 0 & 0 & \dots & 0 & 0 \\ 0 & 0 & h_{K-1} & \dots & h_1 & h_0 & \dots & 0 & 0 \\ 0 & 0 & g_{K-1} & \dots & g_1 & g_0 & \dots & 0 & 0 \\ \vdots & & \vdots & & \vdots & \vdots & \ddots & \vdots & \vdots \\ h_{K-3} & \dots & h_1 & h_0 & 0 & 0 & \dots & h_{K-1} & h_{K-2} \\ g_{K-3} & \dots & g_1 & g_0 & 0 & 0 & \dots & g_{K-1} & g_{K-2} \end{bmatrix}.$$

It is easy to verify that such a matrix consists of many values equal to zero, which have no effect on the final result of the calculations. It is natural then to omit such elements in the computational process, which may result in a significant reduction in the number of calculations. In practice, filter lengths are significantly shorter than the size of the input data, which makes the following relationship $K \ll N$ true. Thus, from the point of view of the number of calculations, the multiplication by A matrix of such form would be equivalent to multiplying the rectangular tall matrix with size of N on K elements by a K -element vector. According to the results presented in paper [13] for such type of matrices the most efficient mass-parallel implementation is the one, in which a number of N parallel threads is executed, and each of them in a sequential manner performs K multiplication and $K - 1$ additions making up a dot product of two vectors. Of course, the resulting value must be written in the output array in the place corresponding with the thread's coordinates within the grid of blocks. Bringing up other results from the same paper (see [13]), the implementation which is slightly worse, is the one in which every parallel thread processes two vectors in a sequential way. In such case, there is no problem with the filter selections, i.e. h and g for even and odd rows respectively, which is a definite advantage, but the number of parallel calculations performed is twice as small. Other investigated approaches that, e.g. do not need synchronization of calculations, take advantage of *atomicAdd()* function, or apply many threads to calculations within one row, were clearly slower. Taking it into consideration, in this paper, the authors decided to use the "one thread per row" approach, which is also the standard and the simplest approach to the task of matrix by vector multiplication (see Fig. 3). Additional parameters of kernel function from Fig. 3 are: input vector array *idt*, output vector array *odt* and filter coefficients *cof* (stored in the same array for both h and g filters).

IV. APPROACH BASED ON LATTICE STRUCTURES

The lattice structures are an effective tool for practical implementations of two-channel banks of finite impulse response filters [7]. In turn, two-channel banks of filters allow for practical realizations of discrete wavelet transforms (DWT)

(see [8]). This means that also wavelet transformations can be

```

void matrix (int K, int N, float *idt, float *odt, float *cof)
{
    float v;
    int id, k1, k2;
    id=blockIdx.x*blockDim.x+threadIdx.x;
    v=0.0f;
    k1=2*(id/2);
    k2=(id%2)*K;
    for (int i=0; i<K; i++)
    {
        v=v+idt[k1]*cof[k2];
        k1=(k1+1)%N;
        k2=k2+1;
    }
    odt[id]=v;
}

```

Fig. 3. Kernel function implementing a matrix-based calculation of DWT. effectively calculated with use of lattice structures. It should be noted that the computational complexity of the direct approach to calculation of DWT, i.e. the one based on a matrix multiplication, requires $\mathcal{L}_{MUL}^{MAT} = NK$ and $\mathcal{L}_{ADD}^{MAT} = N(K - 1)$ multiplications and additions respectively, where $K - 1$ is the order of filters, and N is the size of wavelet transformation. This number can be significantly reduced during the lattice structure factorization process. The mentioned factorization takes advantages of the dependencies between the filters of both banks, which are the inherent consequence of the perfect reconstruction condition imposed on the banks of filters (cf. [7]). The resulting lattice structures allow for calculation of wavelet transforms with the numbers of $\mathcal{L}_{MUL}^{LAT} = \frac{1}{2}N(K + 2)$ multiplications and $\mathcal{L}_{ADD}^{LAT} = \frac{1}{2}NK$ additions. Such a number of operations is almost twice smaller than the number of calculations required by the matrix-based approach.

Although lattice structures have been designed for hardware realizations of the banks of filters in a pipelined mode, it is still possible to construct on the basis of similar factorization the graph structures which are appropriate for mass-parallel calculations (see [9], [11]). In Fig. 4 we show an exemplary lattice structure for filters with length $K = 6$ and transform size of $N = 8$ points. It can be seen that the lattice structure for calculation of DWT requires a number of $K/2 + 1$ sequential steps (even K is assumed). The calculations within the first $K/2$ stages are described by base operations $\Gamma_{i,j}$ (marked with '•') of the general form:

$$\Gamma_{i,j} = \begin{bmatrix} 1 & s_{i,j} \\ t_{i,j} & 1 \end{bmatrix},$$

where $s_{i,j}$, $t_{i,j}$ are the free parameters whose values are determined during the factorization process, $i = 0, 1, \dots, K/2 - 1$ and $j = 0, 1, \dots, N/2 - 1$. It should be noted that a single $\Gamma_{i,j}$ operation requires 2 multiplications and 2 additions. The last stage consists of sole multiplications (described symbolically by '►') by the scaling factors in the number of $N/2$, i.e. by τ_i for $i = 0, 1, \dots, N/2 - 1$. At the output, we obtain a representation of the input signal $x(n)$ for $n = 0, 1, \dots, N - 1$ in two frequency bands: (a) low-frequencies represented by $y_0(k)$, (b) high-frequencies represented by $y_1(k)$ outputs for $k = 0, 1, \dots, N/2 - 1$.

The typical parallel realization of the lattice structure of the form depicted in Fig. 4, assumes the calculation of subsequent

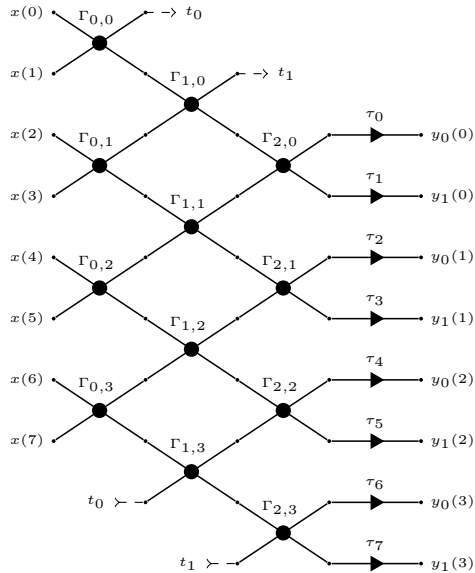


Fig. 4. Lattice structure for $N = 8$ point DWT with filters of length $K = 6$.

$\Gamma_{i,j}$ operators, as well as the pairs of neighbouring scaling multiplications in the last stage, to be performed in separate threads. Of course, calculations in subsequent stages must be also synchronized with use of global synchronization barriers. It gives a total number of $N/2$ computational threads. In Fig. 5, we present the CUDA kernel function used as the implementation of a single stage of a lattice structure. A parameter i , where $i = 0, 1, \dots, K/2$, represents an index of a specific stage, $L = K/2$, and *data* and *params* arrays (allocated in the global DRAM memory of a GPU device) hold processed data and the values of $s_{i,j}$ and $t_{i,j}$ parameters respectively. The global synchronization required between stages is realized on a host computer throughout subsequent kernel calls.

```

void lattice(int i, int L, int N, float*data, float*params)
{
    float t, s, a, b;
    unsigned int k0, k1, k2;
    k1=2*(threadIdx.x+blockDim.x*blockIdx.x)+(i%2);
    k2=(k1+1)%N;
    k0=2*i;
    s=params[k0+0];
    t=params[k0+1];
    a=data[k1];
    b=data[k2];
    if (i<L)
    {
        data[k1]=a+s*b;
        data[k2]=t*a+b;
    }
    else
    {
        data[k1]=s*a;
        data[k2]=t*b;
    }
}

```

Fig. 5. Kernel function implementing a single stage of a lattice structure.

The inverse discrete wavelet transformation (IDWT) can be calculated with use of the analogous structure (see Fig. 4) obtained by reversing the order of lattice stages and by

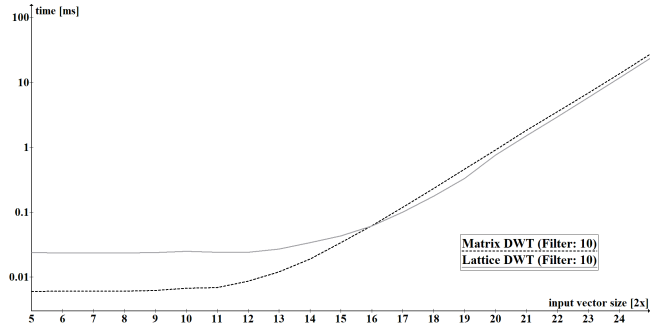


Fig. 6. Comparison of computing time for Matrix and Lattice DWT with filter size: 10 on CUDA architecture - 5.0 Maxwell.

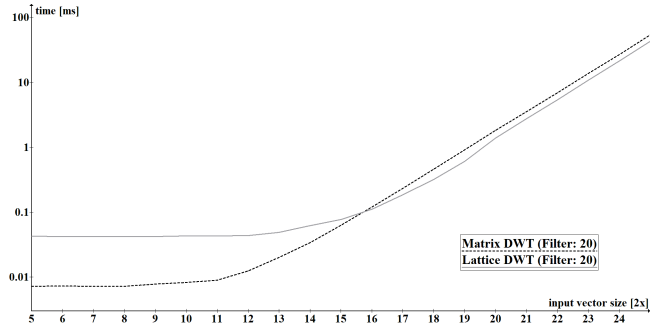


Fig. 7. Comparison of computing time for Matrix and Lattice DWT with filter size: 20 on CUDA architecture - 5.0 Maxwell.

inverting all base operations (cf. [9], [11]).

Another effective approach to calculation of wavelet transforms, in particular biorthogonal transformations with symmetrical filters, is a lifting scheme [7]. It is well known that the lifting structure can be converted to the lattice form (see [19]), also including the one identical to that of Fig. 4. However, these considerations go beyond the scope of this paper and will be the basis for future research.

V. EXPERIMENTAL RESEARCH

In this section experimental research will be presented and discussed in detail. All of experimental results shown below were selected and averaged from many sample runs involving various CUDA devices and architectures. All implementations were made with the use of the NVIDIA CUDA 9.1 Toolkit. In the presented research, portable and stationary computational devices were used. Time measurements were made with the use of methods provided by CUDA API that ensure preciseness and the reported times were sampled directly on the GPUs (except for kernel function call times). The data transfer times from host to device and vice-versa were not included in the presented results. All implementations operated on artificially prepared random data of single-precision floating-point type samples.

After selection of the fastest matrix implementation of the DWT, we started the comparison with the standard lattice implementation. The results of this comparison for the Maxwell 5.0 architecture are presented in Fig's 6–8, and for the Pascal

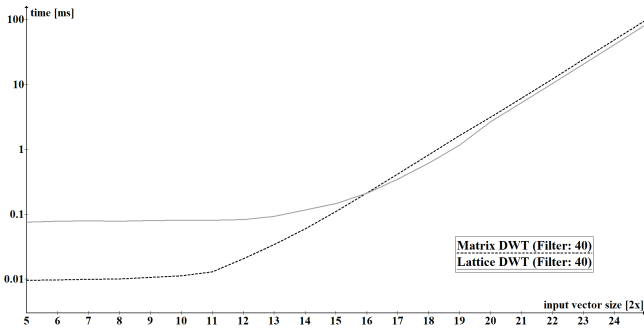


Fig. 8. Comparison of computing time for Matrix and Lattice DWT with filter size: 40 on CUDA architecture - 5.0 Maxwell.

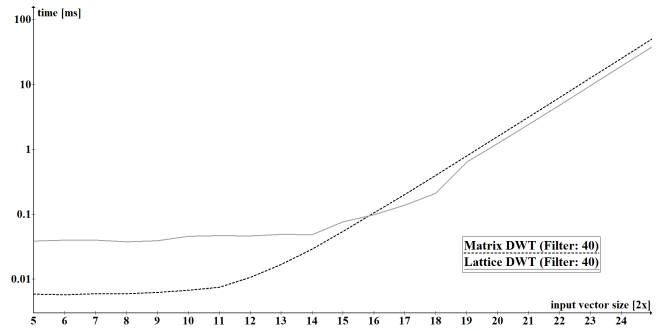


Fig. 11. Comparison of computing time for Matrix and Lattice DWT with filter size: 40 on CUDA architecture - 6.0 Pascal.

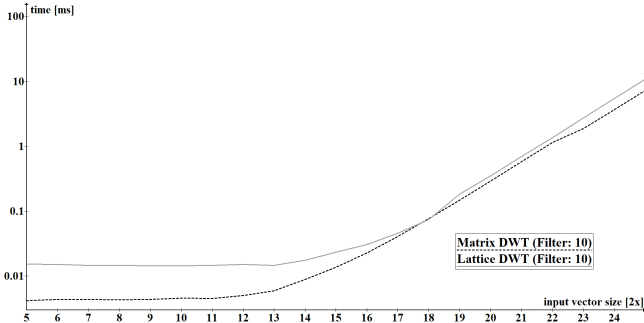


Fig. 9. Comparison of computing time for Matrix and Lattice DWT with filter size: 10 on CUDA architecture - 6.0 Pascal.

6.0 architecture in Fig's 9–11. Because the differences in execution times were small, especially for larger data sets, we've used logarithmic scale to present the execution times results. The division into blocks and threads was made in such a way that each single CUDA block consisted of half of the maximum number of available threads, taking into account the minimum of the DWT size and the maximum number of threads per SM for a given architecture. Such approach appeared to be definitely the most time effective one. For input vectors greater than 2^{10} , the number of threads per block was therefore 512, because the maximum possible value is limited by the architecture of CUDA (the oldest of the considered) up to 1024 threads per block. Figures 12 and 14 show the ratios of execution times of the lattice DWT to matrix

DWT implementations for various data sizes and filter lengths. Figures 13 and 15 show performance profitability boundaries graphs of the two considered DWT computation methods with regard to DWT input data sizes and filter lengths. It can be seen that the differences between DWT lattice and matrix approach are small for large data sizes, nevertheless for such sizes the lattice approach is more time-effective. For data sizes up to around 2^{15} samples the matrix-based approach is considerably faster (even up to 8 to 10 times) compared to the lattice structure-based one. Also it can be observed that the bigger the filter size is the more time-effective the matrix-based DWT computation approach becomes. As indicated in earlier sections such results are consistent in terms of all aspects and limitations of CUDA architecture which take place in each of the considered implementations. At last it's worthwhile noting that the limited time-efficiency of the lattice structure-based approach of the DWT computation in comparison to matrix-based one also is heavily influenced by the necessity of performing global synchronization between subsequent stages of its computational procedure, what is unnecessary in case of the latter approach.

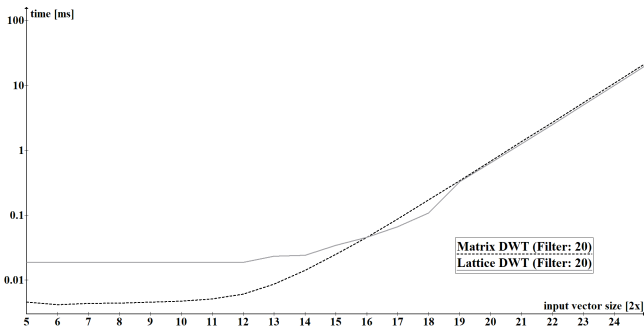


Fig. 10. Comparison of computing time for Matrix and Lattice DWT with filter size: 20 on CUDA architecture - 6.0 Pascal.

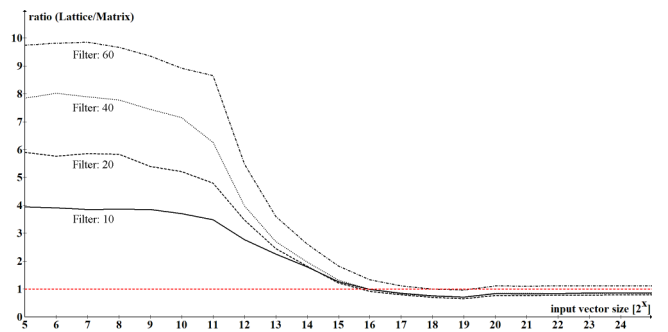


Fig. 12. Lattice DWT to Matrix DWT ratio for chosen filter lengths on CUDA architecture - 5.0 Maxwell.

VI. CONCLUSIONS

In this paper the authors analyze the effectiveness of parallel graphics processing unit realizations of discrete wavelet transform using lattice structure and matrix-based approaches. Experimental verification backed up with GPU architectural considerations has shown that lattice structure DWT computation

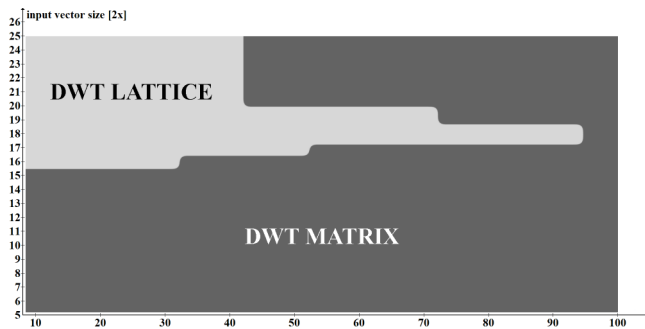


Fig. 13. Performance of matrix and lattice DWT algorithm on CUDA architecture - 5.0 Maxwell.

has a slight time-effectiveness advantage over matrix-based approach for processing and analysis of large data sizes and therefore would be worthwhile consideration for utilization in big data mining and/or processing tasks. However, matrix based approach is structurally simpler and outperforms significantly the lattice structure when one considers smaller data sizes along with larger filter lengths. Based on the presented results one may choose the proper of the two approaches.

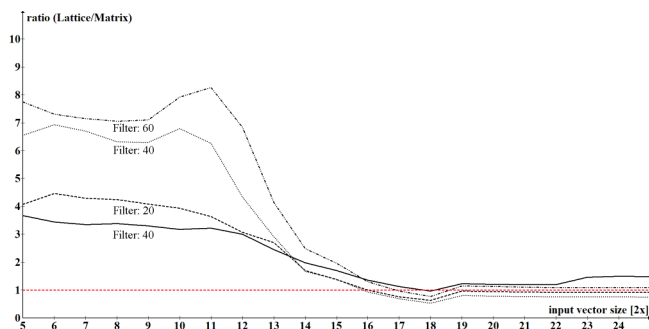


Fig. 14. Lattice DWT to Matrix DWT ratio for chosen filter lengths on CUDA architecture - 6.0 Pascal.

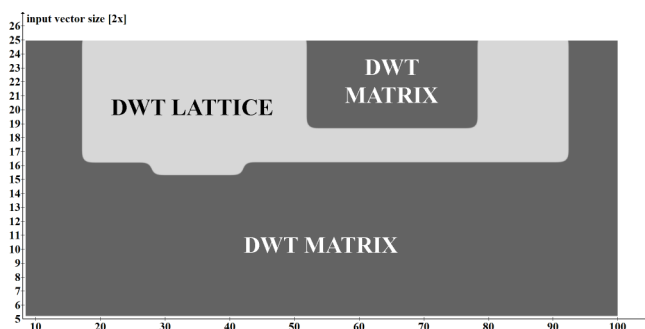


Fig. 15. Performance of matrix and lattice DWT algorithm on CUDA architecture - 6.0 Pascal.

REFERENCES

- [1] A. Nakonechny, Z. Veres, "The Wavelet Based Trained Filter for Image Interpolation", Proc. of IEEE First International Conference on Data Stream Mining & Processing, pp. 218-221, 2016.
- [2] V. Fedak, A. Nakonechny, "Video De-Noising Using Adaptive Wavelet Thresholding", Proc. of IEEE First International Conference on Data Stream Mining & Processing, pp. 222-225, 2016.
- [3] P. Lipiński, J. Stolarek, "Improving Watermark Resistance against Removal Attacks Using Orthogonal Wavelet Adaptation", Proc. 38th Conference on Current Trends in Theory and Practice of Computer Science Location, vol. 7147, pp. 588-599, 2012.
- [4] S. Babichev, "Optimization of Information Preprocessing in Clustering Systems of High Dimension Data", Radio Electronics, Computer Science, Control, no. 2, pp. 135-142, 2014.
- [5] B. Kim, T. McMurry, W. Zhao, R. Wu, A. Berg, "Wavelet-Based Functional Clustering for Patterns of High-Dimensional Dynamic Gene Expression", Journal of Computational Biology, vol. 17, no. 8, pp. 1067-1080, 2010.
- [6] J. Dong, "Data Mining of Time Series Based on Wave Cluster", International Forum on Information Technology and Applications, IFITA '09, DOI: 10.1109/IFITA.2009.132, 2009.
- [7] G. Strang, T. Nguyen, "Wavelets and Filter Banks", Wellesley-Cambridge Press, Wellesley, 1996.
- [8] S. G. Mallat, "A theory for multiresolution signal decomposition: the wavelet representation", IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 11, no. 7, pp. 674-693, 1989.
- [9] M. Yatsymirskyy, "Lattice structures for synthesis and implementation of wavelet transforms", Journal of Applied Computer Science, vol. 17, no. 1, pp. 133-141, 2009.
- [10] M. Yatsymirskyy, "A lattice structure for the two-channel bank of symmetrical biorthogonal filters of lengths $2K+1 / 2K-1$ ", 13-th International Workshop on Computational Problems of Electrical Engineering, Grybów, Poland, 2012.
- [11] D. Puchala, B. Szczepaniak, M. Yatsymirskyy, "Lattice structure for parallel calculation of orthogonal wavelet transform on GPUs with CUDA architecture", Przegląd Elektrotechniczny, vol. R.91, no. 7, pp. 52-54, 2015.
- [12] D. Puchala, K. Stokfiszewski, B. Szczepaniak, M. Yatsymirskyy, "Effectiveness of Fast Fourier Transform implementations on GPU and CPU", Przegląd Elektrotechniczny, vol. 92, no. 7, pp. 69-71, 2016.
- [13] H. Sorensen, "High-Performance Matrix-Vector Multiplication on the GPU", M. Alexander et al. (Eds) Euro-Par 2011 Parallel Processing Workshops, Lecture Notes in Computer Science, vol 7155. Springer, Berlin, Heidelberg, 2012.
- [14] K. Stokfiszewski, K. Wieloch, M. Yatsymirskyy, "The Fast Fourier Transform Partitioning Scheme for GPU's Computation Effectiveness Improvement", Shakhovska N., Stepashko V. (Eds) Advances in Intelligent Systems and Computing II. CSIT 2017, Springer, Cham, vol. 689, no. 1, pp. 511-522, 2017.
- [15] J. Cheng, M. Grossman, T. McKercher, "Professional CUDA C Programming", John Wiley & Sons, Inc. Indianapolis, IN 46256, ISBN: 978-1-118-73932-7, 2014.
- [16] Yatsymirskyy, M., "A Novel Matrix Model of Two Channel Biorthogonal Filter Banks", Metody Informatyki Stosowanej, pp. 205-212, 2011.
- [17] W. van der Laan, A. Jalba, J. Roerdink, "Accelerating Wavelet Lifting on Graphics Hardware Using CUDA", IEEE Transactions on Parallel and Distributed Systems, vol. 22, no. 1, pp. 132-146, 2011.
- [18] L. Cheng, S. Reiji, "An execution time prediction analytical model for GPU with instruction-level and thread-level parallelism awareness", Summer United Workshops on Parallel, Distributed and Cooperative Processing, Hakoshima, Japan, 2011.
- [19] M. Yatsymirskyy, "A lattice structure for the two-channel bank of symmetrical biorthogonal filters of lengths $2K+1/2K-1$ ", Proc. of 13th International Workshop "Computational Problems of Electrical Engineering", Grybów, Poland, September 5-8, 2012.

One Approach of Approximation for Incoming Data Stream in IoT based Monitoring System

Vladyslav Aliexsieiev

Department of Applied Mathematics
Lviv Polytechnic National University
Lviv, Ukraine
vladyslav.i.aliexsieiev@lpnu.ua

Abstract— IoT devices and platforms are a fast growing market. One can mention a number of businesses relying on easy opportunity to build real-time monitoring systems using modern software and IoT hardware solutions. However, the growth has revealed a number of complex problems. Many problems are in area of data processing and storing huge volumes of information. Due to wide use of different kinds of sensors, and even a sets of sensors within each single device, on one hand, practitioners discover unpleasant effects of data losses caused by data packages losses or delays while its transition from sensor to server. On the other hand, huge volumes of data require to use some big data approaches and many startup projects feel the problem of lack of resources. Many of them feel lack of data storage facilities or become unable to support huge data sets due to lack of finance. The paper is focused to research the problem approximation for incoming data stream to make it smaller the volume of data to be stored but to keep it possible to be used. A few approaches to use such data compression via its approximation are discussed with application to IoT based real-time monitoring system.

Keywords— data compression, approximation algorithm, data stream processing, IoT platform, big data

I. INTRODUCTION

A contemporary IT industry trends revealed many new applications of data storages and an IoT became one of the primary trends [1]. Both huge enterprises and small businesses are now dependent from quality of data and quality of data analytics [2]. Some new approaches to understand data and the value of the data had appeared. The industry stepped behind the relational databases and time series has determined new approaches to store and process data [3]. IoT technologies gave an opportunity to design some new platforms for supporting business both with surveillance tools and analytics software applications. This research paper presents some discussion related to IoT bases “real-time” monitoring systems. Common architecture of such platforms are [4]:

- *Device* – a remote computer like Raspberry Pi or any of its alternatives or some custom hardware device that may include a set of sensors (business solution may consist of a whole network of such devices).
- *Internet* – any kind of Internet wireless connection via Wi-Fi, GPRS, 3G/4G or anything else supplied with mobile network (business solution may combine different types of Internet service providers to establish connection).
- *Cloud* – any popular IoT platform to store and process big data.

Adding some software solution to provide data analytics to that IoT platform than it becomes a powerful business tool supporting real-time monitoring. There is a number of companies developing their own platforms or exploiting powerful cloud services to provide their client with such platform as a reliable business solution. Many researchers and practitioners in area of data analysis declare statements similar to [2]: “data accumulation can enable deeper insights and help us to gain more experience and wisdom”. There are many evidences of great performance of time series analysis already and there is a number of solutions for time series databases [3].

However, practical use of such system reveals some serious problems. Many of these problems were predicted earlier. For example, the problem of data uncertainty was known, described and even has some categorization [5]. The era of Big Data has just revealed the complexity of problems, which came with those volume, variety and velocity of big data. Recent researches denoted an importance of data losses problem in monitoring systems [4] and problems of storing big volumes of obsolete data in such systems [6]. One may find some techniques to solve these problems in an analytic manner [7-9]. Some techniques of data clearing allows to aggregate data simultaneously [10], [11].

Unfortunately, there are no common recipes yet in data science to manage with big data sources of any kind. There are some approaches and some of these approaches are well developed, but in some particular cases, there appears the specifics making it difficult or impossible to implement a common solution. Current research can be considered as an alternate or a supplement to those discussions presented in [6]. Unlike to [6] it is offered here not to rely on clustering or quantization, which is appropriate to process obsolete data, but to use approximation as a reliable and a well-developed technique of mathematics.

Generally, we still have the same problem of a large number of devices each with a set of sensors generating huge volume of data. The data from all the devices and all the sensors come to a server (non-relational database). Using these data for online monitoring system allows making some assumptions to ease the solution of the problems.

II. PREREQUISITES AND MEANS FOR SOLVING THE PROBLEM

Let’s define the primary task for the problem solution. First, the aim of the research is to find a way to reduce the number of values in the incoming data flow from the sensors and not to lose the quality of understanding the scope. Second, we’d like to keep some quantitative scope if

possible or to have it in some approximation. Third, we can discuss the ability to recover original data in cases it is necessary. Nevertheless, the last option only remains the option yet.

The matter is, if we consider a real-time monitoring system having a primary target to notify about some critical issues, then it is of less interest to see what exactly normal conditions was surveyed. This allows us to ignore many aspects of value changes within some normal boundaries.

Two sub-problems can be solved simultaneously: 1) an approximation to store values in a database, and 2) an approximation to reduce data volume at the node (peripheral device) or an approximation “on-the-fly” to form smaller packages to be sent to a server. The first problem is rather simple, if to consider only the task to remove the excessive data (to remove less informative values). The second problem is more difficult and requires some actions at the node. The difficulty is that the node will decide about necessity of the values gathered from sensors. This should be made very carefully not to lose an important data. Therefore, the circumstances can be very important to understand whether to implement the approximation at the node. While we have a monitoring system with less analytics purpose, we can assume each node (peripheral device) to gather the same data with respect to its location. This means both problems can be solved successfully.

Now, we can define two key requirements for constructing an appropriate algorithm:

- Simplicity – the algorithm should be easy to implement and fast enough to be used “on-the-fly”.
- Reliability – the algorithm should give a reliable approximation for a data set and hold the information about any abnormal values.

A. Incoming data flow (stream)

Let’s assume the incoming sequence of values to be $f(t)$ with t as a time. Measurements are made with an equal time lapse $\Delta t = const$. This means each next value $f_n := f(t_n)$ is obtained after a fixed period $t_{n+1} = t_n + \Delta t$ and $n = 0 \dots \infty$. This allows considering values as a discrete (Fig. 1). For the purpose of determinateness, the renumbered values presented at Fig. 2. The curve of the input sequence of values presented at Fig. 3. This curve is similar to a signal and one can offer to use some techniques of signal processing. There are many known methods of signal processing and data compression applicable to signals [12]. Unfortunately, we have no evidences of any periodic behavior or repeatable oscillations to be confident to implement those techniques of signal processing. For example, many techniques rely on assumption about definite periodicity in a signal and one of the hardest situations is to use these techniques to process “white noise”. Our “signal” is similar to “white noise”. There are regular appearance of some unpredictable values from sensors. Meanwhile each sensor has some “normal” range of values. This means, there could be a senseless part of values within that range, and some significant values outside the range can be cut as an outliers.

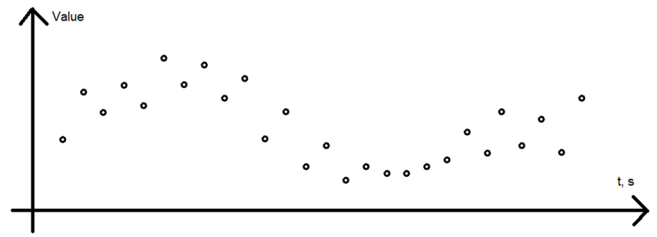


Fig. 1. Input sequence of discrete values (incoming data stream) – measurements with equal intervals of time Δt

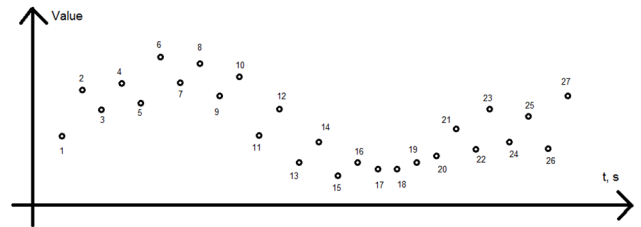


Fig. 2. Number the sequence of values for certainty

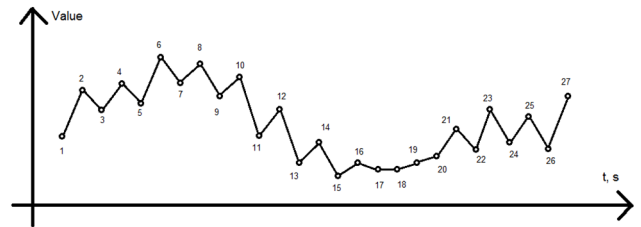


Fig. 3. The curve of the input sequence of values (incoming “signal”)

B. Idea of the algorithm

All values in a data set can be divided into two sub-sets: 1) “maximums” and 2) “minimums”. The “maximums” are the values greater than the previous ones. The “minimums”, otherwise, are the values less than the previous ones. Both the first and the last values in the data set can be marked simultaneously as a “maximum” and a “minimum”. The same simultaneous marking is possible with the consequently equal values. However, this can be an option for the case of equal values to have adequate presentation in resulting approximation. If there were an oscillation character observed, then it would be a rare situation. Thus, the decision about marking the equal values to be made according to necessity. The marking procedure result presented at Fig. 4.

Next, the local extremums can be found within each sub-set as shown at Fig. 5. Values number 1, 6, 18, 23, 25, and 27 are the local extremums among “maximums”. Values number 1, 7, 15, 24, 26, and 27 are the local extremums among “minimums”. Note, the first and the last values are both marked as local extremums.

There can be two strategies to select local extremums: 1) a use of all extremums, 2) a use of selected extremums only.

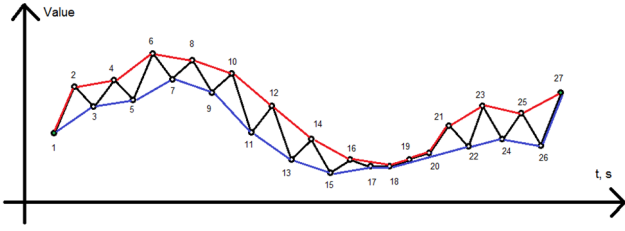


Fig. 4. Split the whole set of values into 2 groups: “maximums” (red, top) and “minimums” (blue, bottom)

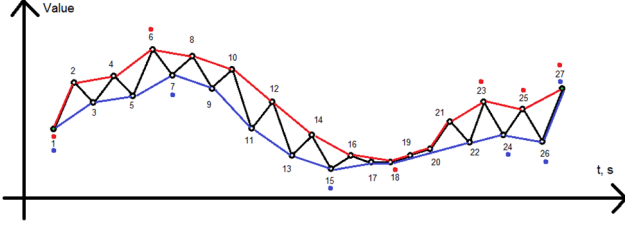


Fig. 5. Extremum selection within each of 2 groups: “maximums” (red, top) and “minimums” (blue, bottom)

According to a strategy of selection of local extremums, one can receive different kind of approximation. Fig. 6 and Fig. 7 present respectively the implementation of first and second strategy. The first one looks to be more accurate, compared to the second one. Nevertheless, each strategy has some advantages and disadvantages, while both give a rough view to data set and a good “compression” or “consolidation”. As it was asserted in [13] the approximation via extremums allows to recover signal using “bell-shaped impulse approximation”.

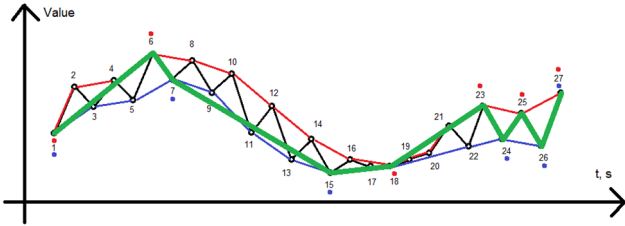


Fig. 6. Approximation for all extremums: “maximums” (red, top) and “minimums” (blue, bottom), approximation (green)

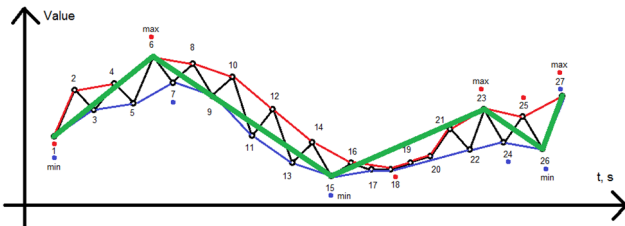


Fig. 7. Approximation for selected extremums – “maximums of maximums” and “minimums of minimums”: “maximums” (red, top) and “minimums” (blue, bottom), approximation (green)

In cases of necessity of further recover of original profile, the algorithm can be replaced to that in [13]. However, in circumstances described above for the purpose of a monitoring system, one may gain great benefits from data consolidation.

III. SOLUTION OF THE EXAMINED PROBLEM

Hence, for certainty, let’s choose an approximation by extremum in the form of a selected extremums (this mean to

use “maximum of maximums and minimums of minimums” according to those shown in Fig. 7). Now, we can consider how this algorithm is formally defined and what are the cases it can be used (implemented).

A. Peculiarities of implementation of approximation with extremums

There are two cases of possible application for consolidation of data: 1) aggregation of data already stored in the database, and 2) data aggregation “on-the-fly”, which can be performed at the node. Both are equivalent to the problems we established initially.

One should note that the offered approach for the algorithm is quite convenient. It requires simultaneously only a few values of data for calculations. This yield the calculations “on-the-fly” with just one previous f_{n-1} value and one current f_n value. It is also necessary to have a three previous values to fix “maximums” ($f_2^{max}, f_1^{max}, f_0^{max}$) and a three previous values to fix “minimums” ($f_2^{min}, f_1^{min}, f_0^{min}$), having indexes 0, 1, and 2 to stand for last, previous to the last and second previous values.

Another aspect is to have actual pair of values presenting both the value and a timestamp. Sure, in case of fixed time intervals $\Delta t = const$ for the values measurement, there is no need to fix the timestamp. Even in case of a time interval or a “window”, just the initial time t_0 supplemented with value order number is enough. However, when we disorder the time series, we need to fix a timestamp for each value. Anyway it is now possible to fix (to store) in the database only the extremums and not the whole sequence of values.

B. The formal algorithm

Now it is possible to describe a formal algorithm with the steps starting from new value f_n has come.

1. A **new value f_n is obtained** (if not then go to Step 3 – the end of the algorithm), and make a data values “shift” accordingly.

1.1. If the new value is greater than the previous $f_n > f_{n-1}$ then the value should be marked as “**maximum**”:
 $f_2^{max} = f_1^{max}$, $f_1^{max} = f_0^{max}$ and $f_0^{max} = f_n$,
 and fix the time $t_1^{max} = t_0^{max}$, $t_0^{max} = t_n$ – the time is needed further to fix the extremum (local maximum).

1.1.1. If $f_2^{max} < f_1^{max}$ and $f_1^{max} > f_0^{max}$ then the value $f_1^{max}(t_1^{max})$ is an extremum:
 $f_k^{extr} = [t_1^{max}, f_1^{max}(t_1^{max})]$, $k = k + 1$.

1.2. If the new value is less than the previous $f_n < f_{n-1}$ then the value should be marked as “**minimum**”:
 $f_2^{min} = f_1^{min}$, $f_1^{min} = f_0^{min}$ and $f_0^{min} = f_n$,
 and fix the time $t_1^{min} = t_0^{min}$, $t_0^{min} = t_n$ – the time is needed further to fix the extremum (local minimum).

1.2.1. If $f_2^{min} > f_1^{min}$ and $f_1^{min} < f_0^{min}$ then the value $f_1^{min}(t_1^{min})$ is an extremum:
 $f_k^{extr} = [t_1^{min}, f_1^{min}(t_1^{min})]$, $k = k + 1$.

1.3. If the new value is equal to the previous $f_n = f_{n-1}$ then the value should be marked **both** as “**maximum**” and “**minimum**”:

- 1.3.1. Execute actions of Step 1.1.
- 1.3.2. Execute actions of Step 1.2.
2. If $k = K$ then the value is a limit (maximum allowed value) and Step 2.1 to be execute, otherwise – Step 2.2.:
 - 2.1. Fix the array of values f_k^{extr} (this means to save in database or send the package to server). Note: it is actually assumed here to have an array of pairs “time–value” $[t_k, f_k]$.
 - 2.2. Go back to Step 1.
3. Execute Step 2.1 over the array of values f_k^{extr} (means to process the rest of the values not fixed yet) and finish the algorithm execution.

IV. RESULTS AND DISCUSSION

The algorithm of approximation by extremums is very easy to understand, easy to implement, and easy to support. There are some problems, and the primary problem is, that the algorithm represents some kind of compression with losses. Nevertheless, those conditions of its application for a monitoring system match the key demand to keep the outliers and not to consider the inner normal range of values (due to its senseless in a common way).

However, there are some approaches to supply ability of approximate signal recovery in case of application of some special techniques. This can a very promising approach for many data science purposes, for IoT based platforms. Due to serious reduce of necessary volume for data storage, one can find it possible to use traditional RDBMS in some areas instead of big data sources.

The practical implementation of the algorithm to real data set at temperature surveillance system gives the average compression up to 10 times compared to initial volume. On one hand, this result can be considered a very particular case of a particular system, but, on the other hand, it relies on a strong mathematics, so it is rather consistent.

V. CONCLUSION

The algorithm in general is quite simple to implement both in case of aggregation of existing data from the

database, and in case of processing data “on-the-fly” at the node (peripheral device). Parameter K allows to set some value analysis “window”, so that one can adjust the accepted volumes of data to be “fixed” (at the database server, or to send the packet from node to server). This ease of use gives a great opportunity to make a software solution even with a “weak” hardware. Meantime, any of preferred task (at the peripheral device or at server) can be solved successfully.

REFERENCES

- [1] A. Oram. *Scaling Data Science for the Industrial Internet of Things*. O’Reily, 2017.
- [2] Y. Lin, and W. Xiao. *Implementing a Smart Data Platform: How Enterprises Survive in the Era of Smart Data*. O’Reily, 2017.
- [3] T. Dunning, and E. Friedman. *Time Series Databases: New Ways to Store and Access Data*. O’Reily, 2015.
- [4] V. Aliksieiev, and O. Gaiduchok, “About the problem of data losses in real-time IoT based monitoring systems,” *Proceedings of International Scientific Conference “Mathematical Modeling” (Borovets, Bulgaria, December 13–16, 2017)*, STUME “Industry 4.0”, Sofia, Bulgaria, Year I, vol. 1/1, pp.10–11, 2017
- [5] C. J. Date. *Database in Depth: Relational Theory for Practitioners*. O’Reilly, CA, 2005.
- [6] V. Aliksieiev, G. Ivasyk, V. Pabyrivskiy, and N. Pabyrivska, “Big data aggregation algorithm for storing obsolete data,” *Proceedings of International Scientific Conference “High Technologies. Business. Society 2018” (Borovets, Bulgaria, March 12–15, 2018)*, STUME “Industry 4.0”, Sofia, Bulgaria, Year II, iss. 1 (3), vol. I “High Technologies”, pp.113–115, 2018.
- [7] P. Bruce, A. Bruce, *Practical Statistics for Data Scientists*. O’Reily, 2017.
- [8] M. Milton. *Head First Data Analysis*. O’Reily, 2009..
- [9] A. B. Downey. *Think Stats*. O’Reily, 2015.
- [10] A. Jain, M. Murty, and P. Flynn, “Data Clustering: A Review,” *ACM Computing Surveys*, vol. 31, no. 3, pp. 264-323, 1999.
- [11] D. Müllner, “Modern hierarchical, agglomerative clustering algorithms,” *ArXiv.org*, 2011. – <https://arxiv.org/pdf/1109.2378.pdf>
- [12] S. W. Smith. *The Scientist and Engineer's Guide to Digital Signal Processing*. California Technical Publishing, 1997.
- [13] N. V. Myasnikova, M. P. Beresten, and M. P. Stroganov, “Approximation of multi extremum functions and its applications to technical systems,” *Herald of higher education institutions. Volga region. Engineering sciences*, no. 2 (18), pp.113–119, 2011. [In Russian]

Software Architecture Design of the Real-Time Processes Monitoring Platform

Anatoliy Batyuk
*Dept. of Automated Control Systems
Institute of Computer Science
and Information Technologies
Lviv Polytechnic National University
Lviv, Ukraine
abatyuk@gmail.com*

Volodymyr Voityshyn
*Dept. of Automated Control Systems
Institute of Computer Science
and Information Technologies
Lviv Polytechnic National University
Lviv, Ukraine
voityshyn@gmail.com*

Volodymyr Verhun
*Dept. of Automated Control Systems
Institute of Computer Science
and Information Technologies
Lviv Polytechnic National University
Lviv, Ukraine
vverhun@gmail.com*

Abstract — Understanding of how business processes are executed in real-life is vitally important for a company. Any process leaves a digital footprint that can be transformed into so-called event logs and analyzed with process mining techniques. A software platform with the purpose of near real-time processes monitoring is implemented. Design of the represented platform is based on the lambda architecture combining online and offline process mining algorithms with advanced analytics based on machine learning.

Keywords — *process mining, event data, event logs, business process management, BPM, XES, lambda architecture*

I. INTRODUCTION

Any event in the surrounding world is not by itself but belongs to some processes. IT systems, that have become ubiquities nowadays, help to automate vast amount of various kind of processes either in personal everyday lives or in huge enterprises. Most main stream software development practices did not consider process nature of the tasks they are devoted to automate “hardcoding” logic of workflow steps in source code. Consequently, it has been developed huge amount of software products which from one hand automate quite complex processes but from the other do not incapsulate any explicit definition of the implemented workflows losing the connection between the implementation and real-life. Nonetheless, within the industry field called business process management (or BPM) it has been developed a lot of practices to deal with workflows including their visual modeling (e.g. BPEL, BPMN, Petri nets etc.) and appropriate software implementations supplying wide range of products from powerful business process management systems (e.g. IBM BPM, Oracle BPM) to software components that can be embedded to a particular application (e.g. jBPM, Activiti, Camunda). Another trend that has had considerable influence on software industry is data science. The goal of applying data science techniques is to make software more intelligent obtaining insights from accumulated data. However, like classical software development practices most data science algorithms do not consider process nature of analyzed data.

Process mining is a discipline that fills in the gap between the mentioned above three industry domains. Significant contribution into creation of the academical core of process mining, its further promoting and encouraging industrial applications has been made in Eindhoven Technical University (The Netherlands) under direction of professor Wil M.P. van der Aalst.

Guiding principle #1 declared in Process Mining Manifesto [1] states that event data (or event logs) is a primary data source for process mining. Like the entire software development industry process mining has faced with the challenge to deal with increasing amount of event data. In practice datasets are not static but are constantly fed with new data. This circumstance requires to handle data streams in near real-time mode and consequently puts process mining techniques into the position when it is necessary to deal with incomplete process instances.

Current paper is devoted to architecture design of the implemented by the authors software platform with the purpose of near real-time processes monitoring. The visualization and analytics modules of the represented system are empowered with advanced process mining and machine learning algorithms.

The rest of the paper is organized as follows: statement of the technical task is provided in section II; architecture significant requirements are specified in section III; the solution architecture design is described in section IV; in section V technical implementation details are provided; section VI contains the results of validation whether the designed system meets the performance requirements; the built-in analytics module is briefly described in section VII; short overview of the already existing process mining software products is provided in section VIII; conclusion remarks are in section IX.

II. TASK STATEMENT

The software platform represented in current paper is general enough to be applied to wide range of practical tasks related to near real-time processes monitoring. However, it is obvious that it is hardly possible to implement a unified software product that can be applied to online process mining tasks in different business domains without modifications. That is why the described system is designed as an extensible platform with wide range of configuration capabilities so that it can be adopted to a particular application needs with minimal efforts, extended with specific features and integrated with other software systems.

As specified above one of the primary requirement is that the system takes event data from continuous data streams. It is assumed that data streams consist of items in XES format [2].

The received event data is stored “forever” in the system’s internal storage. Manual or automate data archiving outside the system is out of scope.

One of the main functional requirements is to support process flow chart visualization discovered by means of process mining techniques [3]. The visualized process model has to be updated in near real-time mode in accordance with receiving events from data streams. A similar system with real-time dashboards is described in [4]. Dealing with process model concept drift [5] is out of scope now and planned for the future.

The analytics module should include features defined in the online process mining framework [6]. The system supports prediction when a process instance completes, suggestions of next steps which are considered as optimal by the system and alerting if actual state of a process instance breaks predefined rules. The specified analytics features function in near real-time mode.

III. NON-FUNCTIONAL REQUIREMENTS

Non-functional requirements to technical architecture of the platform are specified in current section. The requirements listed below are a subset of the quality attributes defined in [7]. The taken software architecture building approach is based on the attribute-driven design method [8].

A. Performance

Characteristics of the performance are defined by the requirement which states that the system has to process event data in near real-time mode (see section II). For current task the latency and throughput are highly important. Latency is the interval from the time of receiving of an event till the time when the end user sees the changes caused by the event (e.g. in the process model visual representation). In turn, throughput refers to number of events processed by the system during a certain period of time.

B. Scalability

In current context scalability stands for ability to vary latency and throughput of the system according the changes of number of received events. From practice standpoint it is necessary to decide whether the system is intended to deal with BigData or “small” data. The reason of necessity to make such decision on the architecture design phase is that the implementation and maintenance cost of a BigData solution is much higher in comparison with the similar solution for “small” data. So, the decision is: the described platform is not intended to deal with BigData. Implementation of the platform modification with BigData support is planned for the future.

C. Interoperability

The represented platform should be so-called cloud agnostic which means that it can be deployed at clouds of different providers (e.g. Azure, AWS, Google Cloud Platform) or use on-premises infrastructure of a customer.

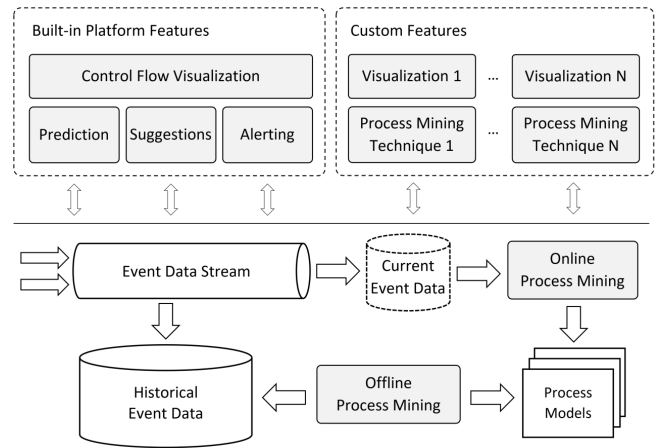


Fig. 1. Architecture concept of the platform

D. Extensibility

As mentioned above the designed software is not a product ready to use without any modifications but it is an extensible platform with predefined architecture and implemented basic built-in functionality. This means that the platform is going to be extended with specific features necessary for a particular customer (e.g. adding an anomaly detection module with appropriate visualization and alerting functionalities).

E. Configurability

In current context configurability means that the platform is flexible enough to be adopted to needs of specific applications without changing the source code. For example, rules for the alerting feature can be defined considering specific of a monitored process.

IV. ARCHITECTURE CONCEPT

High level architecture design of the platform is represented in current section. The concept meets the functional and non-functional requirements specified in sections II and III respectively.

As it is already defined the major requirement is near real-time processing of event data. There are two architecture patterns that address such a task: (a) lambda architecture [9] and (b) kappa architecture [10]. The core idea of the first pattern is that data processing is split into two layers: (a) speed and (b) batch. The speed layer is accountable for handling newly received data in near real-time mode whilst the batch layer deals with accumulated historical data. Kappa architecture is derived from the lambda. The main difference is that batch layer is omitted in kappa architecture simplifying the implementation of the pattern. Hence, kappa architecture is not applicable for the tasks that need batch processing.

The designed platform definitely requires batch layer since process mining techniques (e.g. the implemented process discovery algorithm, see section VII) use historical event data. This is the reason behind choosing lambda architecture as a primary design pattern.

Architecture concept of the platform is depicted on Fig. 1. Parts of the lambda architecture are adopted to the purposes of current task. Offline and online process mining components represent batch and speed layers of the lambda architecture respectively. Another important aspect of the designed architecture is that events of incomplete process instances are kept in a separate storage (“Current Event Data” on Fig. 1) so that relevant data is accessed with minimal latency by the online process mining techniques.

Since lambda architecture is mostly applied to BigData tasks it possible to modify the platform to deal with BigData without changes in its conceptual design. A similar lambda architecture-based BigData system is represented in [11].

V. TECHNICAL SOLUTION

A. Components Model

The components model of the platform (Fig. 2) is the next step of the design process after architecture concept. The model is designed flowing the decision that the platform is not intended to deal with BigData (see section III). Event data stream is supposed to be a message queue. One of the benefits of applying this pattern is that it ensures reliable message delivery. The online process mining algorithms are handlers that listen to the queue for new messages. Received event data are persisted in the database. The main requirement to this database is to be optimized for time series data. Additionally, the most recent event data (including events of incomplete process instances) is cached in an in-memory database which significantly minimizes latency of processing and visualizing this data. Results of execution of the integrated process mining and machine learning algorithms are stored in a NoSQL database. The reason of this decision is that such kind of databases supports unstructured data and are fast on reading. Presentation layer of the platform is implemented as a pluggable single page web application. The server side is composed with microservices and exposes a RESTful API for the frontend.

B. Technology Stack

The service side technology stack is mostly Java-based. One of the reason behind this is that most process mining algorithms are implemented with Java [12]. Another reason is high quality of cross-platform support provided by Java. The frontend side is built with HTML 5 and CSS 3 using the latest JavaScript standard – ECMAScript 6. In particular, process model visualization is implemented with SVG and D3JS. The business rules engine component is intended to address the configurability requirements (see section III). The DMN engine from the Camunda platform is used for its implementation.

VI. PERFORMANCE MEASUREMENT OF THE MESSAGE QUEUE

As a messaging technology RabbitMQ v.3.7.4 was chosen. The reason of taking this particular message broker is that it is an open source mature solution with scalability and high-availability support. Additionally, RabbitMQ is not intended to work with BigData which means that it consumes not so much resources as a similar BigData solution (e.g. Apache Kafka). Since the message queue is a single point of failure of the platform a load test was performed to measure its capabilities. The performance test was executed on the following environment:

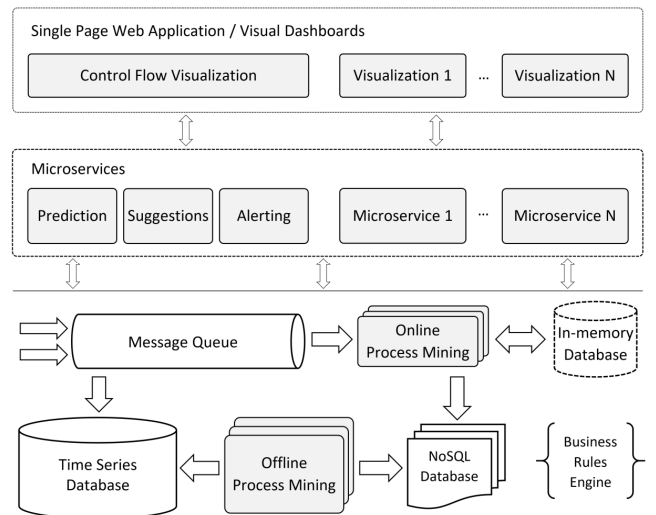


Fig. 2. Components model of the platform

RabbitMQ v.3.7.4 (one producer and one consumer), Linux CentOS 7, 12 GB RAM, Intel Core i7 – 4500U 1.8. GHz. The RabbitMQ management plugin was used to collect the metrics. The results of the test (Table 1) proves that RabbitMQ v.3.7.4 is stable and supports acceptable level of performance.

TABLE I. PERFORMANCE MEASUREMENT OF RABBITMQ V.3.7.4

Message Size, bytes	Measured RabbitMQ Metrics		
	Average Latency, ms	Average Message Publish Rate, msg/s	Average Message Delivery Rate, msg/s
100	3800	55100	20015
300	4100	49800	1893
100000	9055	4400	980

VII. BUILT-IN ANALYTICS FEATURES

As specified in section II flow chart visualization based on a process discovery algorithm is a built-in feature of the platform. One of the oldest and most well-known process discovery technique is the alpha algorithm [6]. It takes event data and produces a Petri net. However, the alpha algorithm is not the best choice for real-life processes with a lot of transitions especially if the results are for business domain experts who are not process mining professionals. The Fuzzy Miner algorithm [13] is more suitable for such cases. The efficiency of this algorithm has been proved by experience of well-known process mining software like Disco [14] and Celonis [15]. Additionally, comparison of the Fuzzy Miner with some other process discovery algorithms is done in [16]. Considering the facts above the offline process discovery implemented within the scope of the platform is based on the Fuzzy Miner algorithm. In turn, the online process discovery implementation follows the ideas outlined in [17]. The built-in prediction analytics is used to forecast completion time of a process instance. The feature is developed upon a combined time series forecasting information technology based on fuzzy experts’ evaluation [18] and analysis of dynamic processes [19]. The suggestions feature recommends an optimal process flows and suits the human behavior (which is important if people are involved into monitored processes). Another requirement to this

feature is the ability of incremental learning from the event data stream. To meet these requirements an implementation of a neuro-fuzzy model [20, 21] is integrated.

VIII. OVERVIEW OF EXISTING PROCESS MINING SOFTWARE

Process mining is a relatively new academic discipline and its software implementations began to gain popularity in the market not so long ago. The oldest process mining tool is ProM [12] which is an open source Java-based framework. Scientists are the target audience of this application. Disco [14] is a commercial process mining tool. It includes the most useful algorithms. This product is used by experts from business domains who are not process mining professionals. Another process mining product is Celonis [15]. It is a fast-growing German startup. Target consumers' audience of Celonis is medium and big enterprises. This product supports offline and online process mining and has connectors to other software, for example SAP [22]. The main difference between the software system developed by the authors and Celonis is that Celonis is a product with a set of features delivered to all its customers whilst the represented platform is a ground for custom development with predefined architecture and initial set of built-in features.

IX. CONCLUSIONS

The implemented platform has been integrated with an energy efficiency management system [23] as an extension with the purpose to monitor real-life processes on the operator control level. The visualization feature has provided visibility on the processes executed within the energy management system and generated alerts once a process instance breaks the predefined restriction rules. From practice standpoint it is necessary to include the following algorithms to the set of built-in features: (a) conformance checking [6] and (b) handling process concept drifts [5]. Another way of the platform's evolution is to design and implement a version with the purpose to support Big Data.

REFERENCES

- [1] W. M. P. van der Aalst, et al., "Process mining manifesto," in Business Process Management Workshops. BPM 2011. Lecture Notes in Business Information Processing, Berlin, Germany, vol. 99, pp. 169-194, 2011.
- [2] IEEE Standard for eXtensible Event Stream (XES) for Achieving Interoperability in Event Logs and Event Streams, IEEE Std 1849-2016, 2016.
- [3] W. M. P. van der Aalst, T. Weijters, and L. Maruster, "Workflow mining: discovering process models from event logs," in IEEE Transactions on Knowledge and Data Engineering, vol. 16, no. 9, pp. 1128-1142, 2004.
- [4] A. Batyuk and V. Voityshyn, "Business Processes Monitoring by Means of Real-Time Visual Dashboards," 6th International Academic Conference on Information, Communication, Society 2017, Lviv, Ukraine, pp. 204-205, 2017.
- [5] R. P. Jagadeesh Chandra Bose, Wil M. P. van der Aalst, I. Žliobaitė, and M. Pechenizkiy, "Handling Concept Drift in Process Mining," in Advanced Information Systems Engineering. CAiSE 2011. Lecture Notes in Computer Science, London, UK, vol. 6741, pp. 391-405, 2011.
- [6] W.M.P. van der Aalst, Process mining: data science in action. Berlin Heidelberg: Springer-Verlag, 2016.
- [7] M. Barbacci, M. H. Klein, T. H. Longstaff, and C. B. Weinstock, Quality Attributes. SEI at Carnegie Mellon University, Pittsburgh, Pennsylvania, Rep. CMU/SEI-95-TR-021, 1995.
- [8] L. Bass, P. Clements, and R. Kazman, Software Architecture in Practice, 3rd ed. Addison-Wesley Professional, 2012.
- [9] "Lambda Architecture", [Online]. Available: <http://lambda-architecture.net/>. [Accessed: 25 Mar 2018].
- [10] "kappa-architecture.com", [Online]. Available: <http://milinda.pathirage.org/kappa-architecture.com/>. [Accessed: 25 Mar 2018].
- [11] A. Batyuk and V. Voityshyn, "Apache storm based on topology for real-time processing of streaming data from social networks," 2016 IEEE First International Conference on Data Stream Mining & Processing (DSMP 2016), Lviv, Ukraine, pp. 345-349, 2016.
- [12] H. M. W. Verbeek, J. C. A. M. Buijs, B. F. van Dongen, and W. M. P. van der Aalst, "ProM 6: the process mining toolkit," in Proceedings of the Business Process Management 2010 Demonstration Track, Hoboken NJ, USA, vol 615, pp. 34-39, 2010.
- [13] W.G. Christian, W.M.P. van der Aalst, "Fuzzy Mining – Adaptive Process Simplification," in Proceedings of the 5th International Conference on Business Process Management (BPM 2007), Brisbane, Australia, vol 4714, pp. 328-343, 2007.
- [14] Ch. W. Günther, and A. Rozinat, "Disco: Discover Your Processes," in Proceedings of the Demonstration Track of the 10th International Conference on Business Process Management (BPM 2012), Tallinn, Estonia, vol 940, pp. 40-44, 2012.
- [15] F. Veit, J. Geyer-Klingenberg, J. Madrzak, M. Haug, and J. Thomson, "The Proactive Insights Engine: Process Mining meets Machine Learning and Artificial Intelligence," in 15th International Conference on Business Process Management (BPM 2017). BPM Demo Track and BPM Dissertation Award, Barcelona, Spain, vol 1920, pages 5, 2017.
- [16] A. Rozinat, "ProM Tips - Which Mining Algorithm Should You Use?", 2010. [Online]. Available: <https://fluxicon.com/blog/2010/10/prom-tips-mining-algorithm/>. [Accessed: 25 Mar 2018].
- [17] A. Burattin, "Process Mining for Stream Data Sources," in Process Mining Techniques in Business Environments. Lecture Notes in Business Information Processing, Cham, Springer, vol 207, pp. 177-204, 2015.
- [18] O. Mulesa, F. Geche, A. Batyuk, and V. Buchok, "Development of Combined Information Technology for Time Series Prediction," in Advances in Intelligent Systems and Computing II (CSIT 2017), Lviv, Ukraine, vol 689, pp. 361-373, 2017.
- [19] P. Bydyuk, A. Gozhyj, I. Kalinina, and V. Gozhyj, "Analysis of uncertainty types for model building and forecasting dynamic processes," in Advances in Intelligent Systems and Computing II (CSIT 2017), Lviv, Ukraine, vol 689, pp.66-82, 2017.
- [20] Ye. Bodyanskiy, I. P. Pliss, D. Peleshko, Yu. Rashkevych, and O. Vynokurova, "Hybrid Generalized Additive Wavelet-Neuro-Fuzzy-System and its Adaptive Learning," in The Eleventh International Conference on Dependability and Complex Systems DepCoS-RELCOMEX., Brunow, Poland, pp. 51-61, 2016.
- [21] Ye. Bodyanskiy, G. Setlak, D. Peleshko, and O. Vynokurova, "Hybrid Generalized Additive Neuro-Fuzzy System and its Adaptive Learning Algorithms," 8th IEEE International Conference on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications, Warsaw, Poland, pp. 328-333, 2015.
- [22] "Showcase: SAP Process Mining by Celonis", [Online]. Available: <https://www.sap.com/developer/showcases/process-mining-by-celonis.html>. [Accessed: 25 Mar 2018].
- [23] T. Teslyuk, I. Tsmots, V. Teslyuk, M. Medykovskyy, and Y. Opotyak, "Architecture and Models for System-Level Computer-Aided Design of the Management System of Energy Efficiency of Technological Processes at the Enterprise in Advances," in Intelligent Systems and Computing II. CSIT 2017, Lviv, Ukraine, vol 689, pp. 538-557, 2017.

Deep Neural Network for Image Recognition Based on the Caffe Framework

Myroslav Komar
Research Institute for Intelligent Computer Systems
Ternopil National Economic University
Ternopil, Ukraine
mko@tneu.edu.ua

Pavlo Yakobchuk
Research Institute for Intelligent Computer Systems
Ternopil National Economic University
Ternopil, Ukraine

Vladimir Golovko
Department of Intelligent Information Technologies
Brest State Technical University
Brest, Republic of Belarus
gva@bstu.by

Vitaliy Dorosh
Research Institute for Intelligent Computer Systems
Ternopil National Economic University
Ternopil, Ukraine

Anatoliy Sachenko
Kazimierz Pulaski University of Technology and Humanities in Radom
Radom, Poland
sachenkoa@yahoo.com

Abstract— Deep Learning of the Neural Networks has become one of the most demanded areas of Information Technology and it has been successfully applied to solving many issues of Artificial Intelligence, for example, speech recognition, computer vision, natural language processing, data visualization. This paper describes the developing the deep neural network model for image recognition and a corresponding experimental research on an example of the MNIST data set. Some practical details for creating the Deep Neural Network and image recognition in the Caffe Framework are given as well.

Keywords— Deep Neural Network, Information Technology, Image Recognition, Artificial Intelligence, Caffe Framework

I. INTRODUCTION

In order to proceed efficiently with large amounts of data at the acceptable time, special information technologies are needed. Nowadays such information technologies can be represented by Deep Neural Networks [1-8], which have the greater efficiency of the non-linear transformation and data representation in comparison with traditional neural networks. A Deep Neural Network performs a deep hierarchical transformation of images in the input space. Moreover the Deep Neural Networks, thanks to the multi-layer architecture, enable to process and analyze the large amount of data, as well as modeling the cognitive processes in various fields. Currently, most high-tech companies in the US (Microsoft, Google, Facebook, Baidu, etc.) use deep neural networks to design the various intelligent systems. According to the scientists of the Massachusetts Institute of Technology, deep neural networks are on the list of the 10 most promising high technologies capable in the near future to largely transform the everyday life of most people on our planet and solve many problems of artificial intelligence, for example, speech recognition, computer vision, natural language processing, data visualization, etc. [9-13, 26].

II. RELATED WORKS

In 2006, Hinton proposed a greedy layer-wise algorithm [1], which became an effective tool for teaching deep neural networks. It was shown that a deep neural network has a greater efficiency of the non-linear transformation and data representation in comparison with a

traditional perceptron. This network performs a deep hierarchical transformation of the input space. As a result, the first hidden layer separates the low-level space of attributes of the input data, the second layer detects the space for attributes for a higher level of abstraction, etc. [14]. Currently there are many works devoted to the recognition of images by means of deep neural networks [15-19]. Authors [15] present the Maxout network in Network architecture. Their approach is based on the convolutional layer and a two layer maxout MLP and it's used to convolve the input and average pooling in all pooling layers. In [16], the architecture of the deep neural network is applied in biology domain. The small receptive fields of convolutional winner-take-all neurons yield large network depth are resulting in roughly as many sparsely connected neural layers. Ikuro Sato in [17] offers an optimal decision rule for a given data sample using classifiers that are trained on extended data. A paper [18] reports to introducing the DropConnect, a generalization of Hinton's Dropout for regularizing large fully-connected layers within neural networks. As a result authors derive a bound on the generalization performance of both Dropout and DropConnect. A simple and effective stochastic pooling strategy is developed [19] to secure over-fitting during the training deep convolutional networks. According to the MNIST [21, 22] the best generalized recognition accuracy was 99.79% [18]. So, we propose below how to improve this value.

III. STRUCTURE OF DEEP NEURAL NETWORK

For the implementation of the above-mentioned architecture, we used Caffe deep learning library [20]. The main advantage of Caffe is the speed of operation. The framework supports CUDA and, if necessary, can switch the processing flow between the processor and the graphics card. The process of training the Deep Neural Network in framework Caffe has been lasted 30 epochs and finished with achieving the given accuracy of learning (Fig. 1). The deep neural network consists of the following layers (Fig. 2): 1st layer – Convolution (out filters: 24, size: 5x5, stride: 1x1); 2nd layer – Convolution (out filters: 12, size: 5x5, stride: 1x1); 3rd layer – Pooling (size: 2x2, stride: 2x2); 4th layer – Convolution (out filters: 8, size: 5x5, stride: 1x1); 5th layer – Convolution (out filters: 4, size: 5x5, stride: 1x1); 6th layer – Pooling (size: 2x2, stride: 2x2); 7th layer – InnerProduct (out:

500, filter: xavier); 8th layer – InnerProduct (out: 0, filter: xavier); 9th layer – Softmax (out: 10, activation: softmax).

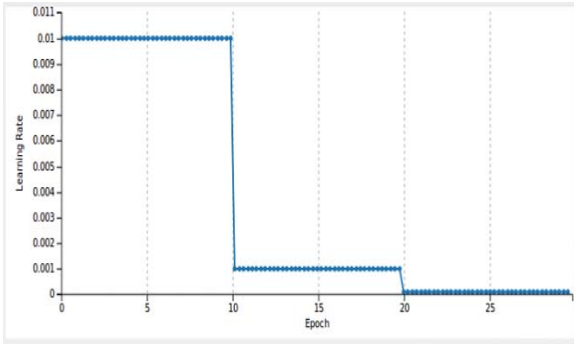


Fig. 1. The process of learning the deep neural network

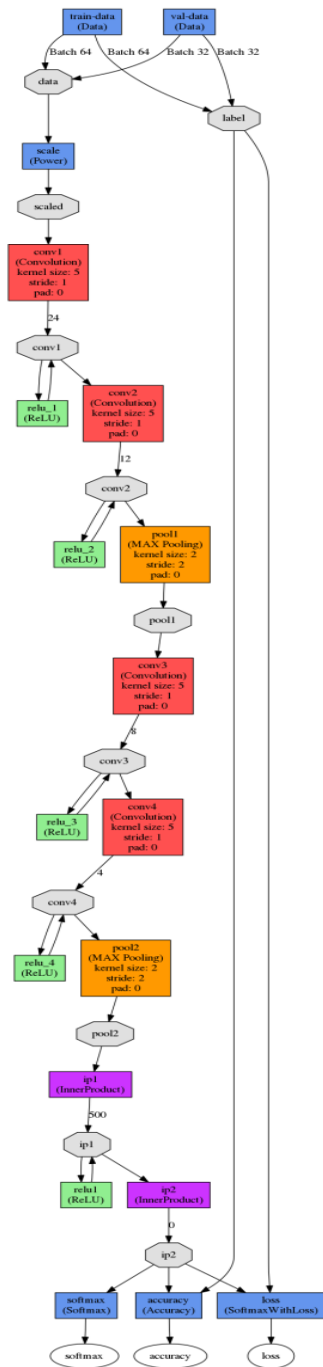


Fig. 2. The structure of the deep neural network

IV. EXPERIMENTAL RESULTS

The image from the test sample of the MNIST data set [21, 22] was used for case study. The MNIST data set consists of 28x28 pixel handwritten digital images organized in 10 classes (0 to 9) with both 60,000 training and 10,000 test samples. Testing on this data set has performed without increasing the data. Results, of the image recognition for the number 0 to 9 from the test sample are show in Fig. 3, the generalized recognition accuracy was 99.93%. The visualization of the image, which is applied to the input of the Deep Neural Network is illustrated by Fig. 4. Fig. 5 shows the visualization of the image processing on the first convolution layer, 24x24x24 (24 functional maps with the element 24x24).

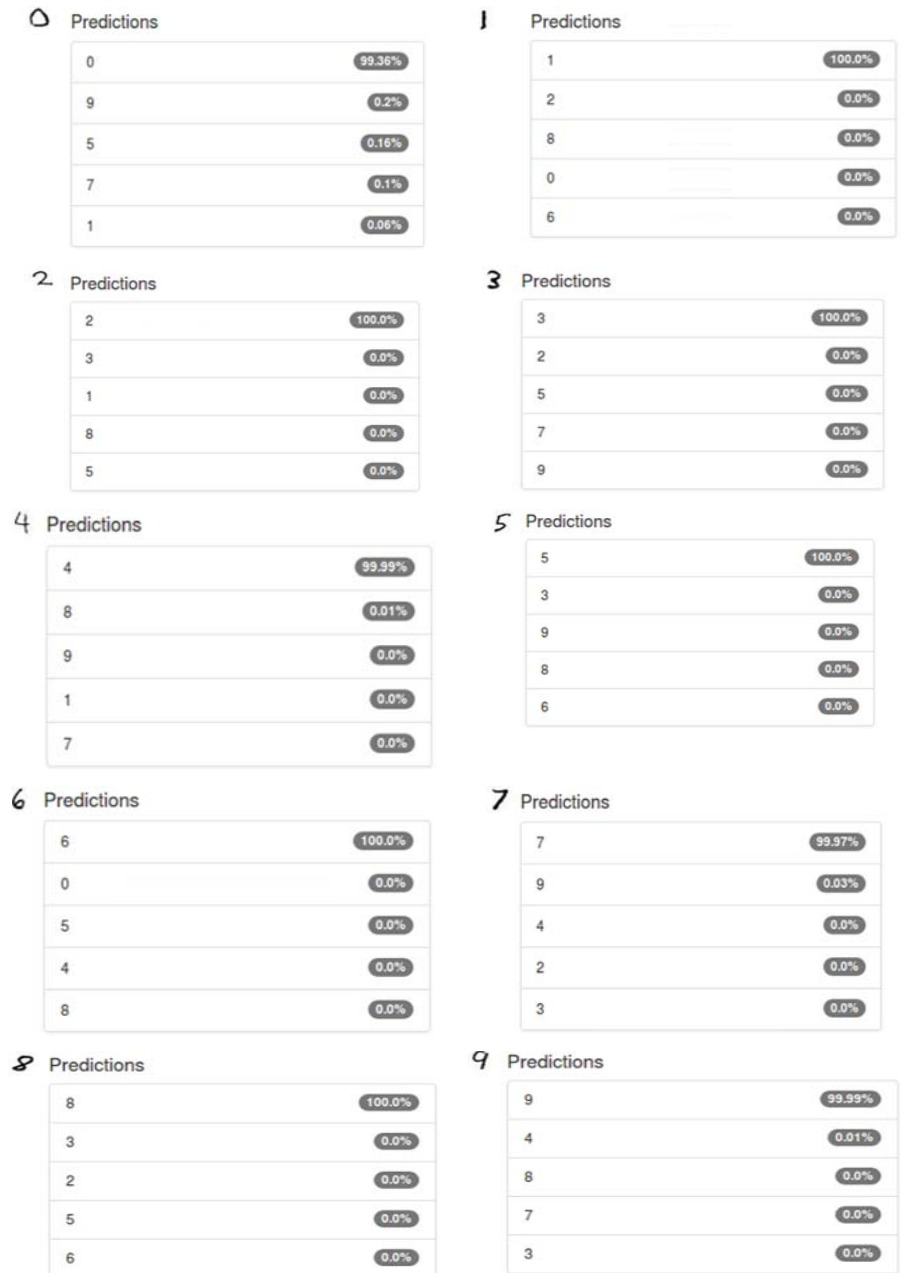


Fig. 3. Results of image recognition from the test sample

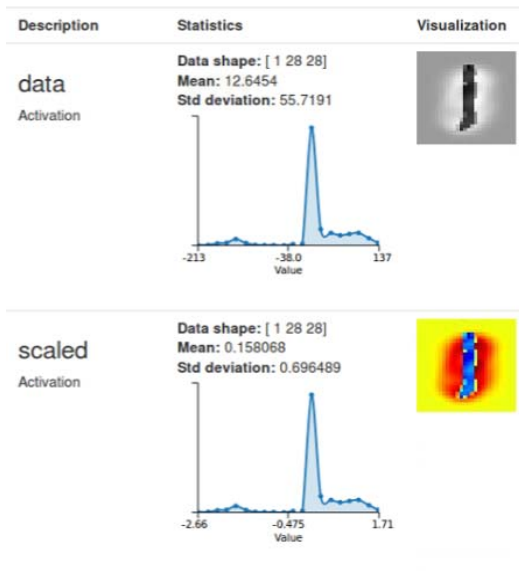


Fig. 4. Visualization of the input image

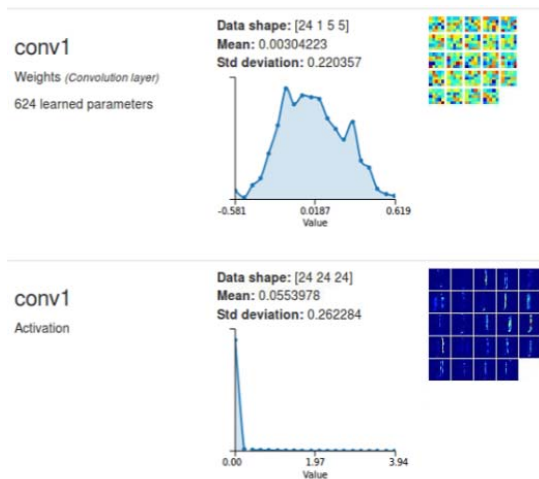


Fig. 5. Image processing on the first convolutional layer

The visualization of image processing on the second convolution layer, 12x20x20 (12 functional maps with the element 20x20) is illustrated by Fig. 6. Fig. 7 shows the visualization of image processing on the first layer of the spatial association.

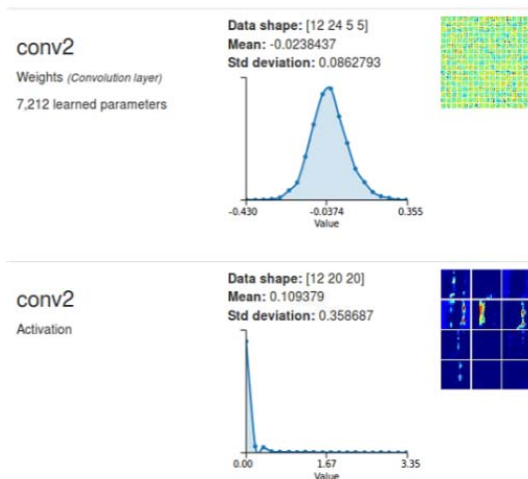


Fig. 6. Image processing on the second convolutional layer

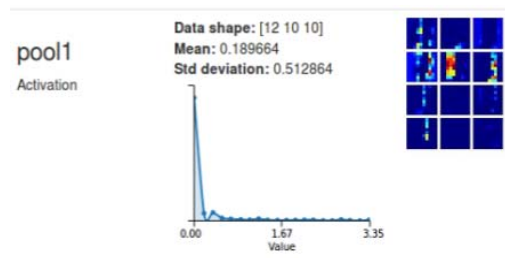


Fig. 7. Image processing on the first layer of the spatial association

The visualization of the image processing on the third convolution layer, 8x6x6 (8 functional maps with the 6x6 element) is illustrated by Fig. 8. Fig. 9 shows the image processing of the image on the fourth convolution layer, 4x2x2 (4 functional maps with the element 2x2).

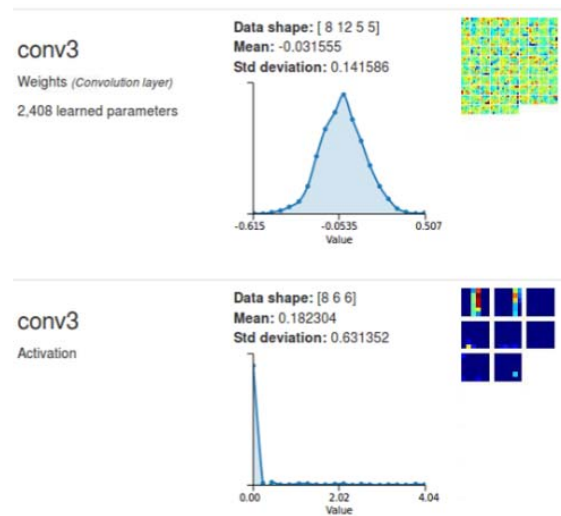


Fig. 8. Image processing on the third convolutional layer

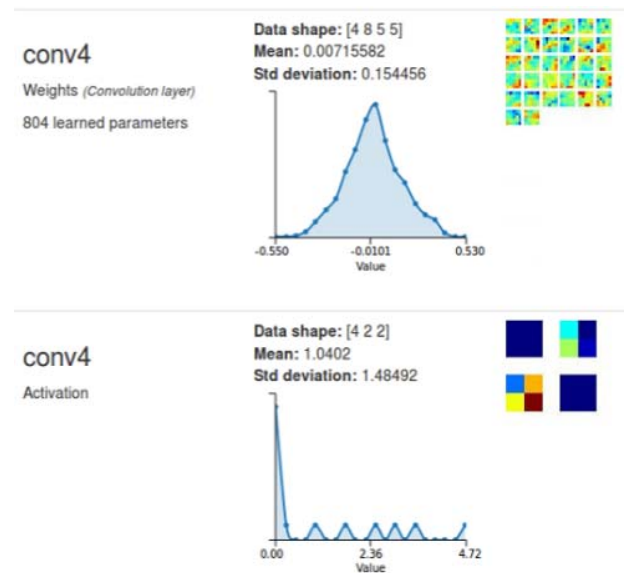


Fig. 9. Image processing on the fourth convolutional layer

In Fig. 10 the visualization of image processing on the second layer of spatial association is illustrated.

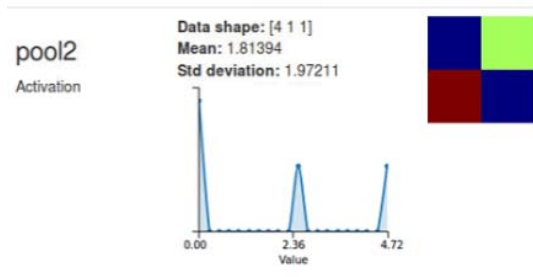


Fig. 10. Image processing on the second layer of the spatial association

The visualization of image processing on the first, second and Softmax full layer is illustrated by Fig. 11, 12 and 13 correspondingly.

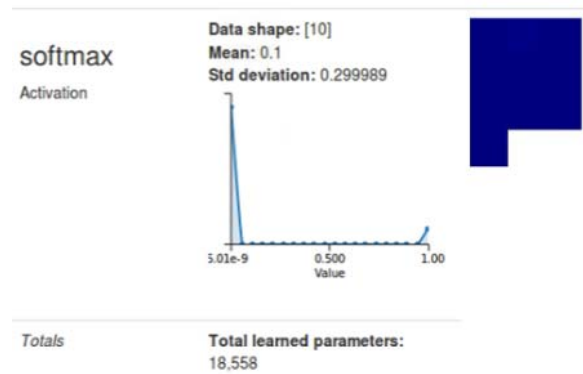


Fig. 13. Image processing on the Softmax layer

As it comes from Fig. 3 above the reached generalized recognition accuracy is equal 99.93%, that is 0.14% better in comparison with a work [18].

V. CONCLUSION AND FUTURE WORK

Authors propose a model of the Deep Neural Network for the recognizing the images of handwritten digits, using the structure of the neural network in the Caffe Framework. Experimental results have been carried out on an example of the MNIST data set and the generalized recognition accuracy was 99.93%.

Employing the deep neural network in Big Data processing is one of perspective direction for a future research. Moreover, it is planned to conduct experimental research on the following data sets CIFAR-10/100 [23], SVHN [24], and ImageNet [25].

ACKNOWLEDGMENT

This work is carried out within the framework of the project of the Ministry of Education and Science of Ukraine «Methods of intellectual processing and analysis of big data based on deep neural networks», 2018-2019 years.

REFERENCES

- [1] G. E. Hinton, S. Osindero, and Y. Teh, "A fast learning algorithm for deep belief nets," *Neural Computation*, vol. 18, pp. 1527–1554, 2006.
- [2] G. E. Hinton, *A practical guide to training restricted Boltzmann machines*, Department of Computer Science, University of Toronto, 2010.
- [3] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521 (7553), pp. 436–444, 2015.
- [4] Y. Bengio, "Learning deep architectures for AI," *Foundations and Trends in Machine Learning*, vol. 2(1), pp. 1–127, 2009.
- [5] X. Glorot, A. Bordes, and Y. Bengio, "Deep sparse rectifier networks," In *Proc. of the 14th Int. Conf. on Artificial Intelligence and Statistics (AISTATS)*, vol. 15, pp. 315–323, 2011.
- [6] V. Golovko, A. Krushchanka, U. Rubanau, and S. Jankowski, "A Learning Technique for Deep Belief Neural Networks," *Communication in Computer and Information Science*, vol. 440, pp. 136–146, 2014.
- [7] V. Golovko, A. Krushchanka, V. Turchenko, S. Jankowski, and D. Treadwell, "A New Technique for Restricted Boltzmann Machine Learning," *8th IEEE International Conference on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications (IDAACS'2015)*, Warsaw, Poland, pp. 182–186, 24–26 September, 2015.
- [8] V. Golovko, A. Krushchanka, and D. Treadwell, "The Nature of Unsupervised Learning in Deep Neural Networks: A New

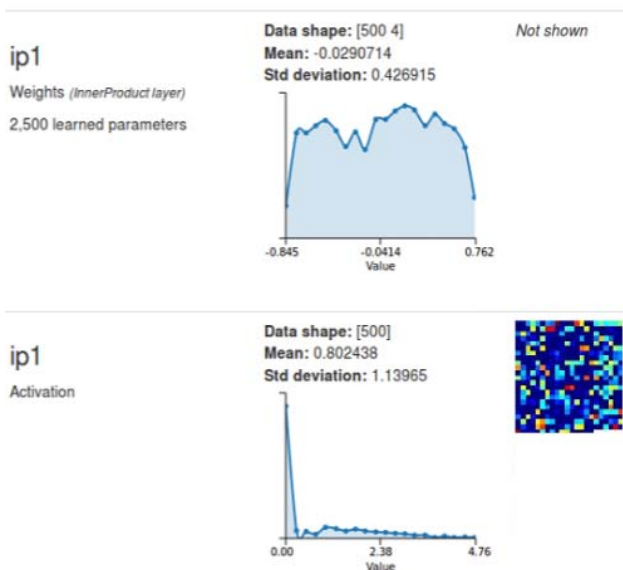


Fig. 11. Image processing on the first full layer

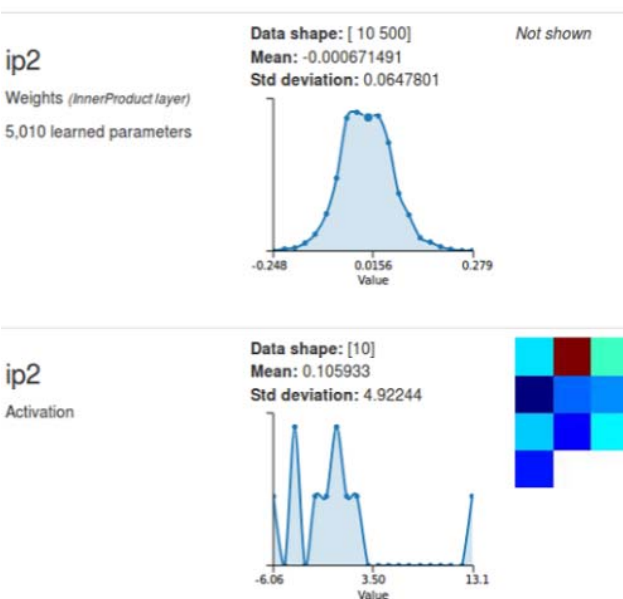


Fig. 12. Image processing on the second full layer

- Understanding and Novel Approach,” *Optical Memory and Neural Networks*, vol. 25(3), pp. 127–141, 2016.
- [9] S. Jankowski, Z. Szymański, U. Dziomin, V. Golovko, and A. Barcz, “Deep learning classifier based on NPCA and orthogonal feature selection,” *International Conference on Photonics Applications in Astronomy, Communications, Industry, and High-Energy Physics Experiments*, Wilga, Poland, pp. 5–9, May 29, 2016.
- [10] G. Hinton, et al., “Deep neural network for acoustic modeling in speech recognition,” *IEEE Signal Processing Magazine*, vol. 29, pp. 82–97, 2012.
- [11] T. Mikolov, A. Deoras, D. Povey, L. Burget, and J. Černocký, “Strategies for training large scale neural network language models,” in *Automatic Speech Recognition and Understanding*, pp. 195–201, 2011.
- [12] A. Krizhevsky, L. Sutskever, and G. Hinton, “ImageNet classification with deep convolutional neural networks,” in *Advances in Neural Information Processing Systems*, vol. 25, pp. 1090–1098, 2012.
- [13] V. Golovko, S. Bezobrazov, A. Kroshchanka, A. Sachenko, M. Komar, and A. Karachka, “Convolutional Neural Network Based Solar Photovoltaic Panel Detection in Satellite Photos,” *9th IEEE International Conference on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications (IDAACS’2017)*, Bucharest, Romania, pp. 14–19, September 21–23, 2017.
- [14] G. Hinton, and R. Salakhutdinov, “Reducing the dimensionality of data with neural networks,” *Science*, vol. 313 (5786), pp. 504–507, 2006.
- [15] Jia-Ren Chang, and Yong-Sheng Chen, “Batch-normalized Maxout Network in Network,” arXiv:1511.02583, 2015.
- [16] D. Ciresan, U. Meier, and J. Schmidhuber, “Multi-column deep neural networks for image classification,” *25th IEEE conference on computer vision and pattern recognition (CVPR)*, New York, pp. 3642–3649, 2012. DOI: 10.1109/CVPR.2012.6248110, 2012.
- [17] I. Sato, H. Nishimura, and K. Yokoi, “APAC: Augmented PAttern Classification with Neural Networks,” arXiv:1505.03229v1, 2015.
- [18] L. Wan, M. Zeiler, S. Zhang, Y. Le Cun, and R. Fergus, “Regularization of Neural Networks using DropConnect,” *Proceedings of the 30th International Conference on Machine Learning*, PMLR, vol. 28(3), pp. 1058–1066, 2013.
- [19] M. D. Zeiler and R. Fergus. “Stochastic pooling for regularization of deep convolutional neural networks,” ArXiv:1301.3557, 2013.
- [20] Caffe Deep Learning Framework, <http://caffe.berkeleyvision.org>, last accessed 15.03.2018.
- [21] The MNIST database, <http://yann.lecun.com/exdb/mnist>, last accessed 15.03.2018.
- [22] Y. Le Cun, L. Bottou, Y. Bengio, and P. Haffner. “Gradientbased learning applied to document recognition,” *Proceedings of the IEEE*, vol. 86(11), pp. 2278–2324, 1998.
- [23] A. Krizhevsky, and G. Hinton, Learning multiple layers of features from tiny images. Technical report, University of Toronto, 1 (4), 7, 2009. <https://www.cs.toronto.edu/~kriz/learning-features-2009-TR.pdf>, last accessed 15.03.2018.
- [24] Y. Netzer, T. Wang, A. Coates, A. Bissacco, B. Wu, and A. Y. Ng, “Reading digits in natural images with unsupervised feature learning,” In *NIPS workshop on deep learning and unsupervised feature learning*, Granada, Spain, vol. 2011, pp. 5. 2011.
- [25] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. FeiFei, “Imagenet: A large-scale hierarchical image database,” In *CVPR09*, pp. 248–255, 2009.
- [26] D. T. V. Dharmajee Rao, and K. V. Ramana, “Winograd’s Inequality: Effectiveness for Efficient Training of Deep Neural Networks,” *Intelligent Systems and Applications*, vol. 6, pp. 49–58, 2018..

A Framework for Semantic Video Content Indexing Using Textual Information

Sadek Mansouri
LATICE Laboratory
Higher institute of computer science
Mednine, Tunisia
mansouri_sadek@hotmail.fr

Mbarek Charhad
Al Madina Al Mounawra, (KSA) 41411
Kingdom of Saudi Arabia.
Taiba university
mbarek.charhad@gmail.com

Ali Rezik
Higher institute of computer science
Gabes University
Mednine, Tunisia
alirekik1@yahoo.com

Mounir Zrigui
LATICE Laboratory FSM of
Monastir
Monastir, Tunisia
mounir.zrigui@fsm.rnu.tn.

Abstract—In these last years, many works have been published in the video indexing and retrieval field. However, except some specific cases such as a sport video where it's possible to estimate the set of important events and concepts in the document, this research is generally limited to analyzing low level content. In this paper, we introduce an approach for semantic video indexing that combines two levels of descriptions. First, we extract automatically textual information from video frames. The second part of our approach consists to exploit linguistic techniques and semantic network in order to extract semantic concepts such as person identity, location name, event type etc. These informations are then used for semantic description of video content. Our proposed approach was tested on video collection of Arabic TV news and experimental results have been satisfying.

Index Terms—Arabic news video, semantic indexing, text detection, conceptual network.

I. INTRODUCTION

The quantity of audiovisual information has increased dramatically with the emergence of the high-speed Internet and TV channels. In addition, the technological advances in recent years in the field of informatics (storage a reas m ore and more considerable, digitization of data, etc.) have helped to simplify the use of data videos in various areas by the public. The complexity of video data at the level structure and heterogeneity has been the source of various research work. The major challenge of the latter is the establishment of systems to allow the user, even casual, to access and interpret easily the video data. In this context, the description of the content of a document video through the indexing process is a decisive step. In effect, the indexing is present upstream of any treatment approach of video data. The indexing is the operation that is to extract a digital signature or text, which describes the content accurately and concisely. The success of this step depends, as well, the success of any process of access to video data. Text embedded especially the artificial text in video frames is one of the important semantic features of the video content analysis. This type of text is artificially added to the video at the time of editing and

provides highlevel information of video content that seems to be a useful clue in the multimedia indexing system. Usually, it provides information about when, where and who elements of the news video events. However, text detection and localization in the video frame is still a challenging problem due to the numerous difficulties resulting from the variety of text features (size, color, and style), the presence of complex background and conditions of video acquisition. The second problem concerns the extraction of knowledge from textual data in order to provide relevant and accurate information. This poses a challenge to the scientific community that must be able to propose effective systems for the extraction of information in particular with the diversity of fields applications and the peculiarity of the studied language .

To treat these various problems, we propose in this article an approach of video indexing using the semantic contents of document. This approach is based on a conceptual description of the contents. Each video document is described by list of concepts (person, localities, etc.). This description makes possible to abstract the semantic content resulting from various sub-media (image, audio, text). . The main challenge is how extract semantic information from text signal in order to provide a high description of video content .

The rest of this paper is organized as follows: In Section 2, we presented state-of-the-art of semantic video indexing systems . Section 3 presents an overview of video indexings levels. In section 4 details the experimentation of proposed approach , followed by conclusions in Section 5.

II. STATE OF THE ART

In this part, we present a categorization of approaches and methods proposed in the literature for video modeling and retrieval. There are two basic classes.

The first class focuses on low-level features extraction [1] [2] from audiovisual information such as color, shape, texture or motion that characterize visual low level content. The major disadvantage of these approaches is the lack of

semantic description. However, users can't express their query to retrieve video segment using semantic description. These kinds of systems don't efficiently resolve the problem of video parsing that exploits semantic content.

Second, the semantic information that makes physical information comprehensible by user. This second level makes possible to support the "interface" between the user and the machine and to exploit thus the video contents more easily. To make possible the complementarity between the two points of views, it will be necessary to design an approach that exploits in the same time the semantic and the signal content [3].

In [6], the authors propose a multilingual information extraction (IE) system for annotating sports videos in English, German, and Dutch using ASR (Automatic speech recognition) tools. The IE components of this system include tools for tokenizing, part-of-speech tagging, knowledge extraction, and coreference resolution. [7], the systems aim is to perform automatic knowledge extraction from Italian TV news. This system also utilizes an ASR tool to obtain the video texts and IE techniques (named entities recognition). Another semantic video annotation application called Rich News has been described in [8], where the authors make use of the resources on the web to enhance the indexing process. The overall system contains the following modules: automatic speech recognition, key-phrase extraction from the speech transcripts and searching the video using key phrases. Moreover, the proposed system allows also manual annotation to ameliorate segmentation results. [9] a system has been implemented to annotate Turkish news video using video text as a source of information and IE techniques including named entity recognition, person entity extraction, co-reference resolution, and semantic event interpretation. For better knowledge, our work presents the first attempt for semantic Arabic news video indexing based on text analysis and information extraction (IE) techniques that subsume low and conceptual features of video content.

III. PROPOSED SYSTEM

In this part, we present an overview of our semantic video indexing system. Fig.1 illustrates the framework of the proposed system, which are based on three levels. The first level puts a focus on low-level processing such as video segmentation, text detection and recognition. The second level seeks for extracting the semantic concepts including named entity such as a name of person, organization, location and event. In the final step, our work is based on the construction and a semantic network that addresses the taxonomic and contextual relations between concepts. This step aims to enhance the semantic content in terms of indexes generated by the second step. We detail the different stages of our proposed system and their goals in the following sub-paragraphs.

A. Level 1: Low-level processing

Key-frames extraction: In this work, we have applied a temporal segmentation based on the following assumption the text in the image requires at least two seconds to be readable

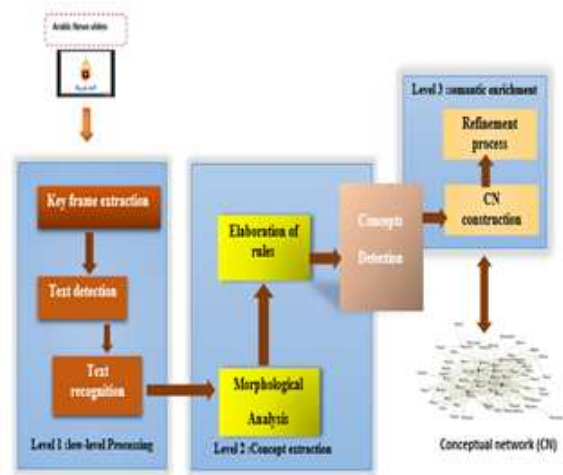


Fig. 1. Proposed system of Arabic news video indexing.

by the user, to generate shots. Then for each video shot, the middle image will be selected as a key-frame.

Text detection and localization : After key-frame extraction, text information is detected and extracted from each key-frame. Our text detection method relies on two necessary steps: text detection and text validation. The first step detects connected components (CC) using a hybrid method which combines MSER and edge information. These CC are then grouped by mathematical morphology operators to form candidate text regions. The second stage aims to remove non-text region using geometric constraints and specific signature of Arabic script called baseline (see Fig.2).

Text recognition: After text detection in the video frame, the next step target is to segment and binarize text region in order to separate it from the rest of the frame using Otsus global thresholding method. An optimal threshold is calculated on the tributions of text pixels and non-text pixels. The method abasis of the grey level histogram by assuming Gaussian disims to maximize the interclass variance. In the last stage, commercial OCR engine ABBYY FineReader has been applied for the recognition of text news. More details are shown in our papers [10] [11].

B. Level 2: Conceptual level

A person who is watching a video summarizes it in general by using concepts (identity of person, name of place, etc.), the subject (politic, sport, business, etc.) and sometimes actions to specify these descriptions. This constitutes a way to video content representation. The target of such representation is mainly to get for video document a list of marked points that facilitate access and re-use of content. Considering the heterogeneity of the content from a point of view data (image, audio, and text) and semantic. Indeed, for each video segment, we can associate multiple possibilities of interpretation that can be assorted by specific / generic relationship. We consider that when we have a description issued from a specific media, it consists a way for categorizing the content. For example,



Fig. 2. Text detection : (a) original image ,(b) MSER extraction in image(a) ,(c) Image mask integrating MSERs and canny edges (d) open result ,(e)candidate text regions ,(f) final result.

when we use the action speak or speaking about we suppose that the description is related to audio content. This makes easier the distinction between the multiple possibilities of interpreting the content of the same video segment. In our work, concepts such as person name, location, organization, etc are extracted automatically using NLP techniques such as Morphological Analysis and linguistic rules.

1) *Morphological Analysis* : Firstly, we segment text in words based on spaces delimiter. Then, we proceed to a In the second step, we parse transcriptions files to extract named entities by comparing each item to the three concepts classes (person identity, the name of a city and organization). This procedure is based on the projection of each news text on the list of keywords called gazetteers. Gazetteers are of a varied nature: lists of first names for the recognition of person names, cities names for the detection of location, etc. Each list is associated with a semantic label which shall be the type of annotation

2) *Elaboration of rules* : Due to Arabic language complexity and specific characteristics, we also exploit a set of Lexical triggers to extract the name of the person, location and organization not covered by the gazetteer resources (see Table1). To do this, we have used three kinds of rules to improve concepts detections process. This task is object of

this publication [12].

TABLE I
A SAMPLE SET OF LEXICAL TRIGGER

Named entity	lexical triggers
Person	وزير , قائد , العقيد , أستاذ , دكتور , مستشار , نائب
Organisation	منظمة , حزب , مؤسسة , جمعية , وزارة , شركة
Location	منطقة , قرية , مدينة , ريف , بلدة , قطاع

The extracted semantic information such as name of person, location, organization and event class is used to annotate the video text and to improve the searching using metadata. The original description are attached to the news video as xml file.

C. Level 3: Semantic enrichment

The semantic enrichment process aims to enrich the semantic interpretation and further enhance the performance of semantic indexing and multimedia retrieval content systems. This task consists of two steps:

1) *Construction of conceptual network*:: This network consists of set concepts which refers to the politic domain and linked by arcs. The latter denote semantic and contextual relations between concepts nodes.

2) *Refinement process*: Given an initial set of indexes $C = c_1, \dots, c_n$, the refinement process consists in selecting the most related concepts among the conceptual network(CN) . In the remainder of this work, we will try to propose a measure which we use for the calculation of the relatedness between a candidates concepts in CN and a given set of indexes C.

IV. EXPERIMENTATION AND RESULT EVALUATION

A. corpus

In order to evaluate the performance of our proposed system in terms of robustness and effectiveness, we used a set of 20 video news (10,000 images) that have been collected from different Arabic TV channels: Aljazeera, Alarabiya, Wataniya 1, Elmayadeen, RT-arabe over the period of September 15 ,2017 until the 5th of December, 2017 and they have a total duration of about two hours. The videos have been automatically transcribed leading to a transcription text of 9704 words. Besides, the named entities extraction phase is done with Farassa ¹platform using Gazetteers and lexical triggers as linguistic resources.

B. Results

1) *text detection*: .A comparative study with previous systems is performed using precision, recall as the evaluation measures. We applied the evaluation method that has been proposed for the AcTiV-DB Test set, together with evaluation results reported in [13]especially many-to-one matches method. Table II shows that the proposed system achieves excellent results for Aljazeera channel and it is able to outperform the other methods .We can notice the excellent precision rate of

¹<http://qatsdemo.cloudapp.net/farasa/>

TABLE II
RESULTS OF THE TEXT DETECTION METHOD

Channel	Method	Precision	Recall
HD(Aljazeera)	Chen [14]	0.67	0.56
	Zayene [4]	0.85	0.83
	our system	0.90	0.87
SD(france 24)	Chen [14]	0.45	0.52
	Zayene [4]	0.75	0.73
	our system	0.71	0.70
SD(RTArabic)	Chen [14]	0.63	0.52
	Zayene [4]	0.73	0.73
	our system	0.75	0.74

our method. This is due to the good rejection ability of false alarms using baseline descriptor. However, this higher score has been decreased For SD channels. This is explained by the fact that the text in these channels is not clearer and the poor quality of graphic text as shown in Fig.3.



Fig. 3. Some detection results from three different SD channels

2) *Concepts Extraction:* As shown in table III, the results may be satisfactory achieving 80.52% as overall of F-measure. The main reason for these results is the use of grammars rules, which permit the detection of Named entities more precisely. For event extraction , the conceptual feature improve the classification results compared to other approach which based only on textual feature .

TABLE III
EXPERIMENTAL RESULTS OF THE CONCEPTS EXTRACTION METHOD

Concept	Precision	Recall	F-measure
Person	83.02%	79.56%	81.25%
Location	80.23%	77.62%	78.90%
Organisation	82.5%	80.35%	81.41%
Overall			80.52%
Event	85 %	80.3%	82.78%

V. CONCLUSIONS

In this paper, we have introduced a semantic approach for Arabic videos news based on text analysis process and concepts extraction techniques. The experimentation and the evaluation results are promising.

In future work, we will try to improve our concept extraction tool by implementing other rules that cover all structure of Arabic text. In addition, we plan also to use other visual features to enhance detection task especially for video frames with low resolutions.

REFERENCES

- [1] C. Zhu, C.-E. Bichot and Liming Chen. Image region description using orthogonal combination of local binary patterns enhanced with color information. *Pattern Recogn.*, vol. 46, no. 7, pages 1949-1963, July 2013.
- [2] R. Vieux, J. Benois-Pineau and J.-P. Domenger. Content based image retrieval using bag-of-regions. In *Proceedings of the 18th international conference on Advances in Multimedia Modeling, MMM, 2012*.
- [3] Yu. Ye, Xu. Rong, X. Yang, Y. Tian: Region Trajectories for Video Semantic Concept Detection. *ICMR 2016: 255-259*.
- [4] O. Zayene, J. Hennebert, S. M. Touj, R. Ingold, and N. E. "Text Detection in Arabic news Video Based on the SWT Operator and Convolutional Auto-encoders". In *Proc of 12th IAPR Workshop on Document Analysis Systems 2016*.
- [5] Ch. Lhioui, A. Zouaghi, M. Zrigui "Realization of Minimum Discursive Units Segmentation of Arab Oral Utterances". *Int. J. Comput. Linguistics Appl.* 7(1): 31-50 (2016)
- [6] H. Saggion, H. Cunningham, K. Bontcheva, D. Maynard, O. Hamza, and Y. Wilks, "Multi-media indexing through multi-source and multi-language information extraction: the MUMIS project," *Data and Knowledge Engineering*, vol. 48, pp. 247264, 2004.
- [7] R. Basili, M. Cammisa, and E. Donati, "RitroveRAI: A web application for semantic indexing and hyperlinking of multimedia news," in *Proceedings of the International Semantic Web Conference (ISWC)*, 2005.
- [8] M. Downman, V. Tablan, H. Cunningham, and B. Popov, "Web-assisted annotation, semantic indexing and search of television and radio news," in *Proceedings of the International Conference on World Wide Web (WWW)*, 2005.
- [9] D. Kucuk, AYazc. "A semi-automatic text-based semantic video annotation system for Turkish facilitating multilingual retrieval". *Expert Systems with Applications*, 40(9), 3398-3411.(2013).
- [10] S. Mansouri, M. Charhad, M. Zrigui: "A Heuristic Approach to Detect and Localize Text on Arabic NewsVideo" *Computacin y Sistemas Journal* (2017) in press.
- [11] S. Mansouri, M. Charhad and M. Zrigui. "Arabic Text Detection in News Video based on Line Segment Detector". *International Journal of Research in Computing Science*(2017).
- [12] S. Mansouri, Ch. Lhioui, M. Charhad and M. Zrigui . "Text-to-concept: a semantic indexing framework for Arabic News videos" *18th International Conference, CICLing 2017*.
- [13] O. Zayene, J. Hennebert, S. M. Touj, R. Ingold, and N. E. BenAmara. "A dataset for arabic text detection, tracking and recognition in news videos-AcTiV". in *Proc. of (ICDAR)*, Nancy, France, 2015.
- [14] C. Huizhong. "Robust Text Detection in Natural Images with Edge Enhanced Maximally Stable External Regions". *IEEE ICPR 2011*.
- [15] G. Lavee, E. Rivlin, and M. Rudzsky. Understanding video events: A survey of methods for automatic interpretation of semantic occurrences in video. *IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews*, 39(5):489504, 2009.
- [16] H. Saggion, H. Cunningham, K. Bontcheva, D. Maynard, O. Hamza, and Y. Wilks. Multimedia indexing through multi-source and multi-language information extraction: The MUMIS project. *Data and Knowledge Engineering*, 48(2):247264, 2004.
- [17] Y. Zhang, Ch. Xu, Y. Rui, J. Wang, and H. Lu. Semantic event extraction from basketball games using multimodal analysis. In *Proceedings of the IEEE Conference on Multimedia and Expo (ICME)*, pages 21902193, 2007.

Topic #2

Dynamic Data Mining & Data Stream Mining

The Autoencoder Based on Generalized Neo-Fuzzy Neuron and its Fast Learning for Deep Neural Networks

Yevgeniy Bodyanskiy
Control System Research Laboratory
Kharkiv National University of Radio Electronics
Kharkiv, Ukraine
yevgeniy.bodyanskiy@nure.ua

Yuriy Rashkevych
Ministry of Education and Science of Ukraine
Kyiv, Ukraine
rashkevychyuriy@gmail.com

Dmytro Peleshko
IT Step University
Lviv, Ukraine
dpeleshko@gmail.com

Olena Vynokurova
Kharkiv National University of Radio Electronics
Kharkiv, Ukraine
IT Step University
Lviv, Ukraine
vynokurova@gmail.com

Abstract— In this paper the autoencoder based on the generalized neo-fuzzy neurons is proposed. Also its fast learning algorithm based on quadratic criterion was proposed. Such system can be used as part of deep learning systems. The proposed autoencoder is characterized by high learning speed and less number of tuned parameters in comparison with well-known autoencoders of “bottle neck” type. The efficiency of proposed approach has been justified based on different benchmarks and real data sets.

Keywords—autoencoder, deep learning network, neo-fuzzy neuron, fast learning algorithm, data compression.

I. INTRODUCTION

Nowadays the deep neural networks (DNN) [1-4] are becoming more widespread for solving the many type Data Mining tasks, first of all, due to the significantly higher quality of information processing in comparison with conventional shallow neural networks (SNN). But such higher quality is achieved at the cost of very slow learning speed. This fact doesn't allow to use existed DNN in the Data Stream Mining tasks when the information is fed sequentially in online mode. In the connection with that, the reducing the learning time is a very actual problem.

The important part of any DNN is the subsystem of input information compression, which is named an autoencoder that provides the dimensionality reduction of the input vectors-patterns without significant loss of an information. Such reduction process allows avoiding the undesirable effects, which are provided by «curse of dimensionality». One of most well-known autoencoders is the autoassociative multilayer perceptron “bottle-neck”, which provides the optimal information compression, but also needs more learning time.

To overcome the difficulty, which is connected with low learning speed, we can use the hybrid systems of computational intelligence instead of the classical neural networks that are constructed based on the elementary perceptron of F. Rosenblatt. The neo-fuzzy neuron (NFN), which was proposed by T. Yamakawa and co-authors in [5-7] can be used as the structural block of such systems. The NFN is defined by high approximation properties and

simplicity of learning process. It should be noticed, the learning process of the NFN can be optimized by speed [8] because the output signals of the NFNs are linearly dependent on the tuned synaptic weights.

The neuro-fuzzy Kolmogorov's network using the NFNs was introduced by authors in [9-13]. Such network has the two layers for information processing and is characterized by the universal approximation properties according to A. Kolmogorov - V. Arnold theorem. Using these networks, the authors in [14] have proposed the neuro-fuzzy model for the dimensionality reduction, which is learned using error backpropagation algorithm with gradient methods. In [15-17] the optimized learning algorithms for two-layer autoencoders based on the neo-fuzzy neurons were proposed, which allow significantly reducing the learning time.

In the same time the hybrid systems of computational intelligence with many inputs and many outputs, which are constructed based on neo-fuzzy neurons, have an abundant number of membership functions. It is possible significantly to reduce the number of these functions, using the so-called generalized neo-fuzzy neuron (GNFN) [18]. GNFN is the extension of NFN for a multidimension case and contains less number of membership functions.

Therefore, in the paper, the architecture of two-layer autoencoder based on GNFN and its optimized learning algorithms are proposed. Such approach allows reducing the time of information preprocessing in DNN.

II. THE ARCHITECTURE OF AUTOENCODER BASED ON GNFN

The proposed autoencoder has the architecture, which is shown on Fig. 1. Such architecture consists of two sequentially connected layers, which are presented by the generalized neo-fuzzy neurons $GNFN^{[1]}$ and $GNFN^{[2]}$. The input signals $x(k) = (x_1(k), \dots, x_i(k), \dots, x_n(k))^T \in R^n$, $k = 1, 2, \dots, N, \dots$ are fed to the $GNFN^{[1]}$, which consists of n multidimensional nonlinear synapses $MNS_i^{[1]}$, $i = 1, 2, \dots, n$. Each of them has the one input, m outputs, h membership functions $\mu_l^{[1]}(x_i(k))$, $l = 1, 2, \dots, h$ and mh

tuned synaptic weights $w_{ji}^{[1]}$, $j = 1, 2, \dots, m$. The architecture of GNFN^[1] is shown on Fig. 2.

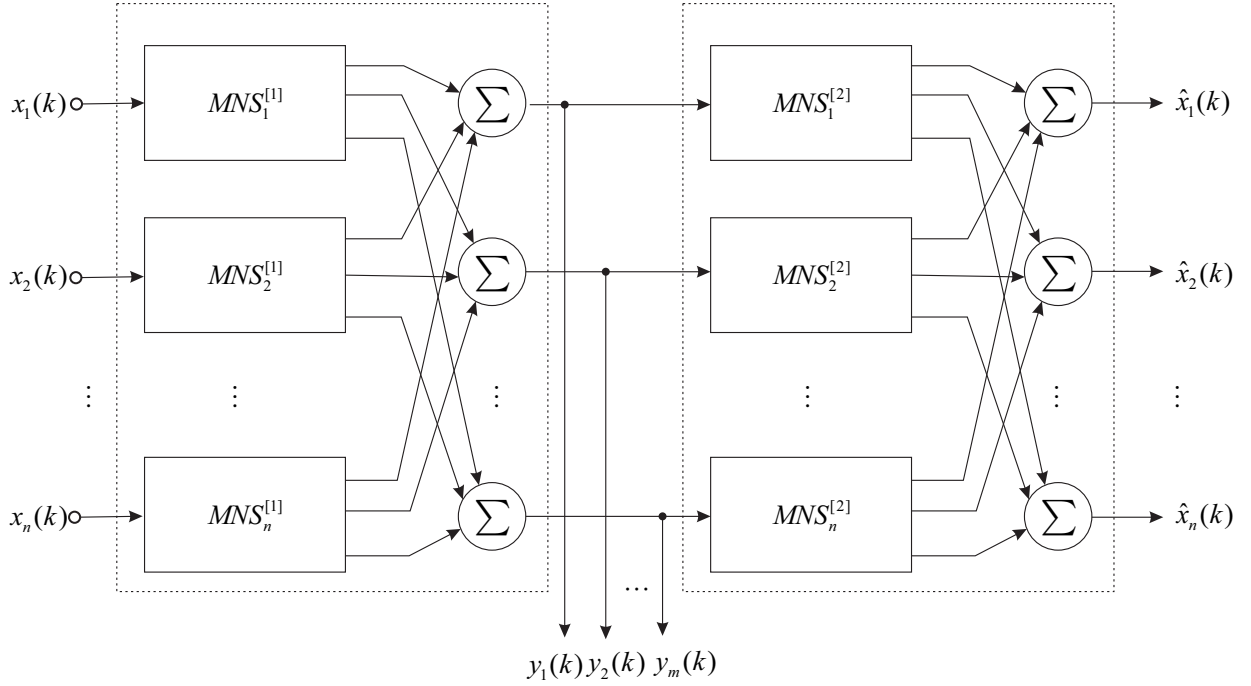


Fig. 1. Autoencoder based on GNFNs

The output of the first autoencoder layer is the compressed vector of signals $y(k) = (y_1(k), \dots, y_j(k), \dots, y_m(k))^T \in R^m$, $m < n$, which at the same time is the output of system in whole. This signal is fed to the inputs of GNFN^[2], which differs from GNFN^[1] only in that it has m inputs, m multidimensional nonlinear synapses $MNS_j^{[2]}$, each of them has one input, n outputs, h membership functions $\mu_{lj}^{[2]}(y_j(k))$, $l = 1, 2, \dots, h$ and nh tuned synaptic weights $w_{ij}^{[2]}$. In total, the autoencoder contains $2nmh$ tuned synaptic weights and $(n+m)h$ membership functions that is significantly less than in the systems, which were described in [14-17].

The output of GNFN^[2] is the recovered vector of input signals $\hat{x}(k) = (\hat{x}_1(k), \dots, \hat{x}_i(k), \dots, \hat{x}_n(k))^T$, at that the less mismatch between $x(k)$ and $\hat{x}(k)$, the higher the quality of the information compression by the autoencoder.

Therefore the autoencoder under consideration is the autoassociative hybrid neo-fuzzy system like «bottle-neck» system.

In general, the proposed system implements the mapping of «input-output» in the form

$$\hat{x}_i(k) = \sum_{j=1}^m \varphi_{ij}^{[2]}(y_j(k)) = \sum_{j=1}^m \varphi_{ij}^{[2]} \left(\sum_{i=1}^n \varphi_{ji}^{[1]}(x_i(k)) \right),$$

$$\forall i = 1, 2, \dots, n$$

where $\varphi_{ji}^{[1]}(\bullet)$, $\varphi_{ij}^{[2]}(\bullet)$ are the nonlinear transformations, which are implemented by the multivariate nonlinear synapses of the autoencoder layers.

These transformations can be written in the form

$$y_j(k) = \sum_{i=1}^n \sum_{l=1}^h w_{ji}^{[1]} \mu_{li}^{[1]}(x_i(k)), \forall j = 1, 2, \dots, m,$$

$$\hat{x}_i(k) = \sum_{j=1}^m \sum_{l=1}^h w_{ij}^{[2]} \mu_{lj}^{[2]}(y_j(k)), \forall i = 1, 2, \dots, n$$

or finally

$$\hat{x}_i(k) = \sum_{j=1}^m \sum_{l=1}^h w_{ij}^{[2]} \mu_{lj}^{[2]} \left(\sum_{i=1}^n \sum_{l=1}^h w_{ji}^{[1]} \mu_{li}^{[1]}(x_i(k)) \right). \quad (1)$$

It should be noted the multidimensional nonlinear synapses in general case describe the Takagi-Sugeno-Kang (zero order) neuro-fuzzy system (Wang-Mendel system), i.e. has high approximation properties, and the transformation (1) describes the autoassociative variation of the neuro-fuzzy Kolmogorov's network, i.e. is the universal approximator.

In the simplest case as the membership function conventional triangular functions can be used, which have been used by the authors of the neo-fuzzy neuron [5-7] and can be written in the form

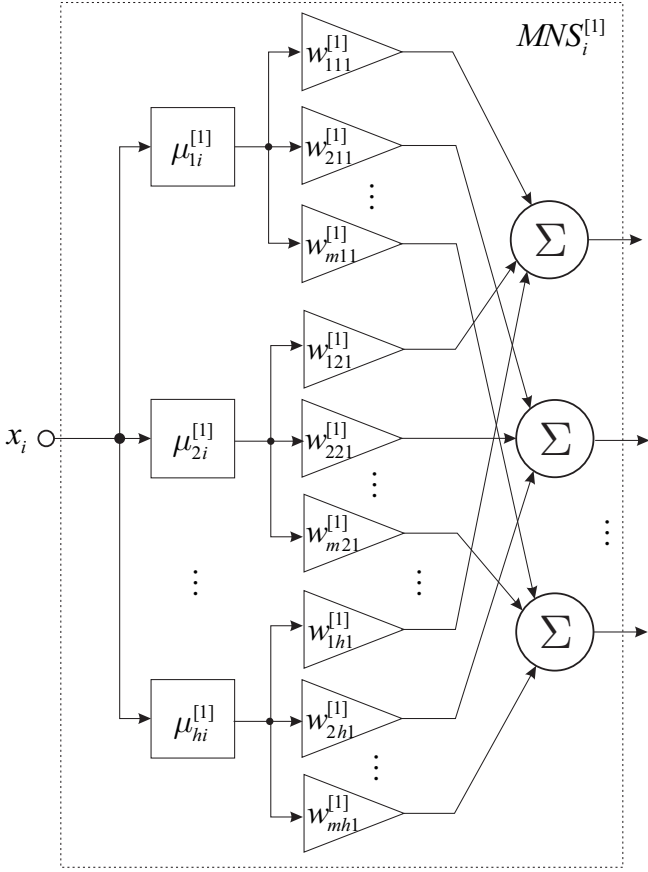


Fig. 2. Multidimensional nonlinear synapse of first layer

$$\mu_{li}^{[1]}(x_i) = \begin{cases} \frac{x_i - \bar{x}_{l-1,i}^{[1]}}{\bar{x}_{l,i}^{[1]} - \bar{x}_{l-1,i}^{[1]}}, & \text{if } x_i \in [\bar{x}_{l-1,i}^{[1]}, \bar{x}_{l,i}^{[1]}], \\ \frac{\bar{x}_{l+1,i}^{[1]} - x_i}{\bar{x}_{l+1,i}^{[1]} - \bar{x}_{l,i}^{[1]}}, & \text{if } x_i \in [\bar{x}_{l,i}^{[1]}, \bar{x}_{l+1,i}^{[1]}], \\ 0 & \text{otherwise} \end{cases} \quad \square \square \square$$

and

$$\mu_{lj}^{[2]}(y_j) = \begin{cases} \frac{y_j - \bar{y}_{l-1,j}^{[2]}}{\bar{y}_{l,j}^{[2]} - \bar{y}_{l-1,j}^{[2]}}, & \text{if } y_j \in [\bar{y}_{l-1,j}^{[2]}, \bar{y}_{l,j}^{[2]}], \\ \frac{\bar{y}_{l+1,j}^{[2]} - y_j}{\bar{y}_{l+1,j}^{[2]} - \bar{y}_{l,j}^{[2]}}, & \text{if } y_j \in [\bar{y}_{l,j}^{[2]}, \bar{y}_{l+1,j}^{[2]}], \\ 0 & \text{otherwise} \end{cases} \quad (3)$$

where $\bar{x}_{l,i}^{[1]}$, $\bar{y}_{l,j}^{[2]}$, $l=1,2,\dots,h$ are the centers of the activation functions, in simplest case they are equidistributed along the axes x_i , y_j .

The membership functions (2), (3) fulfil Ruspini conditions:

$$\begin{cases} \mu_{l-1,i}^{[1]}(x_i) + \mu_{l,i}^{[1]}(x_i) = 1, & \text{if } x_i \in [\bar{x}_{l-1,i}^{[1]}, \bar{x}_{l,i}^{[1]}], \\ \mu_{l,i}^{[1]}(x_i) + \mu_{l+1,i}^{[1]}(x_i) = 1, & \text{if } x_i \in [\bar{x}_{l,i}^{[1]}, \bar{x}_{l+1,i}^{[1]}], \\ \mu_{l-1,j}^{[2]}(y_j) + \mu_{l,j}^{[2]}(y_j) = 1, & \text{if } y_j \in [\bar{y}_{l-1,j}^{[2]}, \bar{y}_{l,j}^{[2]}], \\ \mu_{l,j}^{[2]}(y_j) + \mu_{l+1,j}^{[2]}(y_j) = 1, & \text{if } y_j \in [\bar{y}_{l,j}^{[2]}, \bar{y}_{l+1,j}^{[2]}]. \end{cases}$$

This fact significantly reduces the learning process, because in each instant of time k only two nearest-neighbor membership functions are fired and accordingly that not all synaptic weights are adjusted, but only $4nm$ ones of them.

III. THE LEARNING OF THE GNFN AUTOENCODER

As a rule, the supervised learning process of the hybrid systems of machine learning is reduced to tuning the synaptic weights set with the goal of minimizing the accepted (usually quadratic) learning criterion.

Introducing into consideration the vector of membership functions

$$\begin{aligned} \mu^{[1]}(x(k)) &= (\mu_{11}^{[1]}(x_1(k)), \mu_{21}^{[1]}(x_1(k)), \dots, \mu_{h1}^{[1]}(x_1(k)), \mu_{12}^{[1]}(x_2(k)), \\ &\dots, \mu_{li}^{[1]}(x_i(k)), \dots, \mu_{hn}^{[1]}(x_n(k)))^T, \quad \mu^{[2]}(y(k)) = (\mu_{11}^{[2]}(y_1(k)), \\ &\mu_{21}^{[2]}(y_1(k)), \dots, \mu_{n1}^{[2]}(y_1(k)), \mu_{12}^{[2]}(y_2(k)), \dots, \mu_{lj}^{[2]}(y_j(k)), \dots, \\ &\mu_{hm}^{[2]}(y_m(k)))^T \quad \text{with dimensions } (hn \times 1), (hm \times 1) \end{aligned}$$

respectively and synaptic weights matrices

$$W^{[1]} = \begin{pmatrix} W_{111}^{[1]} & W_{121}^{[1]} & \dots & W_{1hn}^{[1]} \\ W_{211}^{[1]} & W_{221}^{[1]} & \dots & W_{2hn}^{[1]} \\ \vdots & \vdots & \ddots & \vdots \\ W_{m11}^{[1]} & W_{m21}^{[1]} & \dots & W_{mhn}^{[1]} \end{pmatrix},$$

$$W^{[2]} = \begin{pmatrix} W_{111}^{[2]} & W_{121}^{[2]} & \dots & W_{1hm}^{[2]} \\ W_{211}^{[2]} & W_{221}^{[2]} & \dots & W_{2hm}^{[2]} \\ \vdots & \vdots & \ddots & \vdots \\ W_{n11}^{[2]} & W_{n21}^{[2]} & \dots & W_{nhm}^{[2]} \end{pmatrix}$$

of dimensions $(m \times hn)$, $(n \times hm)$ respectively. Then the mapping, which realized by GNFN^[1], can be written in the form

$$y(k) = W^{[1]} \mu^{[1]}(x(k))$$

and GNFN^[2]

$$\hat{x}(k) = W^{[2]} \mu^{[2]}(y(k)).$$

In general, the autoencoder realizes a mapping in the form

$$\hat{x}(k) = W^{[2]} \mu^{[2]}(W^{[1]} \mu^{[1]}(x(k))), \quad (4)$$

which is the generalization of the expression (1). During the learning process the matrix $W^{[1]}$, $W^{[2]}$ have to be obtain, which provide an optimal compression of the initial data set.

For tuning synaptic weights of GNFN^[2] we can introduce into consideration the error of the recovering i -th element of input signal $x_i(k)$ in the form

$$\begin{aligned} e_i(k) &= x_i(k) - \hat{x}_i(k) = \\ &= x_i(k) - \sum_{l=1}^h \sum_{j=1}^m w_{ij}^{[2]}(k-1) \mu_{lj}^{[2]}(y_j(k)) = \\ &= x_i(k) - w_i^{[2]}(k-1) \mu^{[2]}(y(k)) \end{aligned}$$

(here $w_i^{[2]}(k-1)$ is the i -th row of the weights matrix $W^{[2]}$) and standard learning criterion for i -th output

$$E_i(k) = \sum_k e_i^2(k) = \sum_k (x_i(k) - w_i^{[2]}(k-1) \mu^{[2]}(y(k)))^2.$$

The gradient procedure for minimization of the criterion $E_i(k)$ can be written in the general form

$$\begin{aligned} w_i^{[2]}(k) &= w_i^{[2]}(k-1) - \eta^{[2]}(k) \nabla_{w_i^{[2]}} E_i(k) = \\ &= w_i^{[2]}(k-1) - \eta^{[2]}(k) \nabla_{w_i^{[2]}} e_i^2(k) = \\ &= w_i^{[2]}(k-1) + \eta^{[2]}(k) e_i(k) \mu^{[2]T}(y(k)) = \\ &= w_i^{[2]}(k-1) + \\ &\quad + \eta^{[2]}(k) (x_i(k) - w_i^{[2]}(k-1) \mu^{[2]}(y(k))) \mu^{[2]T}(y(k)), \end{aligned} \quad (5)$$

where $\eta^{[2]}(k)$ is a learning rate coefficient of the output layer.

To increase learning process speed based on algorithm in the form (5) we can use either the standard recurrent least square method in the form

$$\begin{cases} w_i^{[2]}(k) = w_i^{[2]}(k-1) + \frac{e_i(k) \mu^{[2]T}(y(k)) P^{[2]}(k-1)}{1 + \mu^{[2]T}(y(k)) P^{[2]}(k-1) \mu^{[2]}(y(k))}, \\ P^{[2]}(k) = P^{[2]}(k-1) - \\ \quad - \frac{P^{[2]}(k-1) \mu^{[2]}(y(k)) \mu^{[2]T}(y(k)) P^{[2]}(k-1)}{1 + \mu^{[2]T}(y(k)) P^{[2]}(k-1) \mu^{[2]}(y(k-1))}, \end{cases}$$

or one-step optimal algorithm in the form [19]:

$$w_i^{[2]}(k) = w_i^{[2]}(k-1) + e_i(k) \mu^{[2]+}(y(k)) \quad (6)$$

(here $\mu^{[2]+}(y(k)) = \mu^{[2]T}(y(k)) \|\mu^{[2]}(y(k))\|^{-2}$) or the algorithm, which has both tracking and smoothing properties [20]:

$$\begin{cases} w_i^{[2]}(k) = w_i^{[2]}(k-1) + (r^2(k))^{-1} e_i(k) \mu^{[2]T}(y(k)), \\ r^{[2]}(k) = \alpha r^{[2]}(k-1) + \|\mu^{[2]}(y(k))\|^2 \end{cases} \quad (7)$$

where $0 \leq \alpha \leq 1$ - forgetting factor.

It can be seen that if $\alpha = 0$, then the algorithm (7) coincides with expression (6).

The tuning of synaptic weights matrix is performed based on the error backpropagation procedure [21]. At that, we can write similarly to (5)

$$\begin{aligned} w_j^{[1]}(k) &= w_j^{[1]}(k-1) - \eta^{[1]}(k) \nabla_{w_j^{[1]}} E_i(k) = \\ &= w_j^{[1]}(k-1) - \eta^{[1]}(k) \nabla_{w_j^{[1]}} e_i^2(k) \end{aligned}$$

or element-wise

$$\begin{aligned} w_{ji}^{[1]}(k) &= w_{ji}^{[1]}(k-1) - \eta^{[1]}(k) \frac{\partial e_i^2(k)}{\partial w_{ji}^{[1]}} = \\ &= w_{ji}^{[1]}(k-1) - \eta^{[1]}(k) \frac{\partial e_i^2(k)}{\partial \hat{x}_i(k)} \frac{\partial \hat{x}_i(k)}{\partial y_j(k)} \frac{\partial y_j(k)}{\partial w_{ji}^{[1]}} = \\ &= w_{ji}^{[1]}(k-1) + \eta^{[1]}(k) e_i(k) \mu_{li}^{[1]}(x_i(k)) \sum_{l=1}^h w_{lj}^{[2]}(k) \frac{\partial \mu_{lj}^{[2]}(y_j(k))}{\partial y_j}. \end{aligned}$$

In the case, if the membership functions' centers in the output layer are uniformly distributed in the line of X-axis we can write

$$\frac{\partial \mu_{lj}^{[2]}(y_j(k))}{\partial y_j} = \begin{cases} (\bar{y}_{l,j}^{[2]} - \bar{y}_{l-1,j}^{[2]})^{-1}, & \text{if } y_j(k) \in [\bar{y}_{l-1,j}^{[2]}, \bar{y}_{l,j}^{[2]}], \\ (\bar{y}_{l,j}^{[2]} - \bar{y}_{l+1,j}^{[2]})^{-1}, & \text{if } y_j(k) \in [\bar{y}_{l,j}^{[2]}, \bar{y}_{l+1,j}^{[2]}], \\ 0 & \text{otherwise} \end{cases}$$

or introducing the notations

$$(\bar{y}_{l,j}^{[2]} - \bar{y}_{l-1,j}^{[2]})^{-1} = \Delta \bar{y},$$

$$(\bar{y}_{l,j}^{[2]} - \bar{y}_{l+1,j}^{[2]})^{-1} = -\Delta \bar{y},$$

we can obtain the compacted expression

$$\frac{\partial \mu_{lj}^{[2]}(y_j(k))}{\partial y_j} = \begin{cases} \Delta \bar{y}, & \text{if } y_j(k) \in [\bar{y}_{l-1,j}^{[2]}, \bar{y}_{l,j}^{[2]}], \\ -\Delta \bar{y}, & \text{if } y_j(k) \in [\bar{y}_{l,j}^{[2]}, \bar{y}_{l+1,j}^{[2]}], \\ 0 & \text{otherwise.} \end{cases}$$

Further, introducing the new notations

$$\sum_{i=1}^h w_{ij}^{[2]}(k) \begin{cases} \Delta \bar{y}, & \text{if } y_j(k) \in [\bar{y}_{i-1,j}^{[2]}, \bar{y}_{i,j}^{[2]}], \\ -\Delta \bar{y}, & \text{if } y_j(k) \in [\bar{y}_{i,j}^{[2]}, \bar{y}_{i+1,j}^{[2]}], \\ 0 & \text{otherwise} \end{cases} = \tilde{w}_{ij}^{[2]}(k),$$

we can write the procedure for adjusting the synaptic weights of the first layer

$$w_{ji}^{[1]}(k) = w_{ji}^{[1]}(k-1) + \eta^{[1]}(k) e_i(k) \mu_{ii}^{[1]}(x_i(k)) \tilde{w}_{ij}^{[2]}(k).$$

Choice of the learning rate parameter can be provided like (7), at that

$$\eta^{[1]}(k) = (r^{[1]}(k))^{-1}; r^{[1]}(k) = \alpha r^{[1]}(k-1) + \|\mu^{[1]}(x(k))\|^2.$$

In distinction from autoencoders, which are described in [14-17] the proposed system contains less number of the membership functions (it reduces its computational implementation) and has a high speed of learning algorithm due to the optimized choice of learning rate parameters.

IV. EXPERIMENTS

The effectiveness of proposed approach has been performed using data sets from UCI Repository of machine learning databases [20]. We take three data set: Iris, Parkinsons, Wine. Iris data set consists of 150 observations with 4 attributes and 3 classes, Wine data set consists of 178 observations with 13 attributes and 3 classes, Parkinsons data set consists of 197 observations with 23 attributes and 3 classes. The obtained results based on proposed autoencoder have been compared with the results based on autoassociative autoencoder ‘‘Bottle Neck’’. The data dimension after compression was 3 components for simplicity of visualization. The results were averaged after 20 times simulation with a different start condition for learning algorithm.

TABLE I. RESULTS OF SIMULATION BASED ON PROPOSED AUTOENCODER

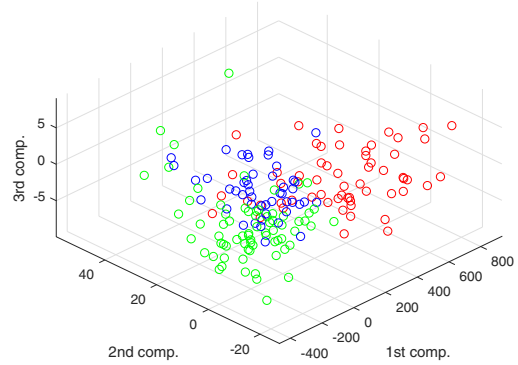
Autoencoders	Data Sets	Error	Learning time, sec.	
			Min	Max
Proposed autoencoder	Iris	0.19	2.31	3.91
	Wine	0.51	2.60	4.12
	Parkinsons	0.83	6.45	7.32
Autoassociative autoencoder ‘‘Bottle Neck’’	Iris	0.486	4.12	6.22
	Wine	0.903	6.44	8.26
	Parkinsons	0.593	9.21	10.98

It should be noticed, data, which are compressed based on proposed autoencoder, are more compact clusters than data, which are compressed using the autoassociative autoencoder ‘‘Bottle Neck’’.

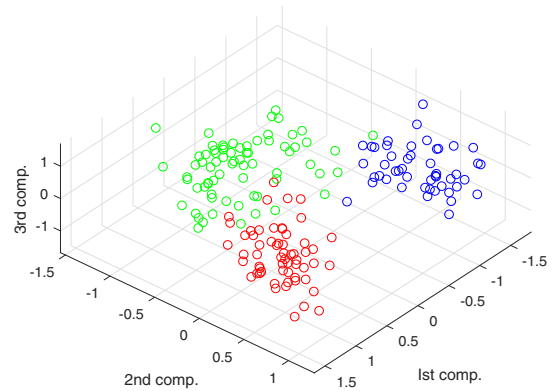
V. CONCLUSIONS

In this paper, the autoencoder based on the generalized neo-fuzzy neuron and its learning algorithm are proposed. Such system can be used as part of deep learning systems or as separated autoencoder for solving compression tasks in the machine learning problems. The proposed autoencoder is

co-called autoassociative ‘‘bottle-neck’’ system of computational intelligence but is characterized by high learning speed and less number of tuned parameters in comparison with well-known autoencoders, that allow using such system in Data Stream Mining. The efficiency of proposed approach has been justified based on different benchmark and real data set, obtained results have confirmed the advantages of the proposed autoencoder based on generalized neo-fuzzy neuron.



a)



b)

Fig. 3. Results of compression based on Wine data set using autoassociative autoencoder ‘‘Bottle Neck’’ (a) and proposed autoencoder (b)

REFERENCES

- [1] Y. LeCun, Y. Bengio, and G.E. Hinton, ‘‘Deep Learning’’. Nature, 2015, v. 521, pp. 436-444.
- [2] D. Graupe, Deep Learning Neural Networks: Design and Case Studies. World Scientific Publishing Company, 2016.
- [3] J. Schmidhuber, ‘‘Deep learning in neural networks: An overview.’’ Neural Networks, 2015, v. 61, pp. 85-117.
- [4] I. Goodfellow, Y. Bengio, and A. Courville, Deep learning. MIT Press, 2016.
- [5] T. Miki, and T. Yamakawa, ‘‘Analog implementation of neo-fuzzy neuron and its on-board learning.’’ In: Mastorakis NE (eds) Computational Intelligence and Application, Piraeus: WSES Press, 1999, pp. 144-149.
- [6] T. Yamakawa, E. Uchino, T. Miki, and H. Kusanagi, ‘‘A neo-fuzzy neuron and its applications to system identification and prediction of the system behavior.’’ In: Proceedings 2-nd International Conference on Fuzzy Logic and Neural Networks (IIZUKA-92), Iizuka, Japan, 17-22 July 1992, pp. 477-483.

- [7] E. Uchino, and T. Yamakawa, "Soft computing based signal prediction, restoration and filtering." In: Ruan D. (eds) *Intelligent Hybrid Systems: Fuzzy Logic, Neural Networks and Genetic Algorithms*, Boston: Kluwer Academic Publishers, 1997, pp. 331–349.
- [8] Ye. Bodyanskiy, I. Kokshenev, and V. Kolodyazhniy, "An adaptive learning algorithm for a neo-fuzzy neuron." In: *Proceedings of 3-rd International Conference of European Union Society for Fuzzy Logic and Technology (EUSFLAT)*, Zittau, Germany, 10-12 September 2003, pp. 375–379
- [9] V. Kolodyazhniy and Ye. Bodyanskiy, "Fuzzy Kolmogorov's Network," in *Lecture Notes in Computer Science*, vol. 3214, M.G. Negoita et al., Eds., Springer-Verlag, 2004, pp.764-771.
- [10] Ye. Bodyanskiy, V. Kolodyazhniy and P. Otto, "Neuro-fuzzy Kolmogorov's network for time-series prediction and pattern classification," in *Lecture Notes in Artificial Intelligence*, vol. 3698, U. Furbach, Ed., Heidelberg: Springer-Verlag, 2005, pp. 191-202.
- [11] V. Kolodyazhniy, Ye. Bodyanskiy and P. Otto, "Universal approximator employing neo-fuzzy neurons," in *Computational Intelligence Theory and Applications*, Ed. B. Reusch, Ed., Berlin-Heidelberg: Springer, 2005, pp. 631-640.
- [12] V. Kolodyazhniy, Ye. Bodyanskiy, V. Poyedyntseva, and A. Stephan "Neuro-fuzzy Kolmogorov's network with a modified perceptron learning rule for classification problems," in *Advances in Soft Computing*, vol. 38, B. Reusch, Ed., Berlin-Heidelberg: Springer-Verlag, 2006, pp. 41-49.
- [13] Ye. Bodyanskiy, Ye. Gorshkov, V. Kolodyazhniy, and V. Poyedyntseva "Neuro-fuzzy Kolmogorov's network," in *Lecture Notes in Computer Science*, vol.3697, W. Duch, J. Kacprzyk, E. Oja, and S. Zadrozny, Eds.,Berlin-Heidelberg: Springer-Verlag, 2005, pp.1-6.
- [14] V. Kolodyazhniy, F. Klawonn, and K. Tschumitschew, "A neuro-fuzzy model for dimensionality reduction and its application" *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems* vol. 15, is. 05, October 2007, pp. 571-593.
- [15] Ye. Bodyanskiy, I. Pliss, and O. Vynokurova, "On-line neuro-fuzzy big data autoencoder for deep neural networks and its rapid learning" In: *Proc. of XXX Int. Conference Problems of Decision making under uncertainties (PDMU 2017)* August 14-19, 2017, Vilnius, Lithuania, P. 20.
- [16] Ye. Bodyanskiy, I. Pliss, D. Peleshko, Yu. Rashkevych, and O. Vynokurova, "Hybrid Generalized Additive Wavelet-Neuro-Fuzzy-System and its Adaptive Learning". In: Eds. Zamojski, W., Mazurkiewicz, J., Sugier, J., Walkowiak, T., Kacprzyk, J., *Dependability Engineering and Complex Systems: Proceedings of the Eleventh International Conference on Dependability and Complex Systems DepCoS-RELCOMEX*. June 27-July 1, 2016, Brunow, Poland, 2016, pp. 51-61.
- [17] O. Vynokurova, Ye. Bodyanskiy, I. Pliss, D. Peleshko, and Yu. Rashkevych. "Neo-fuzzy encoder and its adaptive learning for Big Data processing." *Scientific Journal of RTU, Series "Computer Science" Volume "Information Technology and Management Science" 2017*, vol. 20, pp. 6–11.
- [18] R.P.Landim, B. Rodrigues, S.R. Silva, and W.M. Caminhas, "A neo-fuzzy-neuron with real time training applied to flux observer for an induction motor". In: *Proceedings of IEEE Vth Brazilian Symposium on Neural Networks*, Belo Horizonte, 9-11 Dec 1998, pp. 67-72.
- [19] Ye. Bodyanskiy, D. Peleshko, I. Pliss, and O. Vynokurova. *Hybrid adaptive systems of computational intelligence and their on-line learning in IT energy management tasks Green IT Engineering: Concepts, Models, Complex Systems Architectures*, Eds. By V. Kharchenko, Yu. P Kondratenko, J. Kacprzyk, Series: *Studies in Systems, Decision and Control*, Book 74, Publisher: Springer; 2017, pp 229-244.
- [20] UCI Repository of machine learning databases. CA: University of California, Department of Information and Computer Science. [Online]. Available: <http://www.ics.uci.edu/~mllearn/MLRepository.html>

Shape Evolutions of Poincaré Plots for Electromyograms in Data Acquisition Dynamics

Gennady P. Chuiko

Department of Computer Engineering
Petro Mohyla Black Sea National
University
Mykolayiv, Ukraine
genchuiko@gmail.com

Olga V. Dvornik

Department of Computer Engineering
Petro Mohyla Black Sea National
University
Mykolayiv, Ukraine
olga.v.dvornik@gmail.com

Yevhen S. Darnapuk

Department of Computer Engineering
Petro Mohyla Black Sea National
University
Mykolayiv, Ukraine
yevhen.darnapuk@gmail.com

Abstract— Poincaré plots (PPs) are a known way of study for complex time series. Such are the majority of medical signals. This method is in use here for the study of verified electromyograms (EMGs). EMGs are records of electrical action of muscular and nervous systems. The shapes of PPs for EMGs as well as its standard descriptors are sensitive to the diagnosis. These last describe the variability of the signals. We have studied the changes in the shapes of the PPs during the taping of EMGs. The changes of the standard descriptors were studied too. Three EMGs were considered for diverse diagnoses. They have varied duration but the same sampling rates. We have found the common shape of the PPs stabilizes itself during about the first third of each record. These shapes can change even further, but already remaining self-similar like the fractals. Standard descriptors are changing within the data acquisition. Still, these changes are smoother and less weighty in the last two thirds of each record.

Keywords— data acquisition dynamics; Poincaré Plots; variability; electromyograms; medical signals

I. INTRODUCTION

Electromyograms (EMGs) are records of electric action of muscles. This test is ensuring the high level of diagnosis of the nerves and muscles [1, 2]. It has arisen from the late 1970s by the efforts of American Academy of General Practice. Well-computerized processing is an inherent part in the modern electromyography (EMG) [3]. The database [4] has collected good examples of real EMGs.

Poincaré Plots (PPs) are a kind of return maps. Each result of measurement is plotted as a function of a next one. This simple and effective concept is in use to visualize many complex medical signals now [5]. Long records tapes become visible on a single chart. The longer the record, the more points appear on the chart. One main cloud (or a spot) of points arises as a rule. A shape of the cloud describes the evolution of the system. It allows visualizing the variability of a time series too [6].

There are standard numeric descriptors of PPs shapes, $SD1$, $SD2$ and $R = SD1/SD2$, which have been suggested in [7]. They describe two kinds of the variability of a time series and its randomness. The software [8] presents an example of modern data mining with PPs in cardiology. The mining of EMG data is only starting. Thus, there are now less cutting-edge positions.

The aim of this paper is the study of PPs shapes evolutions in the data record dynamics. We are going to trace the formation of these shapes within the EMGs record. This

will to touch standard descriptors as well. Thus, we are paying attention to the evolution of PPs during the EMG record and planning to link it with the real diagnosis.

II. DATA AND METHODS

The data was borrowed from PhysioNet portal [4]. These EMGs were in use earlier in [3, 5] but there were working with truncated datasets. Here we have used the full datasets. The frequency of the discretization was 4 kHz for all records. Hence, the sampling time interval was equal to 0.00025 s and the same for all records. The sizes of the signals were in mV.

Maple 18 was in use in computer handling of all datasets as well as for graphs plotting [9]. The import of the binary files from PhysioNet to Maple has been described in [10].

The standard descriptors were computed like in [11]:

$$\begin{aligned} SD1 &= \sqrt{2} \cdot SD \left(\frac{s_n - s_{n-1}}{2} \right); \\ SD2 &= \sqrt{2} \cdot SD \left(\frac{s_n + s_{n-1}}{2} \right); \\ &(n = 2, 3, \dots, N) \end{aligned} \quad (1)$$

Where $SD(s)$ denoted the operator of the standard deviation for the time series $\{s_n\}_{n=1..N}$.

First of them ($SD1$) defines the short-time variability of the time series. The second one ($SD2$) describes the long-time variability. Its ratio estimates the random impact in the data [5-7, 11].

Each of the datasets has been segmented on the 100 sections of equal duration (so called percentiles). Yet, these percentiles are varied length for different datasets. Besides, the duration of each percentile was much larger than the time sampling interval.

For instance, the shortest of the percentiles for the healthy patient had the duration 0.121715 s that means above 500 samples. It gives over 500 points in the common Poincaré Plot. The percentiles for patients with myopathy and neuropathy were larger about twice, or even threefold. Thus, one or several consequent percentiles may be reckoned as dynamic parts of the signal.

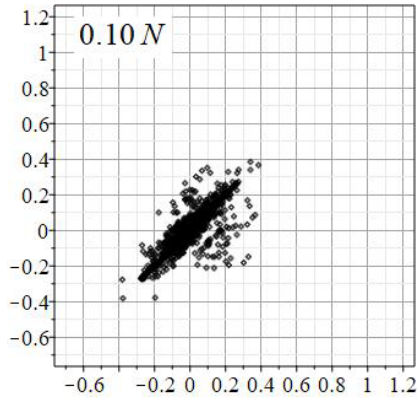
TABLE I. SORT INFORMATION ABOUT PATIENTS [4]

Sex	Age	Short diagnoses	Duration of records, s
male	44	A healthy patient	12.71500
male	57	Myopathy due to long history of polymyositis	27.56425
male	62	Chronic low back pain and neuropathy	36.96450

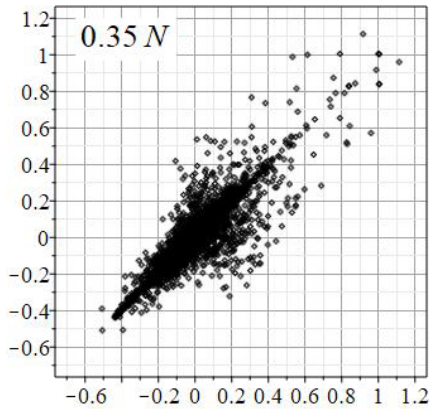
III. RESULTS AND DISCUSSION

A. Healthy Patient

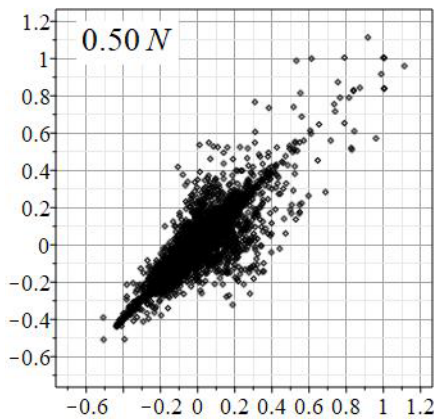
The full record consists of $N = 50860$ samples. Fig. 1 shows the dynamic of the shapes for PPs with varied numbers of points. The typical “comet shape” is observed.



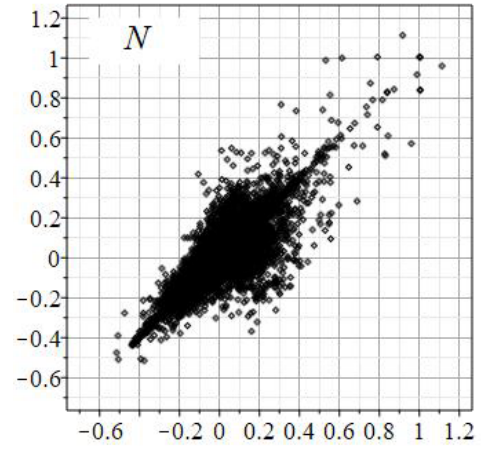
a)



b)



c)



d)

Fig. 1. The changes of Poincaré Plot shape in the process of data acquisition for healthy patient. The number of points is equal to: a) $0.1 \times N$; b) $0.35 \times N$; c) $0.5 \times N$; d) N

One can see the self-similarity of PPs, especially for three last images. Note the numbers of points on these images vary more than twice. The self-similarity was gone only for the PPs with small enough numbers of points (roughly less than $0.15 \times N$, see Fig. 1a).

Fig. 2 presents the dynamics of the standard descriptors for healthy patient.

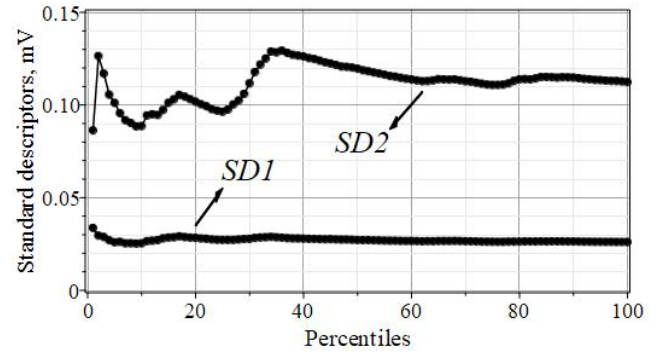


Fig. 2. The dynamics of the standard descriptors for healthy patient PPs

B. Patient with Myopathy

The record consists of $N = 110337$ samples. The self-similarity of the PPs was even clearer expressed, like to the above case, although the shape of PPs was definitely another. That is why we give here only the complete Poincaré Plot and the dynamics of the standard descriptors on the Fig. 3. The location of curves is alike to the Fig. 2.

C. Patient with Neuropathy

This record consists of $N = 147858$ samples. The self-similarity of PPs is inherent in this case too. However, the dynamics of the standard descriptors looks not as smooth as in above sections. The reader can see also specific shape of complete Poincaré Plot in this case (Fig. 4). The location of curves is alike to the Fig. 2 and Fig. 3.

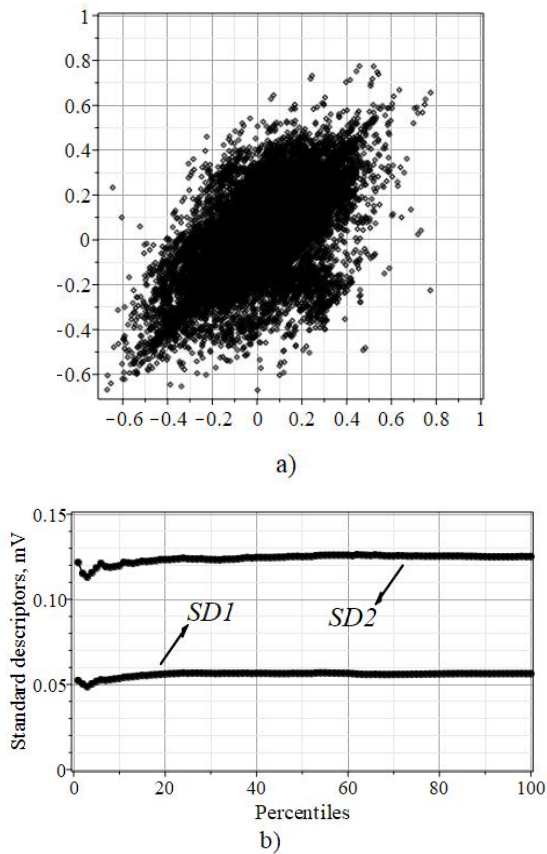


Fig. 3. The results for patient with myopathy: a) the shape of complete Poincaré plot; b) the dynamics of the standard descriptors.

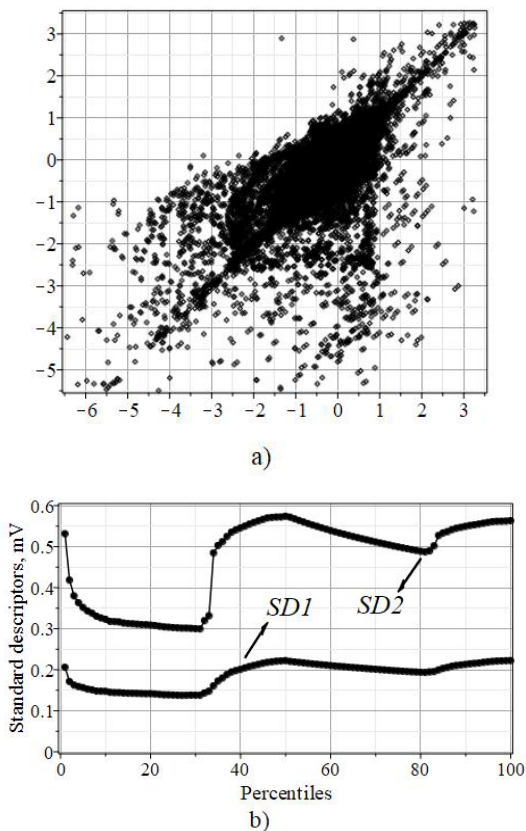


Fig. 4. The results for patient with neuropathy: a) the shape of complete Poincaré plot; b) the dynamics of the standard descriptors

D. Discussion and Comparisons

First, we point out that the shapes of PPs are most likely linked with the diagnosis (see Figures 1, 2 and 3). These specific shapes are very difficult to confuse. This idea needs further testing, yet, it may be useful for diagnosis and clinical decision-making.

We should to point out the similar dynamics of these shapes in the process of data acquisition (see Fig. 1). The shapes of PPs have stabilized after the gathering of about the third of data capacity. It looks as general rule, despite of the difference of shapes and the big difference in the numbers of samples for each data set.

The shape of the each PPs remains self-similar if data is collected further. This suggests the fractal nature of the PPs. Thus, a large enough part of the Poincaré Plot, let us say a half, or one third, is statistically equal to the complete Poincaré Plot [12, 13]. Here we say ‘a half, or one third’ keeping in the mind the number of points in a cloud.

This property permits the digital filtering of PPs with Haar wavelet filters for instance. Filtered PPs will be equal to unfiltered due to the own fractal nature. Haar filters can divide the data set on the two almost independent halves. One of them is the high frequency part of signal while the second one presents the low frequency part. Their scatter plot may be even more informative and convenient as the classic Poincaré Plot.

The behavior of the standard descriptors (se Fig. 2, 3 and 4), in general, confirms that above said. The more or less smooth dependences are inherent in the latter two third of signals. Still, the sizes of the descriptors and the smoothness of their changes with the data capacity are quite dependent on diagnoses.

Let us compare also the randomness of the datasets by the statistical box-plot (Fig. 5). Note that the random effects for the EMG of a healthy patient are clearly not as great as in pathological cases.

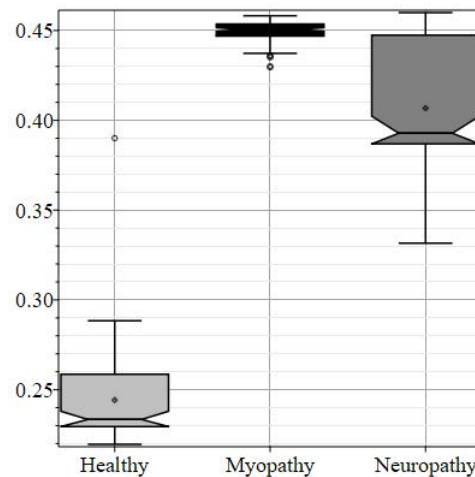


Fig. 5. Statistics box-plot for the ratios $R = SD1 / SD2$. These ratios estimate the randomness of each signal.

IV. CONCLUSIONS

Let summarize the results of this study in several points:

- Shapes of PPs for EMGs become steady after the first third of a data set. The diagnosis and the number of samples in the data did not have much weight for this rule.
- PPs have demonstrated self-similarity, behaved like typical fractals, in the process of further data gathering. Similar results have been reported earlier for the PPs of instantaneous heart rhythm [14].
- The dependences of standard descriptors of PPs on the number of samples have confirmed the above conclusions.

ACKNOWLEDGMENT

The authors are thankful to Associate Professor of Lviv Commercial Academy Dr. Markyan Girnyk for the help and the access to a licensed copy of Maple 18.

REFERENCES

- [1] K.R. Mills, "The basics of electromyography", *J. Neurol. Neurosurg. Psychiatry*, vol. 76, pp. ii32-ii35, 2005. DOI: 10.1136/jnnp.2005.069211.
- [2] M.B.I. Reaz, M.S. Hussain, F. Mohd-Yasin, "Techniques of EMG signal analysis: detection, processing, classification and applications", *Biol. Proced. Online*, vol. 8, pp. 11-35, 2006. DOI: 10.1251/bpo115.
- [3] G.P. Chuiko, I.A. Shyian, "Processing and analysis of electroneuromyograms with Maple tools", *Biomedical engineering and electronics*. [Online]. n. 10, pp. 1-8, 2015. Available: <http://biofbs.esrae.ru/pdf/2015/3/1006.pdf>.
- [4] Examples of Electromyograms. [Online]. Available: <https://www.physionet.org/physiobank/database/emgdb/>. DOI: 10.13026/C24S3D.
- [5] A. Kitlas-Golińska, "Poincaré Plots in Analysis of Selected Biomedical Signals", *Studies in Logic, Grammar and Rhetoric*, vol. 35, pp. 117-127, 2013. DOI: 10.2478/slgr-2013-0031.
- [6] R.A. Hoshi, C.M. Pastre, L.C.M. Vanderlei, M.F. Godoy, "Poincaré plot indexes of heart rate variability: Relationships with other nonlinear variables", *Auton. Neurosci. Basic Clin.*, vol. 177, pp. 271-274, 2013. DOI: 10.1016/j.autneu.2013.05.004.
- [7] M.P. Tulppo, T.H. Makikallio, T.E. Takala, T. Seppanen, H.V. Huikuri, "Quantitative beat-to-beat analysis of heart rate dynamics during exercise", *Am. J. Physiol.*, vol. 271, pp. H244-H252, 1996. DOI: 10.1152/ajpheart.1996.271.1.H244.
- [8] Poincaré Graph. BTL Cardiopoint-Poincaré graph. [Online]. Available: https://files.btlnet.com/product-document/9792e3d5-3dbf-45d8-9e84-5c964a6a8602/BTL-Cardiopoint_WP_Poincaré_graph_EN400_9792e3d5-3dbf-45d8-9e84-5c964a6a8602_original.pdf.
- [9] Review of New Features in Maple 18. [Online]. Available: <https://www.wolfram.com/mathematica/compare-mathematica/files/ReviewOfMaple18.pdf>.
- [10] G.P. Chuiko, D.A. Galyak, I.A. Shyian, "Interface elements of scientific Web-resource PhysioNet and import data to computer mathematics system Maple 17", *Medical Informatics and Engineering*, vol. 3, pp. 84-88, 2015. DOI: 10.11603/mie.1996-1960.2015.3.5008.
- [11] J. Piskorski, P. Guzik, "Filtering Poincaré plot", *Comput. Methods Sci. Technology*, vol. 11, pp. 39-48, 2005. DOI: 10.12921/cmst.2005.11.01.39-48.
- [12] B. Mandelbrot, "How Long Is the Coast of Britain? Statistical Self-Similarity and Fractional Dimension", *Science*, vol. 156, pp. 636-638, 1967. DOI: 10.1126/science.156.3775.636.
- [13] P. Bourke, "Box counting fractal dimension of volumetric data". [Online]. Available: <http://paulbourke.net/fractals/cubecount/>.
- [14] A.N. Kudinov, D.Y. Lebedev, V.N. Ryzzykov, V.P. Zvetkov, I.V. Zvetkov, A.P. Ivanov, "Self-similarity of the scatter plot of instantaneous heart rhythm", *Vestnik TvGU. Seriya: Prikladnaya matematika [Herald of Tver State University. Series: Applied Mathematics]*, n. 3, pp. 105-115, 2014. Available: <http://pmk-vestnik.tversu.ru/issues/2014-3/vestnik-pmk-2014-3-kudinov.pdf> (in Russian).

Game Model for Data Stream Clustering

Petro Kravets
Computer Science Department
Lviv Polytechnic National University
Lviv, Ukraine
Petro.O.Kravets@lpnu.ua

Abstract—In this article, the stochastic game model for data stream clustering is offered. Players represent numerical values of the clustering data. The essence of the game is that players perform a self-learning random move from one cluster to another in order to minimize the differences between the data of the same cluster. To solve the game, an adaptive recursive method has been developed. Computer modeling confirms the convergence of the game method with certain limitations of its parameters.

Keywords—data stream clustering; stochastic game model; adaptive game method.

I. INTRODUCTION

The clustering is a partition of the set of objects into subsets depending on their similarity. The separated subsets are called as clusters. Elements of one cluster have the general properties. Elements of different clusters considerably differ among themselves.

The clustering is used for problem-solving of the intellectual analysis and visualization of the data, grouping, and recognition of images, extraction of new knowledge and for information search. The clustering purpose consists in the finding of groups of similar objects in the set [1].

The clustering of objects also is used in chemistry, biology, medicine, sociology, pedagogics, psychology, philology, marketing, signal processing, pattern recognition, scientific discipline of documentation, computer science, scientific work and other areas of human activity for data structure in a cluster form for the purpose of their ordering and the group analysis.

The general clustering scheme is such: extraction of characteristics of objects; definition of the metric affinity of objects; partition of a set of objects on clusters; interpretation of clustering results.

Let each object $x \in X$ from a set of objects $X = (x_1, x_2, \dots, x_L)$ is described by a vector of properties $x = (x[1], x[2], \dots, x[k])$, which can be quantitative or qualitative characteristics of the object.

In problems of data stream clustering vectors of properties change in time: $x = (x_t | t=1, 2, \dots)$ [2, 3]. As a rule, the law of change of object properties is unknown a priori.

The similarity of two objects x_i also x_j is defined by the metrics of their affinity $\delta(x_i, x_j)$ in space of characteristics. As the metrics, the Euclidean distance, the Tchebyshev distance, the Manhattan distance, the percentage of

inconsistency, the Pierce correlation factor and others are used.

Generally, the clustering of objects it is possible to consider as a problem of optimum distribution of objects on groups. The minimization of a root-mean-square error of clusters setting can be the criterion of optimization:

$$\delta = \sum_{j=1}^N \sum_{i=1}^{C_j} \|x_i^{(j)} - \bar{x}_j\|^2 \rightarrow \min,$$

where $x_i^{(j)}$ is the point belonging to j -th cluster; \bar{x}_j is the center of the j -th cluster; C_j is a number of elements of the j -th cluster.

Substantial interpretation of the generated clusters for a finding of factors or the reasons of a grouping of objects in clusters is the final stage of clustering. For estimation of the quality of clustering, involve experts from corresponding subject domains.

The data intended for clustering, as a rule, contains uncertainty elements in practical applications. It can be indistinctly specified characteristics of the objects, missed attributes of objects in databases, noisy signals etc. In the uncertainty conditions apply fuzzy clustering, adaptive clustering, genetic algorithms, neural networks without the teacher learning.

The data clustering is formulated as a competitive or cooperative problem of assigning an object to one or another cluster. Problems of a competition and cooperation of objects are studied by the theory of games [4], and in uncertainty conditions, they are studied by the theory of stochastic games [5]. Therefore actual from the scientific, informative and practical points of view there are applications of methods of stochastic games for the data clustering in the conditions of uncertainty.

Construction of a game model of the data clustering with uncertainty elements is the goal of this paper. For purpose achievement it is necessary to solve such problems: to carry out a formulation of a problem game of the data clustering, to develop an adaptive game method and algorithm for solving the problem, to develop computer program model, to analyze and interpretation of the received results.

II. GAME PROBLEM STATEMENT

Let $X = \{x_1, x_2, \dots, x_L\}$ is the set by coordinates of points $x \in R^k$ in k -dimensional parametrical space. Coordinates of points define the normalized characteristic vector intended

for objects clustering. In this set, it is necessary to separate N clusters

$$\left\{ Y_n, n=1..N \left| \begin{array}{l} \bigcup_{n=1..N} Y_n = X, \\ Y_i \cap Y_j = \emptyset \quad \forall (i, j) \in \{1..N\} \end{array} \right. \right\}$$

by criteria

$$\frac{1}{C_n} \sum_{x \in Y_n} \|x_i - x_j\| \rightarrow \min, \quad n=1..N \quad (1)$$

where $C_n = |Y_n|$ is a quantity of elements which enter into a cluster Y_n ; $\|\cdot\| \in R^1$ is the Euclidean distance of a vector.

Let parameters of objects are vector random variables with stationary normal distribution:

$$x \sim Normal(m_x, d_x) \in R^k,$$

where m_x is an expectation value; d_x is a dispersion.

The separation of clusters Y_n ($n=1..N$) in the set X will be done using the stochastic game method described by the tuple $(I, A^i, \Xi^i | i \in I)$, where I is a set of players; $L = |I|$ is a quantity of players; $A^i = \{a^i[1], \dots, a^i[N]\}$ is a set of pure strategies of the i -th player which define a choice of one of clusters; N is a quantity of strategies of the i -th player ($N < L$); $\Xi^i: A \rightarrow R^1$ is a lose function of the i -th player; $A = \times_{i \in I} A^i$ is a set of the combined strategies.

The game essence consists in the random moving of players from one cluster to another. For this purpose during time moments $t=1, 2, \dots$, each player on the basis of the generator of random events independently of others chooses a pure strategy $a^i \in A^i$ which defines its accessory to the corresponding cluster. According to (1), after the realization of the combined variant $a \in A$, players receive random losses $\xi^i(a)$ with a priori unknown stochastic characteristics:

$$\xi_t^i = \frac{1}{C_t^i} \sum_{j \in I} \chi(a_i^i = a_i^j) \|x_i - x_j\| \quad \forall i \in I, \quad (2)$$

where $C_t^i = \sum_{j \in I} \chi(a_i^i = a_i^j)$ is a current quantity of elements of a cluster which contains the i -th player; $\chi(*) \in \{0, 1\}$ is an indicator of the event.

The efficiency of a game course is defined by functions of average losses:

$$\Xi_t^i = \frac{1}{t} \sum_{\tau=1}^t \xi_\tau^i \quad \forall i \in I. \quad (3)$$

The game purpose consists in minimization of the system of functions of average losses (3) in time:

$$\overline{\lim}_{t \rightarrow \infty} \Xi_t^i \rightarrow \min \quad \forall i \in I. \quad (4)$$

So, on the basis of a supervision of current losses $\{\xi_n^i\}$ each player $i \in I$ should learn to choose pure strategy $\{a_i^i\}$ so that with time course $t=1, 2, \dots$ to provide the performance of criteria system (4).

The game problem solutions will satisfy one of the conditions of collective balance, for example, on Nash or Pareto, depending on a method of formation of a sequence of strategies $\{a_i^i\} \forall i \in I$.

III. METHOD OF PROBLEM SOLVING

Stochastic game solving we will execute by means of adaptive recurrent transformation of vectors $p_i^i \forall i \in I$ of the mixed strategies.

Construction of a method of stochastic game solving we will carry out on the basis of stochastic approximation of a complementary slackness condition of a determined game, correct for the mixed strategies in a balance point on Nash [5].

For this purpose, we will define a polylinear function of average losses for the determined game:

$$V^i(p) = \sum_{a \in A} v^i(a) \prod_{j \in I, a^j \in a} p^j(a^j),$$

where $v(a) = M\{\xi_t^i(a)\}$.

Then the vector of a complementary slackness condition (CS) will be of the form:

$$\vec{CS}^i = \nabla_{p^i} V^i(p) - e^N V^i(p) = 0 \quad \forall i \in I,$$

where $\nabla_{p^i} V^i(p)$ is a gradient of the polylinear function of average losses; $e^N = (1_j | j=1..N)$ is a vector whose all components are equal to 1; $p \in S^M$ is the combined mixed strategy of players set on a convex unit simplex S^M ($M = N^L$).

To take account of the solutions in vertices of the unit simplex we will execute weighing of a CS^i -vector by elements of a vector p^i of the mixed strategies:

$$diag(p^i)(\vec{CS}^i) = 0 \quad \forall i \in I, \quad (5)$$

where $diag(p^i)$ it is the square diagonal matrix of an order N constructed of elements of a vector p^i .

Considering that

$$\begin{aligned} & \text{diag}(p^i)[\nabla_{p^i} V^i - e^N V^i] = \\ & = E\{\xi_t^i [e(a_t^i) - p_t^i] | p_t^i = p^i\} \end{aligned}$$

on the basis of a method of stochastic approximation [6] we will receive recurrent expression:

$$p_{t+1}^i = \pi_{\varepsilon_{t+1}}^N \{p_t^i - \gamma_t \xi_t^i (e(a_t^i) - p_t^i)\} \quad \forall i \in I, \quad (6)$$

where E is an expectation symbol; $\pi_{\varepsilon_{t+1}}^N$ is a projector on N -dimensional ε_t -simplex $S_{\varepsilon_{t+1}}^N$ [5]; $\gamma_t > 0$, and $\varepsilon_t > 0$ are monotonously descending sequences of positive values; $e(a_t^i)$ is the unit vector specifying in a choice of pure strategy $a_t^i = a^i \in A^i$.

Parameters γ_t and ε_t can be calculated as follows:

$$\gamma_t = \gamma t^{-\alpha}, \quad \varepsilon_t = \varepsilon t^{-\beta}, \quad (7)$$

where $\gamma > 0$; $\alpha > 0$; $\varepsilon > 0$; $\beta > 0$.

Convergence of strategies (6) to optimum values with probability 1 and in the root-mean-square is defined by the ratio of parameters γ_t and ε_t which should satisfy fundamental conditions of stochastic approximation [6].

Projection on expanded an ε_t -simplex $S_{\varepsilon_{t+1}}^N$ provides the performance of the condition $p_t^i[j] \geq \varepsilon_t$, $j = 1..N$ necessary for completeness of the statistical information on chosen pure strategies, and the parameter $\varepsilon_t \rightarrow 0$ is used as an additional element for controlling the convergence of the recurrent method.

The choice of pure strategy $a_t^i[k] \quad \forall i \in I$ is carried out by players on the basis of dynamic random distributions (6):

$$k = \arg \left(\min_{k=1..N} \sum_{j=1}^k p_t^i(a^i(j)) > \omega \right) \in \{1..N\}, \quad (8)$$

where $\omega \in [0, 1]$ it is the real random number with the uniform distribution law.

The stochastic game begins from not learned vectors of the mixed strategies with a value of elements $p_0^i(j) = 1/N$, where $j = 1..N$. During following moments of time the dynamics of vectors of the mixed strategies are defined by a Markovian recurrent method (6) – (8).

So, at the moment of time t each player on the basis of the mixed strategy p_t^i chooses a pure strategy a_t^i and until the moment of time $t+1$ receives current loss ξ_t^i then calculates the mixed strategy p_{t+1}^i according to (6).

Thanks to the dynamic reorganization of the mixed strategies based on the processing of current losses, the method (6) – (8) provide an adaptive choice of pure strategies in time.

Quality of game of the data clustering is estimated by:

1) the average loss function:

$$\Xi_t = \frac{1}{L} \sum_{i=1}^L \Xi_t^i, \quad (9)$$

where $L = |I|$ is a cardinality of a set of players;

2) the function of the average norm of mixed player strategies:

$$\Delta_t = \frac{1}{tL} \sum_{\tau=1}^t \sum_{i=1}^L \|p_\tau^i\|. \quad (10)$$

The algorithm of Stochastic Game Solving

1. To set initial values of parameters: $t = 0$ is an initial moment of time; N is a quantity of pure strategies of players (otherwise it is a number of clusters Y_n , $n = 1..N$); $L = |I|$ is a quantity of players; $X = \{x_1, x_2, \dots, x_L\}$ is a set of objects intended for clustering; k is a quantity of characteristic factors of objects $x \in R^k$; $m_x = (m_x[1], m_x[2], \dots, m_x[k])$ is an expectation value of parameters of object $x \in X$; $d_x = (d_x[1], d_x[2], \dots, d_x[k])$ is a dispersion of parameters of object $x \in X$; $A^i = \{a^i[1], a^i[2], \dots, a^i[N]\}$, $a^i(j) = j$, $i = 1..L$, $j = 1..N$ is a vectors of pure strategies of players; $p_0^i = (1/N, \dots, 1/N)$, $i = 1..L$ is an initial mixed strategies of players; $\gamma > 0$ is a parameter of a step of learning; $\alpha \in (0, 1]$ is an order of a step of learning; ε is an ε -simplex parameter; $\beta > 0$ is an order of an ε -simplex expansion rate; t_{\max} is a maximum quantity of steps of a method.

2. To choice variants of actions $a_t^i \in A^i$ of players $i = 1..L$ according to (8).

3. Get current property values of objects as random variables with the normal distribution law:

$$x_t = m_x + \sqrt{d_x} \left(\sum_{j=1}^{12} \omega_{j,t} - 6 \right),$$

where $\omega_{j,t} \in [0, 1]$ it is the real random number with the uniform distribution law.

4. To calculate the value of current losses ξ_t^i , $i = 1..L$ according to (2).

5. To calculate the value of parameters γ_t and ε_t , according to (7).

6. To calculate elements of vectors of the mixed strategies p_t^i , $i = 1..L$ according to (6).

7. To calculate quality characteristics Ξ_t (9) and Δ_t (10) of the data clustering.

8. To set the following moment of time $t := t + 1$.

9. If $t < t_{\max}$ then go to a step 2, else to stop.

IV. RESULTS OF COMPUTER MODELLING

We will solve a stochastic game by means of a recurrent method (6) – (8) for test parameters: $k=2$, $N=2$, $A^i = \{1,2\}$, $\gamma=1$, $\varepsilon=0.999/N$, $\alpha=0.3$, $\beta=2$, $t_{\max}=10^5$, $d=0.01$.

Let in the base set $X = \{Y_1, Y_2\}$ two non-empty subsets $Y_1 \cap Y_2 = \emptyset$ are visualized such that intracluster distances are less than intercluster distances. Elements of these subsets are received as the random points generated on a plane on the normal distribution law for different mathematical expectations.

On Fig. 1 graphs of functions Ξ_t of average losses of players and average norm Δ_t of the mixed strategies which characterize the convergence of stochastic game of data clustering are represented in logarithmic scale.

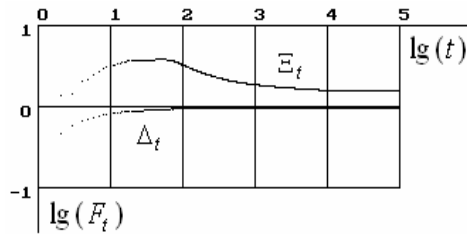


Fig. 1. Characteristics of solving the stochastic game in pure strategies

The game method (6) – (8) provides minimization of the function Δ_t of average losses in time. The function of the average norm of mixed strategies reaches the logarithmic zero, which illustrates the obtaining of the game's solving in pure strategies.

Dependence of average quantity of game learning steps \bar{t} on the parameter α is shown on Fig. 2. Value \bar{t} is averaged on realizations of random processes.

The moment of a game stop is defined by a condition of the approach of the average norm of mixed strategies to 1 ($\Delta_t \geq 0.99$) and correct assignment of elements of the set X to one of the clusters Y_1 or Y_2 (how these clusters are visualized in the set X).

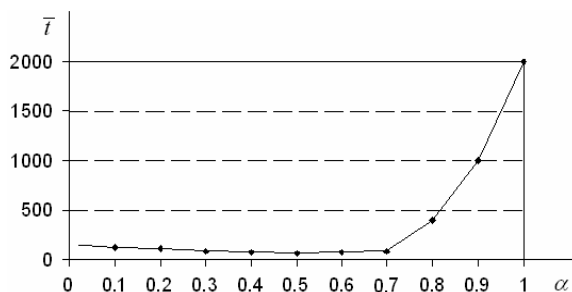


Fig. 2. Influence of the parameter α on the game convergence

For a solved problem, the growth of parameter α from 0 to 0.7 does not lead to considerable deterioration of

stochastic game convergence. Considerable growth of the average quantity of game steps occurs at $\alpha > 0.7$.

The order of convergence rate of a game method is defined by a parity of parameters α and β . For the convergence of the offered method, it is necessary that these parameters satisfy the conditions of stochastic approximation [6]. Dependence of average quantity of steps \bar{t} of clustering game from a dispersion d_x of parameters of objects $x \in X$ it is representing by the diagram on Fig. 3.

Value of a dispersion $d_x \in [0;50]$ does not a material effect on the quantity of the steps necessary for the data clustering by means of a game method (6) – (8). For values $d_x > 50$ of a dispersion, considerable growth of the average quantity of game steps necessary for correct adding of elements of the set X to one of the clusters Y_1 or Y_2 at the level $\Delta_t \geq 0.99$ of game learning is observed.

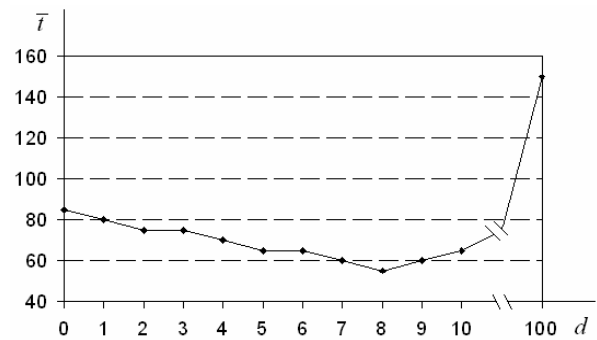


Fig. 3. Influence of the dispersion on the game convergence

The boundary elements of the subsets can be assigned to both cluster Y_1 and cluster Y_2 , that is, clusters can intersect.

Let in the general set X there are the points $y \in Y$ placed on equally spaced from subsets $Y_1 - Y$ and, $Y_2 - Y$ that is, $|s(y, Y_1 - Y) - s(y, Y_2 - Y)| < \varepsilon$, where $s(y, Z) = \min_{z \in Z} \|y - z\|$. Then the method (6) – (8) provides solving the game in mixed strategies as shown on Fig. 4.

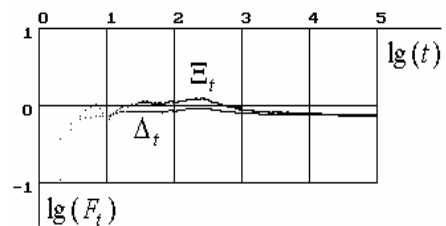


Fig. 4. Characteristics of solving the stochastic game in mixed strategies

On Fig. 4 it is visible that the function Δ_t of the average norm does not reach the logarithmic zero indicating that the game is solved in mixed strategies.

The growth of cardinality of a set X and corresponding growth of the number of players leads to a reduction of convergence rate of the stochastic game, which appears in the growth of the quantity of the steps necessary for the data clustering.

On Fig. 5 the graph of the dependence of average quantity of steps of stochastic game learning on the number of clustering objects is represented. The data intended for clustering is received randomly by means of the normal distribution law of coordinates of points on a plane. It is generated two concentrations of points with parameters of the normal distribution. The moment \bar{t} of the game termination is defined by a condition $\Delta_t \geq 0.99$. The obtained results are averaged on $k_{\text{exp}} = 100$ experiments.

By results of experiments, it is visible that with an increase in the quantity of clustering objects the quantity of the steps necessary for stochastic game learning increases.

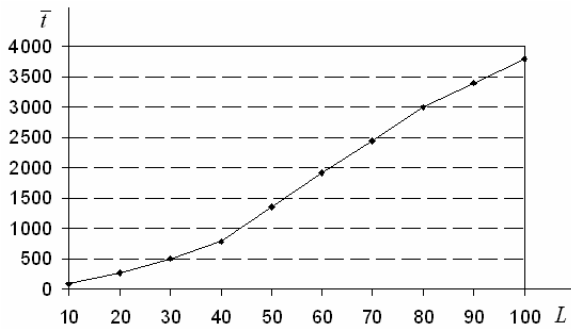


Fig. 5. Dependence of the average quantity of game steps on the number of clustering points

Achievement of the characteristics of the stochastic game convergence, which is acceptable in practice, is determined by fine-tuning of the parameters of the game method within the framework of the basic relations given by the theory of stochastic approximation.

V. CONCLUSIONS

In this paper, the new game model for data stream clustering is proposed. An adaptive recursive method was

constructed to solve the game. Random moving of points on a plane simulates data streams.

Convergence of a game method depends on the dimension of the stochastic game, the intensity of the noise and the parity of parameters of the game method.

The efficiency of the game of data clustering decreases at an increase of the number of players and noise intensity.

Simulation veracity proves repeatability of values of average characteristics of the stochastic game obtained for various realizations of random variables.

The offered game method of the data clustering belongs to a class of methods, which are based on the processing of reactions of the environment. This method has a relatively small (power-law) order of convergence rate due to the a priori uncertainty of the system.

This limitation can be overcome by the high performance of modern computer and possibility a game problem parallelization.

REFERENCES

- [1] A. Jain, M. Murty, and P. Flynn, "Data Clustering: A Review", *ACM Computing Surveys*, vol 31, no. 3, pp. 264-323, September 1999.
- [2] D. Barbara, "Requirements for clustering data streams", *ACM SIGKDD Explorations Newsletter*, vol. 3, №. 2, pp. 23-27, 2003.
- [3] J. Chandrika, and K.R. Ananda Kumar, "Dynamic Clustering Of High-Speed Data Streams", *International Journal of Computer Science Issues*, vol. 9, iss. 2, №. 1, pp. 224-228, 2012.
- [4] T. Roughgarden, E. Tardos and V. V. Vazirani. *Algorithmic Game Theory*, edited by Noam Nisan, Cambridge University Press, 2007.
- [5] A. Nazin, and A. Poznyak, *Adaptive Choice of Variants*, Moscow, Nauka, 1986 (in Russian).
- [6] H. J. Kushner, G. George Yin, *Stochastic Approximation and Recursive Algorithms and Applications*. New York: Springer Verlag, 2003.

Defining Author's Style for Plagiarism Detection in Academic Environment

Victoria Vysotska
Information Systems and Network Department
Lviv Polytechnic National University
Lviv, Ukraine
Victoria.A.Vysotska@lpnu.ua

Vasyl Lytvyn
Information Systems and Network Department
Lviv Polytechnic National University
Lviv, Ukraine
Vasyl.V.Lytvyn@lpnu.ua

Yevhen Burov
Information Systems and Network Department
Lviv Polytechnic National University
Lviv, Ukraine
Yevhen.V.Burov@lpnu.ua

Andriy Demchuk
Information Systems and Network Department
Lviv Polytechnic National University
Lviv, Ukraine
Andrii.B.Demchuk@lpnu.ua

Abstract—The usage of linguometry and stylometry technologies for author's style detection is discussed. The statistical linguistic analysis of author's text uses the advantages of content monitoring based on NLP methods for stop words definition. They are used in stylometry for definition of attribution degree of analyzed text to the specific author. The formal approach for Ukrainian language text author's style definition is proposed.

Keywords—*plagiarism, plagiarism detection, author's style, statistic linguistic analysis, quantitative linguistics, authorship attribution.*

I. INTRODUCTION

The strategic task of scientific and educational organizations is the formation of scientific elite, able to contribute to economic growth by introducing innovative products in industry, agriculture, medical services and information technologies. They would be able to support the sustainable and continuous growth of Ukrainian economy. The effectiveness of education and science is a constant concern for government and public organizations. Plagiarism negatively affects the quality of education and science [1]. Academic integrity is a rather broad term. It is understood as a set of ethical principles and rules defined by laws which should be followed by all participants of educational process in study, scientific research with the ultimate purpose of creating trust in objectivity of educational or scientific accomplishments. Attaining academic integrity requires a systemic approach combining the implementation of various organizational, educational, and technical measures. Among important technical solutions in this area is the creation of repository of academic texts of organization and National repository of academic texts (NRAT); and using information system (IS) for plagiarism detection in scientific articles which detects similarities based on data from NRAT and Internet. In article [2] plagiarism is subdivided in four kinds, each of them having the distinct purpose. Depending of activity type and application area, plagiarism can be:

- professional (acquisition of intellectual, creative and professional achievements of other people with a professional purpose;

- educational and scientific (appropriation of intellectual property uniquely in process of attaining a scientific degree, or qualification certification, or confirming existing degree);
- social (similar to professional, but not occurring in professional environment);
- normative (appropriation of results of methodical, scientific work, laws and bills developed by other people, also practical solutions).

Depending on form, we establish following types of plagiarism:

- Full or partial copying of author's work using linguistic, lexical and technological interpretation (implemented by most of existing plagiarism detectors).
- Appropriation of the main idea (hard to detect, because there are no methods in existing systems allowing to extract and analyze meaning of document).
- Plagiarism implying the use of references, while citing other works, referencing non-existing or non-relevant sources, presenting facts from other research without referencing it (partially implemented in some plagiarism detecting systems);
- When plagiarized work is not created by putative author himself, but bought from certain organization or person producing such works commercially and attributed to buyer. The detection of such type of plagiarism requires the definition of original writer's style using as a sample his other works. This important problem is addressed in this article.

II. BACKGROUND ANALYSIS

For automatic detection of language formatted excerpts of text are analyzed: letters are ordered in diminishing order according to frequencies of their appearance in excerpt (frequencies are shown), capitals and small letters are not distinguished. We can analyze the data and detect the

language of text using one of three methods presented in [3-9]:

- Frequencies of vowels and consonants in text,
- Frequencies of voiced, voiceless, nasal and other types of letters and their combinations,
- Frequencies of letter's usage.

In order to research peculiarities of author's style various individualized lexical metrics were defined. Among them are continuity, lexical variety, syntactic complexity, indices of concentration and uniqueness for fragments from author's text and other text. Next, using those metrics, internal dynamics of text is analyzed and finally, the degree of appropriation of analyzed text to a specific author is defined [10]. In order to define the degree of plagiarism a summary grouping table is built. In this table are placed, calculated previously for sets of texts with similar content, the average group values of continuity, lexical variety, syntactic complexity, indices of concentration and uniqueness [11].

The area of standard deviations is defined and thus the lexical similarity of analyzed fragments is compared to benchmark [12]. The author of text is defined based on analysis of his formatted text [13]. Words are placed in falling order according to their usage frequency in text fragment. The type of speech (direct or indirect) is noted. Common names are removed from text fragment. Author of text fragment or of a whole text is defined when possible using frequency dictionaries [14]. The analysis of authorship is based of definition of differences between individual author styles [15]. The uniqueness of style make author's language dynamic, memorable, easier to understand. However it is important to differentiate between individual style characteristics and characteristics common to authors [16]. The degree of authorship attribution of particular work includes credibility, authenticity of such work characteristics as author, time and place of creation. This degree is defined based on analysis of stylistic and technical distinctions [17].

III. STATEMENT OF PROBLEM

Our work aims to detect plagiarism in scientific articles by analysis of author style on a deeper level. We are treating this task as a text classification problem (fig. 1), where every document belongs only to a specific category/class/author. As initial dataset we use one-author articles of 100 Ukrainian scientists, 10 articles from each.

Thus, we analyze the collection of 1000 texts. For each category benchmarks which correspond to the style are defined. For each of researched text the coefficient of attribution to specific categories is calculated. For author's style definition the methods of machine learning were used (Bayesian classifier and reference vector method). Using our dataset we also obtained such quantitative characteristics of author's style as coefficient of text variability, degree of syntactic complexity, continuity, and uniqueness and concentration indices [3-5].

The stage of text preprocessing is rather time consuming and consists of primary processing of linguistic data (building of distribution sequences, calculating statistics and other linguometric characteristics), lexicographical processing of textual data (creating frequential and alphabetically frequential dictionaries for texts of specific

author, dictionaries of concordances, referential words, and author's style keywords) and applying methods of stylometry for authorship attribution [18-24].

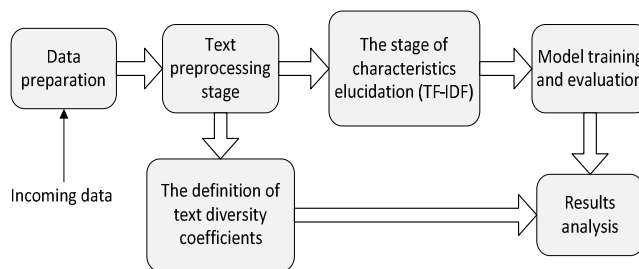


Fig. 1. The general scheme of authorship attribution workflow

It is practically impossible to do this without preceding analysis of author's works represented in form of benchmark. The linguistic analysis and interpretation of stylistic properties and regularities of author's style is done using methods of content analysis and has following steps: the correct selection of texts; lemmatization of text units; removing heterogeneity of text units; building dictionaries and based on them creating statistical distributions within corresponding frequential linguistic dictionary scales; search of parameters which adequately represent the structure of frequential dictionary; reviewing parameters for effectiveness; mathematical modeling and analysis of lexical and statistical distributions; creating statistical classifications; results interpretation and definition of benchmarked coefficients of text diversity for the specific author. An important part of this process is the definition of correct keyword dictionaries of author's style.

We developed a system allowing to select language(s) used in text. The access to the process of keywords definition based of thematic words stems is provided by Victana (victana.lviv.ua) resource [25]. When building such dictionaries from Ukrainian texts the word stems should be taken without inflections.

IV. THE METHOD OF AUTHOR'S STYLE DEFINITION FOR TEXTUAL CONTENT

The linguistic and statistical basis for research of text attribution is based on works [3, 18-23]:

1) primary processing of linguistic data (building distribution sequences, calculating statistics, statistical metrics and other linguometric parameters),

2) lexicographic processing of textual data (creation of frequential and alphabetically frequential dictionaries, dictionaries of concordances, referential words, and author's style keywords).

The usage of linguometric methods for statistical description of text supports research in the domain of authorship [24]. The method of linguistic analysis and interpretation of stylistic features and regularities of writer's style (or style of specific historical period) uses algorithm 1.

Algorithm 1. The linguistic analysis and interpretation of stylistic features and regularities of specific author's style.

Stage 1. Texts selection. It is important how texts are selected and the size of selection: for reliable definition of characteristics the required selection size is at least 18 thousand words [26-31].

Stage 2. The lemmatization of textual units. Unification of different words according to lemmas of language [5].

Stage 3. Removing disparity of textual units. Resolving the problem of text units' disparity, such as the usage of direct and indirect speeches.

Stage 4. Building a system of frequential dictionaries and, based on it, creation of the statistical distributions within corresponding frequency scales. Frequential dictionary is a type of dictionary where the usage frequency of specific language object is noted (word composition, word form, collocations, idioms, phrases) in different texts of specific size. Usually the absolute and relative frequency of language units usage is shown, in fall down order [3].

Stage 5. Search for parameters which adequately reflect the structure of frequential dictionary. The number of parameters is different. For example, for description of French texts from XVII century 51 parameters were proposed [26-31]. The common parameters found in [27-30] allow to formulate several linguo-statistical methods for text research:

- the method of supporting words (calculation of usage frequency and percent of auxiliary words [18-22]: such as pronouns, prepositions, conjunctions and determiners;
- the method of punctuation marks (only calculate the number of internal and external punctuation marks);
- the method of words (only calculate the number of words of specified length);
- the method of sentences (only calculate the number of sentences of specific length);
- the syntactic method (calculate the number of punctuation marks, words and sentences of specified length);
- the combined method (the combination of supporting words method and syntactic method).

Stage 6. Checking the effectiveness of parameters. Applying known methods for selected parameters effectiveness verification.

Stage 7. Mathematical modeling of lexico-statistical distributions. Applying known methods of mathematical modeling of lexico-statistical distribution.

Stage 8. Building of statistical classifications (author's benchmarks) reflecting the stylistic regularities within works of specified author or specific literary epoch (or the sequence of literary epochs).

Stage 9. The interpretation of obtained results from the perspective of historical and linguistic knowledge, general and historical stylistics.

Using algorithm 1 we perform the task of authorship attribution which can be described as follows. Suppose, that there is a benchmark statistically reflecting all author's works. We should evaluate the belonging of specific texts to benchmark using corresponding methods. Let's consider as example the works of Author 1 and her publications from [32-40]. Assume that author's benchmark is already built, that tasks of selecting texts, lemmatization and non-uniformness problems are already resolved and frequential dictionary is created [3, 41-48]. We will use the method of supporting words for attribution and will represent results as

correlation coefficients and also graphically. Separately, we will also note the evolution of relevance of one of text's parameters – auxiliary words- for authorship attribution: pronouns, conjunctions, prepositions [49-64].

For individual writer's style definition the auxiliary words are significant because they are not influenced by book theme or content [3]. We will assume that this parameter is effective for authorship attribution and will use a list of 71 auxiliary stop-words [41-48, 55-69].

V. THE RESULTS OF EXPERIMENTS

Let's take four arbitrary fragments from [3-5, 25], formatted depending on attribution method: from each fragment only prepositions, conjunctions and pronouns are used. The overall number of word usages is calculated and common names are excluded. For each of fragments we define the absolute frequency (AF) and relative frequency (RF) of auxiliary word appearance and also the relative frequency of this word appearance in benchmark. Figure 2 shows the graphical representation of relative frequency of stop words appearance in Fragment 1 and in benchmark. The coefficient of correlation for auxiliary words in this case is $R_{e-V1}=0,6076$.

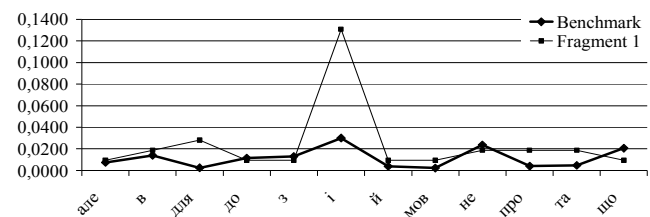


Fig. 2. Relative frequency of auxiliary words appearance in fragment 1 and benchmark

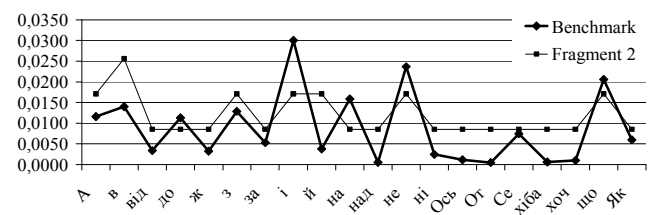


Fig. 3. Relative frequency of auxiliary words appearance in fragment 2 and benchmark.

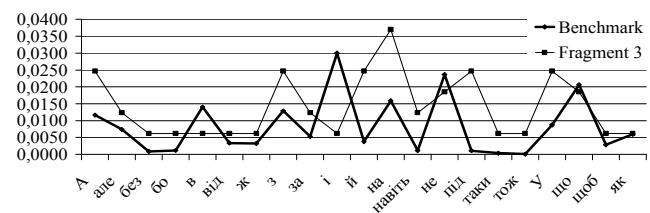


Fig. 4. Relative frequency of auxiliary words appearance in fragment 3 and benchmark

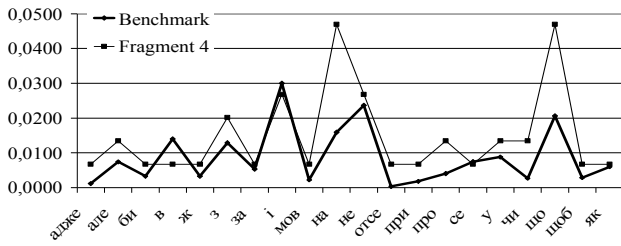


Fig. 5. Relative frequency of auxiliary words appearance in fragment 4 and benchmark

The graphical representation of relative frequency of auxiliary words appearance in fragment 2 and benchmark is shown on fig. 3. The coefficient of correlation in this case is $R_{e-y2}=0,7066$. The graphical representation of relative frequency of auxiliary words appearance in fragment 3 and in benchmark is shown on fig. 4. The coefficient of correlation in this case is $R_{e-y3}=0,2810$. The relative frequencies of auxiliary words in fragment 4 and benchmark are presented in fig.5. The coefficient of correlation is $R_{e-y4}=0,7326$. The correlation coefficients for each auxiliary word for fragments 1-4 are summarized in Table I. After the analysis of correlation coefficients for auxiliary words we conclude that probability of relatedness of fragments to benchmark is largest for fragment 4, the next is fragment 2 and then fragments 1 and 3. Let's note, that for all four fragments there are high correlations for prepositions. This could be interpreted as prepositions not influencing the author's style. Additionally, for selected text fragments, we analyzed the frequency of only pronouns and only conjunctions appearance and calculated corresponding correlation coefficients. The comparison of results is shown in Table II.

TABLE I. CORRELATION COEFFICIENTS FOR AUXILIARY WORDS

Fragment	Pronoun	Conjunction	Preposition
1	$R_{e-y1z}=0,72$	$R_{e-y1s}=0,79$	$R_{e-y1c}=1$
2	$R_{e-y2z}=0,4928$	$R_{e-y2s}=0,5714$	$R_{e-y2c}=0,9580$
3	$R_{e-y3z}=0,1517$	$R_{e-y3s}=0,1624$	$R_{e-y3c}=0,8800$
4	$R_{e-y4z}=0,5639$	$R_{e-y4s}=0,9544$	$R_{e-y4c}=0,9594$

TABLE II. CORRELATION COEFFICIENTS FOR EVERY FRAGMENT

Coefficient	Fragment 1	Fragment 2	Fragment 3	Fragment 4
R_{e-y}	0,6076	0,7066	0,2810	0,7326
R_{e-y}	0,6900	0,4913	0,2254	0,6905

Fragment 4 is still has the biggest probability of being related to benchmark, after it with a small gap go fragments 1, 2, 3. The order is the same as was in previous experiment. For confirmation of results we refer to [25], which provided fragments for experiment. Thus, the usage of supporting words method lead to following results: among researched fragments the greatest probability of relatedness to benchmark is obtained for the fragment which in fact belongs to it [25]. Other results also confirm the applicability of supporting words method for texts authorship attribution. Accordingly, in first experiment the second greater probability of being related to benchmark is assigned to fragment of another text by the same author. Fragment 1 which also belongs to benchmark, lost to fragment 4 only one tenth in correlation coefficient value. Also adequate is the result for fragment 3, which is represents a text nearly

one hundred years distant in its creation time from benchmark. The hypothesis made in [25] that prepositions as method parameters have no influence on author's style definition lead to reduction of correlation coefficients, but put the probability of relatedness to benchmark in right order. Most of all, the difference between correlation coefficients for fragments 1 and 4 was considerably reduced and had value of 0,0005. However, for confirmation or refutation of claim that prepositions are not an important factor in authorship definition additional research should be done.

VI. CONCLUSION

In this work the importance of textual works analysis for plagiarism was shown, types of plagiarism were reviewed and approach to detect author's style was described. The authors researched machine learning methods and linguistic metrics for definition of author's style based on corpus of Ukrainian academic articles. The publications authored by several people present a separate problem. The definition of style in this case is more complex because styles of different authors are superimposing. In this case it is difficult to detect whether the academic work was created by commercial writer.

REFERENCES

- [1] L. Luparenko, "Instrumentariy vyyavlennya plahiatsu v naukovykh robotakh: analiz prohramnykh rishen," *Informatsiyni tekhnolohiyi i zasoby navchannya*, Vol. 2(40), 2014, pp. 151-169.
- [2] V. Petrenko, "Ponyattya ta vydy plahiatsu," *Chasopys tsyvilistyky*, Vol. 14, 2013, pp. 128-131.
- [3] V. Lytvyn, V. Vysotska, P. Pukach, I. Bobyk, and D. Uhryn, "Development of a method for the recognition of author's style in the Ukrainian language texts based on linguometry, stylemetry and glottochronology," *Eastern-European Journal of Enterprise Technologies*, vol. 4/2(88), pp. 10-18, 2017.
- [4] V. Lytvyn, V. Vysotska, P. Pukach, O. Brodyak, and D. Ugryn, "Development of a method for determining the keywords in the slavic language texts based on the technology of web mining," *Eastern-European Journal of Enterprise Technologies*, vol. 2/2(86), pp. 4-12, 2017.
- [5] V. Lytvyn, P. Pukach, I. Bobyk, and V. Vysotska, "The method of formation of the status of personality understanding based on the content analysis," *Eastern-European Journal of Enterprise Technologies*, vol.5/2(83), pp. 4-12, 2016.
- [6] B. Mobasher, *Data mining for web personalization*, The adaptive web, Springer Berlin Heidelberg, 2007, pp. 90-135.
- [7] C. E. Dinuca, and D. Ciobanu, "Web Content Mining," *University of Petrosani, Economics*, pp. 85-92, 2012.
- [8] G. Xu, Y. Zhang, and L. Li. *Web content mining. Web Mining and Social Networking*. Springer US, 2011, pp. 71-87.
- [9] I. Khomytska, and V. Teslyuk, "The Method of Statistical Analysis of the Scientific, Colloquial, Belles-Lettres and Newspaper Styles on the Phonological Level," *Advances in Intelligent Systems and Computing*, pp. 149-163, 2017.
- [10] I. Khomytska, and V. Teslyuk, "Specifics of Phonostatistical Structure of the Scientific Style in English Style System," *Computer Science and Information Technologies, Proc. of the XI-th Int. Conf. CSIT*, pp. 129-131, 2016.
- [11] I. Khomytska, V. Teslyuk, "Modelling of Phonostatistical Structures of English Backlingual Phoneme Group in Style System", *Proceedings of CADMS*, pp. 324-327, 2017.
- [12] I. Khomytska, V. Teslyuk, "Modelling of Phonostatistical Structures of the Colloquial and Newspaper Styles in English Sonorant Phoneme Group", *Computer Science and Information Technologies, CSIT*, pp. 67-70, 2017.
- [13] A. Anisimov, and A. Marchenko, "Sistema obrabotki tekstov na yestestvennom yazyke," *Iskusstvennyy intellekt*, 4, 157-163, 2002.
- [14] V. Perebiynis, "Matematychna lingvistyka," *Ukrains'ka yentsiklopediya*, pp. 287-302, 2000.
- [15] S. Buk, "Osnovy statystychnoy lingvistyky," I. Franka LNU, 2008.
- [16] M. Kochergan, "Vstup do movoznavstva," *Akademiya*, 2005.

- [17] V. Perebinyis, *Statystychni metody dlya lingvistiv*. Nova kniga, 2013.
- [18] M.A. Goncalves, "Streams, Structures, Spaces, Scenarios, and Societies (5S): A Formal Model for Digital Library Framework and Its Applications", PhD thesis, 2004.
- [19] A. Varfolomeyev, "Psikhosemantika slova i lingvostatistika teksta," Kaliningrad, KGU, 2000.
- [20] R. Manekin, "Kognitivnaya stilometriya: k postanovke problemy", (<http://www.manekin.narod.ru/hist/styl.htm>).
- [21] B. Rusyn, O. Lutsyk, O. Lysak, A. Lukeniuk, and L. Pohreliuk, "Lossless Image Compression in the Remote Sensing Applications", *Data stream mining & processing, DSMP*, pp.195-198, 2016.
- [22] Kowalik Dagmara, "Polish vocational competence standards for the needs of adult education and the European labour market", *International Conference on Advanced Information Engineering and Education Science, ICAEES*, pp. 95-98, 2013.
- [23] R. Tkachenko, P. Tkachenko, I. Izonin, Y. Tsymbal, "Learning-based image scaling using neural-like structure of geometric transformation paradigm", In *Studies in Computational Intelligence*, vol. 730, Springer Verlag, pp. 537-565, 2018.
- [24] Y. Rashkevych, D. Peleshko, O. Vynokurova, I. Izonin and N. Lotoshynska, "Single-frame image super-resolution based on singular square matrix operator," 2017 IEEE First Ukraine Conference on Electrical and Computer Engineering (UKRCON), pp. 944-948, 2017.
- [25] Victana, (<http://victana.lviv.ua/index.php/kliuchovi-slova>).
- [26] O. Kanishcheva, V. Vysotska, L. Chyrun, and A. Gozhyj, "Method of Integration and Content Management of the Information Resources Network," *Advances in Intelligent Systems and Computing*, Springer, 2017, pp. 204-216.
- [27] J. Su, V. Vysotska, A. Sachenko, V. Lytvyn, and Y. Burov, "Information resources processing using linguistic analysis of textual content," *Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications*, 9th IEEE International Conference, Bucharest, Romania, pp. 573-578, 2017.
- [28] V. Lytvyn, V. Vysotska, O. Veres, I. Rishnyak, and H. Rishnyak, "The Risk Management Modelling in Multi Project Environment," *Computer Science and Information Technologies, XII-th Int. Conf. CSIT*, pp. 32-35, 2017.
- [29] M. Korobchinsky, V. Vysotska, L. Chyrun, and L. Chyrun, "Peculiarities of Content Forming and Analysis in Internet Newspaper Covering Music News." *Computer Science and Information Technologies, XII-th Int. Conf. CSIT*, pp. 52-57, 2017.
- [30] O. Naum, L. Chyrun, O. Kanishcheva, and V. Vysotska, "Intellectual System Design for Content Formation," *Computer Science and Information Technologies, XII-th Int. Conf. CSIT*, pp. 131-138, 2017.
- [31] V. Lytvyn, V. Vysotska, Y. Burov, O. Veres, and I. Rishnyak, "The Contextual Search Method Based on Domain Thesaurus," *Advances in Intelligent Systems and Computing*, pp. 310-319, 2017.
- [32] M. Davydov, and O. Lozynska, "Information System for Translation into Ukrainian Sign Language on Mobile Devices," *Computer Science and Information Technologies, CSIT*, pp. 48-51, 2017.
- [33] Jivani Anjali Ganesh. "A Comparative Study of Stemming Algorithms," *Int. J. Comp. Tech. Appl.*, vol. 2(6), 1930-1938, 2011.
- [34] A. Mishler, E.S. Crabb, S. Paletz, B. Hefright, and E. Golonka, "Using Structural Topic Modeling to Detect Events and Cluster Twitter Users in the Ukrainian Crisis," *Communications in Computer and Information Science*, vol. 528. Springer, pp. 639-644, 2015.
- [35] M. Davydov, and O. Lozynska, "Mathematical Method of Translation into Ukrainian Sign Language Based on Ontologies," *Advances in Intelligent Systems and Computing*, vol. 689, pp. 89-100, 2018.
- [36] O. Chernukha, and Y. Bilushchak, "Mathematical modeling of random concentration field and its second moments in a semispace with erlangian distribution of layered inclusions", *Task Quarterly*, vol. 20(3), pp. 295-334, 2016.
- [37] K. Kowalska, Cai Di, and S. Wade, "Sentiment analysis of polish texts," *Inter. J. of Computer and Communication Engineering*, vol. 1.1, pp. 39-41, 2012.
- [38] N. Kotsyba, "The current state of work on the Polish-Ukrainian Parallel Corpus," *Organization and Development of Digital Lexical Resources*, pp. 55-60, 2009.
- [39] M. Davydov, and O. Lozynska, "Linguistic Models of Assistive Computer Technologies for Cognition and Communication," *Computer Science and Information Technologies, XI-th Int. Conf. CSIT*, pp. 171-175, 2017.
- [40] V. Lytvyn, V. Vysotska, O. Veres, I. Rishnyak, and H. Rishnyak, "Classification Methods of Text Documents Using Ontology Based Approach," *Advances in Intelligent Systems and Computing*, Springer, 512, pp. 229-240, 2017.
- [41] V. Vysotska, "Linguistic Analysis of Textual Commercial Content for Information Resources Processing," *Modern Problems of Radio Engineering, Telecommunications and Computer Science, TCSET*, pp. 709-713, 2016.
- [42] V. Vysotska, L. Chyrun, and L. Chyrun, "Information Technology of Processing Information Resources in Electronic Content Commerce Systems," *Computer Science and Information Technologies, CSIT*, pp. 212-222, 2016.
- [43] V. Vysotska, L. Chyrun, and L. Chyrun, "The Commercial Content Digest Formation and Distributional Process," *Computer Science and Information Technologies, XI-th Int. Conf. CSIT*, pp. 186-189, 2016.
- [44] V. Lytvyn, V. Vysotska, O. Veres, I. Rishnyak, and H. Rishnyak, "Content linguistic analysis methods for textual documents classification," *Computer Science and Information Technologies, CSIT*, pp. 190-192, 2016.
- [45] V. Lytvyn, and V. Vysotska, "Designing architecture of electronic content commerce system," *Computer Science and Information Technologies, X-th Int. Conf. CSIT*, 115-119, 2015.
- [46] V. Vysotska, and L. Chyrun, "Analysis features of information resources processing," *Computer Science and Information Technologies*, pp. 124-128, 2015.
- [47] Vasyl Lytvyn, Victoria Vysotska, Dmytro Dosyn, Roman Holoschuk, and Zoriana Rybchak, "Application of Sentence Parsing for Determining Keywords in Ukrainian Texts," *Computer Science and Information Technologies*, pp. 326-331, 2017.
- [48] O. Maksymiv, T. Rak, and D. Peleshko, "Video-based Flame Detection using LBP-based Descriptor: Influences of Classifiers Variety on Detection Efficiency," *International Journal of Intelligent Systems and Applications*, vol. 9(2), pp. 42-48, 2017.
- [49] D. Peleshko, T. Rak, and I. Izonin, "Image Superresolution via Divergence Matrix and Automatic Detection of Crossover," *International Journal of Intelligent Systems and Application*, vol. 8(12), pp. 1-8, 2016.
- [50] O. Bazylyk, P. Taradaha, O. Nadobko, L. Chyrun, and T. Shestakevych, "The results of software complex OPTAN use for modeling and optimization of standard engineering processes of printed circuit boards manufacturing," *Proceedings of the 11th International Conference TCSET*, pp. 107-108, 2012.
- [51] A. Bondariev, M. Kiselychnyk, O. Nadobko, L. Nedostup, L. Chyrun, and T. Shestakevych, "The software complex development for modeling and optimizing of processes of radio-engineering equipment quality providing at the stage of manufacture," *Proceedings of the 11th International Conference TCSET*, pp. 159, 2012.
- [52] V. Riznyk, "Multi-modular Optimum Coding Systems Based on Remarkable Geometric Properties of Space," *Advances in Intelligent Systems and Computing*, 512, pp. 129-148, 2017.
- [53] V. Teslyuk, V. Beregovskiy, P. Denysyuk, T. Teslyuk, and A. Lozynskiy, "Development and Implementation of the Technical Accident Prevention Subsystem for the Smart Home System," *International Journal of Intelligent Systems and Applications*, vol. 10(1), pp. 1-8, 2018.
- [54] T. Basyuk, "The main reasons of attendance falling of internet resource," *CSIT*, pp. 91-93, 2015.
- [55] V. Pasichnyk, and T. Shestakevych, "The model of data analysis of the psychophysiological survey results," *Advances in Intelligent Systems and Computing*, vol. 512, pp. 271-281, 2017.
- [56] P. Zhezhnych, and O. Markiv, "Linguistic Comparison Quality Evaluation of Web-Site Content with Tourism Documentation Objects," *Advances in Intelligent Systems and Computing*, vol. 689, pp. 656-667, 2018.
- [57] P. Kravets, "The game method for orthonormal systems construction", *The Experience of Designing and Application of CAD Systems in Microelectronics, CADSM 2007*.
- [58] P. Kravets, "Game Model of Dragonfly Animat Self-Learning", *Perspective Technologies and Methods in MEMS Design*, pp. 195-201, 2016.
- [59] P. Kravets, "The control agent with fuzzy logic", *Perspective Technologies and Methods in MEMS Design*, pp. 40-41, 2010.
- [60] Google - word2vec, (github.com/danielfrg/word2vec/blob/master/examples/word2vec.ipynb).
- [61] V. Lytvyn, V. Vysotska, P. Pukach, M. Vovk, and D. Ugryn, "Method of functioning of intelligent agents, designed to solve action planning problems based on ontological approach", *Eastern-European Journal of Enterprise Technologies*, vol. 3/2(87), pp. 11-17, 2017.
- [62] V. Lytvyn, V. Vysotska, L. Chyrun, and L. Chyrun, "Distance Learning Method for Modern Youth Promotion and Involvement in

- Independent Scientific Researches”, *Data Stream Mining & Processing*, pp. 269-274, 2016.
- [63] V. Vysotska, I. Rishnyak, and L. Chyrun, “Analysis and evaluation of risks in electronic commerce”, *CAD Systems in Microelectronics*, 9th International Conference, pp. 332-333, 2007.
- [64] V. Lytvyn, V. Vysotska, O. Veres, I. Rishnyak, and H. Rishnyak, “The Risk Management Modelling in Multi Project Environment”, *Computer Science and Information Technologies*, pp. 32-35, 2017.
- [65] M.A. Gonçalves, E.A. Fox, L.T.nWatson, and N.A. Kipp, “Streams, Structures, Spaces, Scenarios, Societies (5S): A Formal Model for Digital Libraries”, *ACM Transactions on Information Systems (TOIS)*, vol. 22(2), pp. 270-312, 2004.
- [66] A. Pérez, M. Enrech, “Virtual Library Services for a Virtual University: User-Oriented Virtual Sites in an Open Library”, *EADTU*, Paris, 1999.
- [67] A. Pérez, M. Enrech, “Defining library services for a virtual community”, *Libraries Without Walls Conference*, Lesvos, Grecia, Centre for research in Library and Information Management, 1999.
- [68] V. Pasichnyk, T. Shestakevych, “The method of education format ascertaining in program system of inclusive education support”, *Computer Science and Information Technologies*, 279-284, 2017.
- [69] V. Pasichnyk, T. Shestakevych, “The application of multivariate data analysis technology to support inclusive education”, *Computer Science and Information Technologies*, pp. 88-90, 2015.

Intelligent Analysis of Data Systems for Defects in Underground Gas Pipeline

Volodymyr Yuzevych
*Department of Electrophysical Methods
of nondestructive testing
Karpenko Physico-Mechanical Institute
of the National Academy of Sciences of
Ukraine*
Lviv, Ukraine
yuzevych@ukr.net

Ruslan Skrynkovskyy
*Department of Business Economy and
Information Technology
Lviv University of Business and Law*
Lviv, Ukraine
uan_lviv@ukr.net

Bohdan Koman
*Department of System Design
Ivan Franko National University of
Lviv*
Lviv, Ukraine
sonce_28@ukr.net

Abstract—A method of functioning of intelligent software and hardware complex for monitoring system of an underground gas pipeline and cathodic protection devices using data and knowledge bases is proposed.

Keywords—data mining, gas pipeline, intelligent software, hardware, monitoring, cathodic protection, databases.

I. INTRODUCTION

Efficiency of protection of underground steel gas main-pipelines (USGP) against corrosion depends on a state and quality of pipes. The pipes are protected by coatings and the cathodic protection equipment (CPE). The surface of the pipes becomes covered with defects during operation, in particular, the corrosive ones. Coefficient of corrosion inhibition depends on the defects.

Defects of the pipelines (USGP) can be divided into two classes [1]:

1 – defects of continuity (defects of a material) featuring with local discontinuity of the material (steel);

2 – shape defects revealing as local changes of the pipeline elements during production process or operation.

We will restrict ourselves to consideration of the defect system of a pipeline (USGP) and processes of the corrosion-mechanical fracture, in particular, stress corrosion cracking (SCC), which are appropriate to analyze on the base of a software and hardware complex.

The problem is to evaluate efficiency and quality of the software and hardware complex (SHC) to ensure conditions of the pipeline operation with accounting for the CPE, the SCC processes, and application of data and knowledge bases.

II. AN ANALYSIS OF RECENT PUBLICATIONS ON THE PROBLEM

We have addressed a perspective of application of the functioning method of intelligent SHC to monitor and ensure safety of the pipeline operation by means of databases [2]. Such database comprises the following: data of continuous monitoring of the pipeline actual state, technical documentation data, results of internal pipe diagnostics, data of electrometric measurements, data of visual and dimensional control, data of periodic non-destructive measurements, corrosive and mechanical characteristics of a

metal, criteria of limit states of pipeline systems with damages [3].

The information presented deals with a technique of control of complex interacting processes allowing development of an intelligent system based on knowledge about peculiarities of the interaction of process participants in a specific subject field [3]. The technique is developed on the base of models, methods and algorithms of the ontological analysis and data processing. The novelty of corresponding mathematical model underlying the intelligent system consists in a combination of various mechanisms of logical conclusion, when making a decision in a problem situation, based on all available information about the subject area [3].

Diagnostic weight and value of attributes proposed to use for the process optimization have been formulated [4].

A new approach is proposed which solves the problem of automated intelligent diagnostic using machine learning techniques [5].

III. FORMULATION OF THE RESEARCH GOALS

The purpose of the study is to develop methodological and theoretical foundations of diagnostics of states of a complex technical system such as the system of defects of a pipeline (SDP), and corresponding technological processes based on analysis and processing of knowledge under conditions of uncertainty as well as development of new intelligent methods and tools of multi-criterial diagnostics of the object's states (SDP).

The diagnostics relates to the pipeline critical situation. The critical situation (state) emerging during the process of the pipeline operation reveal itself as a destruction of the pipe metal due to the stress corrosion cracking (SCC). Main informational parameters of the pipeline are pipe diameter, wall thickness, internal pressure, yield and strength limits of material (steel), and energy characteristics of the surface layers.

IV. INFORMATION ON COMBINATION OF CURRENT AND POTENTIAL MEASUREMENTS

Recognition of the object's parameters of state (SDP) is performed by methods of solution finding on the base of logical rules and precedents of solutions of the diagnostics problems.

A problem of the value of information obtained during diagnosis of underground gas pipelines by means of contactless current measurements (CCM) is raised [6].

The implementation of the CCM method in the device for measuring polarization potential (MPP) provides an opportunity to use the MPP in detection of the USGP damaged insulation both at alternating current (Pearson's method) and by the potential difference (gradient method) on the surface of the soil [6].

To determine places of damage of the pipe protective cover, two electrodes are placed on the soil surface above the pipeline and alternative voltage V_{gg} (Pearson's method) as well as dc voltage U_{gg} (method of transversal gradient of potential) are measured [6]. Combination of these measurements allows determining polarization potential U_P by the formula [6]:

$$U_P = U_{mg} - V_{mg} \times U_{gg} / V_{gg} . \quad (1)$$

Here, U_{MG} , V_{MG} – correspondingly the constant potential difference between metal (pipe) and comparison electrode placed in corrosive environment.

To implement the CCM method, a hardware with electronic memory to measure constant and variable electrical voltages as well as polarization potential (PP) [6] was designed. The MPP method can be used to control pipelines and other metal constructions by parameters indicated in (1). In the version of equipment in [6], a GPS module was installed allowing additionally to fix place and time of control. This facilitates considerably processing and documenting the data arrays, in particular, combination of measurements of potential with measurements of current during their processing and determination of the USGP parameters.

In formula (1), second term is an ohmic component of the potential characterizing state of the protective coating. Ratio of the ac voltage to the dc one, $V_{gg} / U_{gg} = k$, is the measured harmonic factor showing ratio of loss of the measured ac component to the dc one [6]. It is used to assess losses of the dc component of the CPE on the pipe sections and to determine distribution of the current density of cathodic protection.

The use of GPS modulus in the developed device provides automatic fixation of geographical coordinates and time of the current measurement as well as the USGP depth, which greatly facilitates processing and documentation of the inspection results [6]. It is necessary for computing the current density, its losses, and transition resistance "pipe-ground" in different parts of the USGP. The CMC together with MPP allow determining distributions of the current density of cathodic protection, resistivity of soil surrounding the pipe, and resistivity of the protective insulation on different sections of the USGP [6].

V. PRINCIPLES OF DIAGNOSTING THE PIPELINE DEFECTS SYSTEM

Under conditions of operation and monitoring, conditions of functioning of objects of the USGP system and CPE change. Let us consider project of the USGP monitoring which main element is the SDP monitoring.

Underground pipelines are in soil environment. Anticorrosion coating of pipelines can be metal and film-type. Between metal of the pipeline and coating, defects of the cavern type are formed. Water from the soil environment penetrates into the defects that is the aqueous electrolyte solution is created. Electrochemical reactions and adsorption processes are characteristic for such type of defects [7].

There are presented mathematical relationships and methods of evaluation of physical and electrochemical characteristics of the interphase layer at the metal-solution boundary of the electrolyte together with corresponding algorithms and software supplemented by the base of numerical data [7]. These data are information means that is a base of information technology of selection and processing data concerning assessment of energy characteristics of interphase layers and overstrain characterizing the boundary metal-environment and conditions of metal plastic deformation near the cavern tip [7]. The database of such kind allows particularly to describe state of the interphase metal layer with adsorbed impurities in corrosive environment of the soil electrolyte type.

Number of parameters and volume of information required for ordering and evaluation in order to take a decision in critical situation are growing. For this purpose, both the database and knowledge base are used.

The knowledge base is part of the decision support system (DSS). The DSS has to contain information, which is partially analogous to the [2] and characterizing the following:

- conditions of implementation of physical and chemical processes;
- probable results of physical and chemical processes;
- conditions of the CPE functioning;
- time periods related to the USGP reliability control;
- degree of formalization of the decision making process with accounting for normative documentation;
- probable critical situations, associated with the risk of the USGMP and CPE system, with a note of the reasons and conditions of their occurrence, and measures to correct deviations from the operating modes.

The most important for information selection and processing is the last sixth paragraph, associated with making decisions for the models of knowledge representation and their following usage in the knowledge base.

To implement procedures of monitoring the USGP system and CPE with accounting for a feedback, the diagnostic weight of an attribute is introduced. If in the result of the study it is revealed that the attribute k_j has for the given object a value of k_{js} , then this value is called the realization of the attribute k_j [8]. Then information about specific state (diagnosis) D_i ($i=1, 2, \dots, n$ – total number of states under consideration), which possesses a state of the attribute k_{js} , can be defined by the formula in the information theory [8]:

$$Z_{Di}(k_{js}) = \ln(P(D_i / k_{js}) / P(D_i)), \quad (2)$$

where $P(D_i/k_{js})$ – probability of the state D_i provided that the attribute k_j obtained value of k_{js} ;

$P(D_i)$ – a priori probability of state.

For specific calculations, the diagnostic weight of presence of the attribute k_j within the interval s can be written in an equivalent form relatively to (2) analogously to [8, 9]:

$$Z_{D_i}(k_{js}) = \ln(P(k_{js}/D_i)/P(k_{js})). \quad (3)$$

Equivalency of relationships (2) and (3) follows from identity [8, 9] based on the probability theory:

$$P(k_{js}) \cdot P(D_i/k_{js}) = P(D_i) \cdot P(k_{js}/D_i) = P(k_{js}D_i). \quad (4)$$

From the point of view of information theory [8, 9], the value $Z_{D_i}(k_{js})$ characterizes information about the D_i state, which has attributes k_{js} . In expression (4), the $P(k_{js}/D_i)$ value is the probability of occurrence of interval s of the attribute k_j for an element of the system in the D_i state, and $P(k_{js})$ is the probability of simultaneous occurrence of corresponding interval of each attribute in every state considered.

The $P(k_{js})$ value is determined according to [8, 9]:

$$P(k_{js}) = \sum_{i=1}^n P(D_i) \cdot P(k_{js}/D_i). \quad (5)$$

With accounting for the relationship (5), we will obtain resulting expression:

$$Z_{D_i}(k_{js}) = \ln \left(P(k_{js}/D_i) / \left(\sum_{i=1}^n P(D_i) \cdot P(k_{js}/D_i) \right) \right). \quad (6)$$

Let us introduce important concept of the information theory – information or diagnostic value of the $Z_{D_i}(k_j)$ study for the USGP system and CPE. The $Z_{D_i}(k_j)$ diagnostic value by the attribute k_j for the state D_i is a volume of information introduced by all variants of realizations of this attribute in establishing a corresponding state [9]. Expression $Z_{D_i}(k_j)$ for the m-bit attribute is proposed to write in form of [9]:

$$Z_{D_i}(k_j) = \sum_{s=1}^m P(k_{js}/D_i) Z_{D_i}(k_{js});$$

$$Z_{D_i}(k_j) = \sum_{s=1}^m P(k_{js}/D_i) \log \frac{P(k_{js}/D_i)}{\sum_{i=1}^n P(D_i) \cdot P(k_{js}/D_i)}. \quad (7)$$

Diagnostic value of the process of diagnosing $Z_{D_i}(k_j)$ for the USGP system and CPE considers all possible realizations of the attribute, which corresponds to a specific state (diagnosis) and is certain mean expected value, more precisely, it is mathematical expectation of the information value introduced by separate realizations of the attribute in this state [8]. Because the $Z_{D_i}(k_j)$ is attributed to a single specific state, it is commonly referred to as the partial diagnostic value by the attribute k_j [4, 9].

A technique to predict situations of the USGP system and CPE is proposed, consisting in accounting for main

informative parameters by means of artificial neural networks, as well as were defined directions of application of the “data mining” methodology to control limit states of pipes on the base of the strength and yield criteria.

The neural network with direct signal distribution to assess defects on the gas pipe surface is proposed. A mean square error averaged by the number of the neural network output variables and calculated on the base of predicted and real values of the test sample by formula [10] was used to determine efficiency of the neural network being studied:

$$E = \frac{1}{N \times K} \sum_{i=1}^N \sum_{k=1}^K (y_{ij}^R - y_{ij}^P)^2. \quad (8)$$

Here, y_{ij}^R – value of the i -th output variable of neural network for j -th training or test example;

y_{ij}^P – predicted value of the i -th output variable of neural network for j -th training or test example;

N – number of examples in the training or test sample;

K – number of output variables of the neural network.

Activation functions are selected by an exhaustive search over the given set [11] including:

- threshold;
- symmetric threshold;
- sigmoid;
- stepwise sigmoid;
- symmetric sigmoid;
- stepwise symmetric sigmoid;
- Gaussian;
- symmetric Gaussian;
- stepwise Gaussian;
- Elliot function;
- symmetric Elliot function;
- linear function;
- symmetric linear function;
- symmetric sine;
- symmetric cosine;
- sine;
- cosine.

To optimize information flows and improve the project configuration, let us use the quality functional $J(P_k, FB(P_k))$ with accounting for sensitivity coefficient β_R and a feed-back [12]:

$$J(P_k, FB(P_k)) = \int_{t_0}^{t_k} f(\bar{y}, \bar{u}, \bar{s}, \bar{\beta}_R) dt \Rightarrow opt, \quad (9)$$

where \bar{y} – vector of specific impacts upon the OP ($y_j(t)$ – vector components (key parameters for the USGP system and CPE), $j = 1, 2, \dots, n$);

\bar{u} – control vector of information flows;

\bar{s} – vector of indeterminate perturbations;

P_k – information flows ($k = 1, 2, \dots, m$);

m – total number of information flows P_k considered in the given project;

$[t_0, t_k]$ – time interval, in which the process is considered (formation of optimal values of parameters corresponding to P_k);

$f(\bar{y}, \bar{u}, \bar{s}, \beta_R)$ – function reflecting quality index of the project;

β_R – sensitivity coefficient;

$FB(P_k)$ – function characterizing feed-back between flows P_k and project's environment with accounting for sensitivity coefficient β_R and expert opinions, of experts;

opt – optimization symbol;

t – time.

To implement above processes, it is proposed to use the intelligent predictive control system of technological processes (IPCS TP) pooled into unified information complex analogously to [13]. Here, we recommend to combine the information complex, equipment for measurements of constant and alternating voltages and determination of polarization potential [6] into unified information space.

This allows the following:

- to expand considerably spectrum of tasks associated with control system;
- provide personnel with information on state of hardware and software;
- decrease risks and improve reliability of the measuring hardware complex;
- to increase considerably general informativity of components of corresponding information space;
- to automate procedure of formation of normative-technical documentation;
- to attract information on strength criteria of elements of metal constructions with consideration of energy characteristics of surface layers [14].

During the process of monitoring the USGP system and CPE, values of physical and chemical parameters are recorded through set time intervals and form a system of interrelated time series. The possibility of free access to information saved in databases is an important perspective, because in the future only small part of the general volume of information can be necessary.

For the time being, prediction of parameters characterizing the state of a technical object is made as a rule on the base of classical models of auto-regression – integrated moving average [15]. Existing approaches to

information processing in control systems as a rule do not provide required accuracy of prediction, leading to increase of probability of making erroneous decision in the object control [15]. Therefore, the prediction algorithms need to be improved. Let us formulate the problem of a variable values in a general case for a discrete time analogously to [13]:

$$Y_{k+\tau} = F_Z(Y_{k-m}, \varphi(Y_k, V_k, y_j, k)) . \quad (10)$$

Here, $Y_{k+\tau}$ – vector of prediction for anticipation interval;

τ , k – actual time interval (clock time);

$Y_{k+\tau}$ – vector of parameter values (e.g. one of the parameters $y_j(t)$) with memory depth m ;

Y_k – vector of values of prehistory of corresponding parameter; V_k – white (Gauss) noise;

$F_Z(\cdot)$ – generalized transformation function (method, algorithm);

$\varphi(\cdot)$ – linear independent functions characterizing properties of the time series.

Prediction of critical state of the system studied can be done by a criterion of the prediction error minimum:

$$\varepsilon = |Y_{k+\tau} - Y_{k+\tau 0}| \Rightarrow \min \text{ when } \tau = \tau_c, \quad (11)$$

where $Y_{k+\tau 0}$ – actual values of technological variable;

τ_c – set value of anticipation interval.

The solution of the problem of prediction for the USGP system and CPE (i.e. finding generalized function F_Z and parameters $y_j(t)$) is formed by means of interpolation of a temporary series and extrapolation of values of predicted series by its initial values by means of functions $\varphi(\cdot)$ in order to ensure criteria (8) and (9).

Results of prediction of the technical state allow not only foresee places of failure of the structural elements and outage of gas transportation system but also to determine the optimal periodicity of running diagnostics procedure and repairs.

Application of algorithms for predicting technical condition of a pipeline allows considerably improve the efficiency of the technical diagnostic system, which in turn leads to improvement of the operation of monitoring equipment described in [6]. In this case, risks of occurrence of the main crack will be reduced, and such situation will assist to provide the desired level of security of gas transportation.

When running procedure of predicting technical state of the USGP system and CPE, not only data obtained in the actual time are used. Results of previous measurements are considered too. Knowledge of previous results of the object diagnosing indicates the need to use databases with the large volume of memory. That said, saving all information about the state of the USGP system and CPE for the whole period of operation in databases is not worthwhile because of many reasons.

In particular, results of measurements of currents and potentials prior to a pipeline repair and afterwards may not correlate between themselves. The reason can be that after

repair some elements of structures can be replaced, for example pipes in some sections. As a consequence, data processing can produce incorrect results of the prediction, and thus a certainty of the results of diagnostic procedure of the USGP system and CPE will be minimal. Another reason can be excessively large volume of data.

To improve the efficiency of the intelligent information system (IIS), it is practical to include an inductive component (IC) into the IIC structure, which allows automatically complete the database [16]. This, in turn, set forth a problem of unification of deductive and inductive formalisms into a single system that leads to the necessity to develop structural and functional models of the IIC with the inductive component (IIC IC), as well as algorithms of efficient search of a solution as a base for development of the quality IIC independently of their complexity and character [16]. We propose to use the basic intelligent component in the process of neural network spectral analysis, which is able to adapt to requirements of a specific sensor [17]. An artificial neural network does not give possibility to create new, unique architecture, but only allows to bring in an ordinary software and hardware for calculations. Such component features with high reliability, ability to adapt to a specific application, and usage of design principles ensuring the possibility of a simple expansion of the intelligent component potential by means of completion with new algorithmic solutions [17].

A large amount of experimental data is provided in article [6]. Processing the results of experiment [6] allows to determine distributions of cathodic protection current densities, specific resistances of surrounding the pipe soil and protective isolation on the different sections of the USGP. Corresponding information will allow to predict the resource of separate areas of gas pipeline and determine the terms of repair.

VI. CONCLUSIONS

A method of functioning of intelligent software and hardware complex for the monitoring system of an underground steel gas pipelines (USGP) and cathodic protection equipment (CPE) using data and knowledge bases is proposed. The data and knowledge bases for monitoring the USGP system and CPE comprise the following:

- data of continuous monitoring of information about actual state of the system of corrosion defects,
- data of normative-technical documentation, data of diagnosing underground gas pipelines by means of contactless current measurements,
- data on critical risk-related cases for the USGP system and CPE with indication of reasons and conditions of their occurrence, and also procedures concerning correction of deviations from a pipeline operating modes,
- data on the control measurements of currents and potentials,
- data of nondestructive control with accounting for the stress corrosion cracking (SCC),
- corrosion-mechanical characteristics of a metal,
- strength and yield criteria for a pipe material.

REFERENCES

- [1] A. Cosham, and P. Hopkins, "An Overview of the pipeline defect assessment manual (PDAM)," proceedings of 4th International Pipeline Technology Conference, Oostende, Belgium, pp. 1-12, May 2004.
- [2] N. G. Gubanov, S. V. Susarev, Yu. I. Steblev, and V. I. Batishev, "The method of functioning of an intelligent software and hardware complex for monitoring and ensuring the safety of pipeline operation using a database," Proceedings of the XIX International Conference "Complex Systems: Control and Modeling Problems", Samara, Russia, pp. 96-102, September 2017.
- [3] O. V. Barmina, and N. O. Nikulina, "Intelligent system for interactive business processes management in project-oriented organizations," *Ontology of designing*. vol. 7, no. 1 (23), pp. 48-65, 2017. doi: <https://doi.org/10.18287/2223-9537-2017-7-1-48-65> .
- [4] V. Yuzevych, O. Klyuvak, and R. Skrynkovskyy, "Diagnostics of the system of interaction between the government and business in terms of public e-procurement," *Economic Annals-XXI*, vol. 160, no. 7-8, pp. 39-44, Oct. 2016. doi: <https://doi.org/10.21003/ea.v160-08> .
- [5] N. O. Komleva, K. S. Cherneha, B. I. Tymchenko, and O. M. Komlevoy, "Intellectual approach application for pulmonary diagnosis," 2016 IEEE First International Conference on Data Stream Mining & Processing (DSMP), Aug. 2016. doi: <https://doi.org/10.1109/dsmp.2016.7583505> .
- [6] R. M. Dzhala, B. Y. Verbenets', M. I. Mel'nyk, A. B. Mytsyk, R. S. Savula, and O. M. Semenyuk, "New Methods for the Corrosion Monitoring of Underground Pipelines According to the Measurements of Currents and Potentials," *Materials Science*, vol. 52, no. 5, pp. 732-741, Mar. 2017. doi: <https://doi.org/10.1007/s11003-017-0016-8> .
- [7] R. Dzhala, V. Yuzevych, and M. Melnyk, "Modeling the adsorption connections and their influence on informational parameters of metal-electrolyte interface," *Bulletin of the Lviv Polytechnic National University, Series "Computer Sciences and Information Technologies"*, no. 826, pp. 185-190, 2015.
- [8] C. E. Shannon, "A Mathematical Theory of Communication," *Bell System Technical Journal*, vol. 27, no. 4, pp. 623-656, Oct. 1948. doi: <https://doi.org/10.1002/j.1538-7305.1948.tb00917.x> .
- [9] V. A. Sokolov, "Diagnostic weight of signs and diagnostic value of examination in recognition of the states of elements of building system," *Engineering and Construction Journal*, no. 3(13), pp. 27-31, 2010.
- [10] N. A. Matveeva, L. Y. Martynovych, and U. V. Lazorenko, "Choice of the optimal neural network for determining defects in composite materials," *Bulletin of the Kherson National Technical University*, no. 3(50), pp. 66-70, 2014.
- [11] P. Sibi, S. Allwyn Jones, and P. Siddarth, "Analysis of Different Activation Functions Using Back Propagation Neural Networks," *Journal of Theoretical and Applied Information Technology*, vol. 47, no. 3, pp. 1264-1268, 2013. <http://www.jatit.org/volumes/Vol47No3/61Vol47No3.pdf> .
- [12] N. Krap, V. Yuzevych, "Neural Networks as a tool for managing the configurations of tourist flow projects," *Management of Development of Complex Systems*, no. 14, pp. 37-40, 2013.
- [13] I. Gulina, A. Martynenko, A. Gulin, "Construction of intelligent predictive control systems for nonlinear technological processes," *Information Processing Systems*, no. 3 (149), pp. 101-105, 2017.
- [14] V. M. Yuzevych, R. M. Dzhala, and B. P. Koman, "Analysis of Metal Corrosion under Conditions of Mechanical Impacts and Aggressive Environments," *Metallofizika i Noveishie Tekhnologii*, vol. 39, no. 12, pp. 1655-1667, Mar. 2018. doi: <https://doi.org/10.15407/mfint.39.12.1655> .
- [15] G. E. P. Box, G. M. Jenkins, *Time series analysis: forecasting and control*. San Francisco, CA: Holden-Day. 1976.
- [16] O. Nelles, *Nonlinear System Identification: From Classical Approaches to Neural and Fuzzy Models*. Berlin: Springer, 2001. <https://www.springer.com/us/book/9783540673699> .
- [17] J. H. Holland, *Adaptation in natural and artificial systems. An introductory analysis with application to biology, control and artificial intelligence*, London: Bradford book edition, 1994.

Content Analysis Method for Cut Formation of Human Psychological State

Liliya Chyrun
Information Systems and Network Department
Lviv Polytechnic National University
Lviv, Ukraine
Liliya.B.Chyrun@lpnu.ua

Iaroslav Kis
Information Systems and Network Department
Lviv Polytechnic National University
Lviv, Ukraine
Yaroslav.P.Kis@lpnu.ua

Victoria Vysotska
Information Systems and Network Department
Lviv Polytechnic National University
Lviv, Ukraine
Victoria.A.Vysotska@lpnu.ua

Lyubomyr Chyrun
Information Systems and Network Department
Lviv Polytechnic National University
Lviv, Ukraine
Lyubomyr.V.Chyrun@lpnu.ua

Abstract—Negative factors in shaping the content complicate the process of finding the necessary data when scanning various sources. Increase in volume and change of relevance or dynamics of the content streams (systematic and irregular updates) leads to duplication, information pollution and redundant results in content search. Comprehension and generalization of large dynamic content streams which are continuously generated by Internet resources requires new methods/approaches to search such as content monitoring. Input information for content monitoring is a text in any natural language presented as a sequence of symbols, whilst output information is provided as tables of sections, sentences and lexemes of the analyzed text.

Keywords—analysis of information resources, content-analysis, rating evaluation, content management system.

I. INTRODUCTION

Nowadays the evolution of Internet-resources, which presently displace television, newspapers and magazines, is taking place [1-3]. This is why the need to create an information product targeted to user's needs and satisfaction increases. There are many information resources providing information about world and local news, but many of them are not adaptive or easy to use [4-7]. Foreign resources are much more informative and easier to use than the Ukrainian ones, which are cluttered with excessive content and advertisements. Currently there are not many Ukrainian resources and their development and improvement is important for users. Most of them do not have adaptive layout, which greatly reduces the number of users, as nowadays not only personal computers have Internet access, but also electronic media with different screen resolutions, such as mobile phones, smartphones, tablets, etc are equipped with one. Fewer and fewer new systems and software products are being ordered, as there are many different and similar systems or programs on the market. Therefore, ready system models or their templates enter mass market. However, questions related to the area of expertise of an enterprise or a person and the field of science or industry almost always arise [8-10]. This creates demand for improving existing systems and programs. In order to improve information systems for Formation of Psychological State of Human (ISFPSH) mathematical linguistics and content analysis of text data sets are used. The problem is

the lack of common standardized approach to architecture design and the development of ISFPSH as well as the data processing within system [11-12]. The absence of a general classification leads to issues with defining and forming unified methods for processing data, which in turn causes the problem of creating appropriate software [13-20]. This justifies the purpose, topicality, expediency and direction of research. The peculiarities of the use of ISFPSH are as follows: openness - access for all companies as well as users; globalism - access from anywhere in the world; absence of time restrictions - access at any time; directness - low barriers for entering market; direct user interaction - reduction of distribution channels and the elimination of intermediaries; automatic processing of requests. The urgency of implementing ISFPSH is due to the following factors: globalization leads to business information needs and quick access to this information for a successful business; time savings in obtaining the necessary content, e-commerce content personalization, integration of ISFPSH. The importance and urgency of building an ISFPSH require the theoretical foundations investigation and practical summaries and proposals development. The aim is to develop an information resource with automatic filling content according to user requirements, the need to set the following tasks: automatic generation of commercial content; Automatic collection and content creation; automatic formatting of content; key words and concepts identification; categorization of content; duplicate content identification; digest content formation; selective dissemination of content. It is necessary to develop a general structure of analysis system of social networks users' profiles and their activities to form a cut of individual's psychological state on the basis of "Big Five" model [11]. To complete this, it is advisable to use methods and means of information resources processing on the Internet environment. [12].

II. A FORMAL MODEL OF THE SYSTEM OF FORMATION OF THE STATUS OF PSYCHOLOGICAL STATE OF A PERSON

We will present the system model S of the status development of psychological state of a person based on the content analysis of the text data sets of this individual (for example, comments in social networks) by the tuple [12, 21-25]

$$S = \langle X, Ident, C, ContProc, Q, Const, PrCont, PersPref, AutAd, ContIntegr, Y \rangle \quad (1)$$

where X is the incoming data from personalities of social networks, the psychological state of whom is analyzed (history, profile, posts, comments, likes, community, etc.), $Ident$ is the process of identification of the system users and personalization of personalities, C is the content of the system, $ContProc$ is the process of initial processing of the content (content and spam filtering, spam identification, analysis, saving, elimination of duplication, content blocking, etc.), Q are the requests from users, $Const$ is the process of provision of consistency of the content, $PrCont$ is the provision of analysis of private content, $PersPref$ is the analysis of personal preferences and personal data of the user, $AutAd$ is the provision of analysis of automatic settings and user profile updates, $ContIntegr$ is the provision of integration of data from other systems, including those from other social networks, Y are the results of the users' queries concerning the status of psychological state of a human/person/individual.

1. Algorithm 1. Cut Formation of Psychological State for Person.
2. User authorization/authentication. Program completion in case of authentication error.
3. New research start.
4. Search for user. If the user is not found, it is necessary to carry out re-search or to complete process.
5. Search for information. If the access is closed, it is necessary to carry out re-search of another user or to complete process.
6. Information gathering.
7. Analysis of the psychological state of a person simultaneously in six categories.
8. Results representation on the screen.
9. Obtaining findings and recommendations.
10. Work completion.

The process of generating answer for the user S in the form of the status of psychological state of the analyzed human/person/individual by the main characteristics of the big five is described by superposition of the main functions (input data of one function are the original data of another one) from (1) as follows

$$Y = ContProc \circ PersPref \circ Const \circ AutAd \circ ContIntegr \circ Ident, \quad (2)$$

in this case, the main process is $ContProc$, which is described by the formula

$$Y = ContProc(X, Q, C) = ContAnal \circ ContSav \circ ContBlock \circ ContDupl \circ ContSpFilt \circ Spldent, \quad (3)$$

where $ContAnal$ is the content analysis, input data/requests, $ContSav$ is the saving of content/results, $ContBlock$ is the

content blocking, $ContDupl$ is the elimination of duplication, $ContSpFilt$ is the filtration of content/spam, $Spldent$ is the identification of content/spam. We suggest using the algorithms of analysis of the syntax of the Ukrainian and English-language text for processing and content analysis (algorithm 2) of large arrays of text data for finding and analyzing the marked words [21-25].

We will focus on the features of this particular social network as the source of data for analysis and determination of personality dispositions. With this purpose we will present the main processes of the S system as $PersPref$, $AutAd$, $Ident$ and will detail them by superposition

$$C^{St} = PersPref \circ Const \circ AutAd \circ ContIntegr \circ Ident, \quad (4)$$

where C^{St} is the content as a result of statistical data of the activity of a personality.

The process $PersPref$ of the analysis of personal preferences and personal data of the user will be represented by superposition

$$C^S = PersPref(X, Q, C^{Pf}, C^{Pl}, C^{Sp}, C^{US}, C^{Pc}, C^{Mr}) = MatchPred \circ GamPred \circ ProfProc \circ SitChan \circ GamModer \circ SitAdm, \quad (5)$$

where $C^{Mr} \subseteq C$ is the set of marked words in the content of analyzed personality, $ProfProc$ stands for processing the user profile and the profiles of participants of the experiment, $SitChan$ is the editing of dictionaries, $GamModer$ is the moderation of the rules of content monitoring of text data arrays of a specific individual, content analysis to find the marked words, analysis of the text's syntax and semantics, as well as the rules of formation of the status of psychological state of a personality, $SitAdm$ is the system administration, $GamPred$ is the obtaining of result of formation of the status of psychological state of a personality based on the associative rules, $MatchPred$ is the formation of the status of psychological state of a personality based on associative rules [11-12].

The component of the rules of content-monitoring $GamModer$ is the content search and content analysis of the text. The content analysis is aimed at searching for the content in the data set by universal linguistic units. The unit of account is a quantitative measure of the unit of analysis, which allows registering the frequency (regularity) of occurrence of indicator of the category of analysis in the text. Then the text is analyzed for the presence of certain marked words and the results are categorized according to psychological metrics (consciousness, friendliness, extraversion, emotionality and openness to experience) [12], namely

$$C^S = MatchPred(X, Q, C, P, D, B) = Opn \circ Cns \circ Ext \circ Agr \circ Nrt \circ Filt, \quad (6)$$

where $Filt$ is the process of filtration of the original text, P is the glossary of rules, D are the dictionaries for classification of the text by psychological dispositions of a personality, B is the dictionary of blocked words, C^S is the result of analysis

of text arrays data and construction of the "Big Five" model [11], i.e. the hierarchical model of a personality by the five features. In particular, such features are *the openness to experience* $C^{Opn}=Opn(C^{Filt}, U^{Opn}, P, D)$ through parameters U^{Opn} (u_1^{Opn} is the frequency of occurrence of words associated with benevolence/malevolence, u_2^{Opn} is the frequency of occurrence of words associated with trust/mistrust, u_3^{Opn} is the frequency of occurrence of words associated with warmth/hostility, u_4^{Opn} is the frequency of occurrence of words associated with sincerity/selfishness); *integrity* $C^{Cns}=Cns(C^{Filt}, U^{Cns}, P, D)$ through parameters U^{Cn} (u_1^{Cns} is the spontaneity/deliberation, u_2^{Cns} is the creativity/narrow-mindedness, u_3^{Cns} is the distinction/mediocrity, u_4^{Cns} is the liberality/parochialism); *extraversion* $C^{Ext}=Ext(C^{Filt}, U^{Ext}, P, D)$ through parameters U^{Ext} (u_1^{Ext} is the

sociability/unsociability, u_2^{Ext} is the assertiveness/tranquility, u_3^{Ext} is the activity/passivity); *amiability* $C^{Arg}=Arg(C^{Filt}, U^{Arg}, P, D)$ through parameters U^{Arg} (u_1^{Arg} is the orderliness/negligence, u_2^{Arg} is the thoroughness/carelessness, u_3^{Arg} is the unreliability/reliability) and *neuroticism* $C^{Nrt}=Nrt(C^{Filt}, U^{Nrt}, P, D)$ through parameters U^{Nrt} (u_1^{Nrt} is the relaxation/nervousness, u_2^{Nrt} is the poise/depression, u_3^{Nrt} is the resistance/irritability) [12].

The use cases diagram (Fig. 1) uses two basic types of entities: use cases and actors, among which are the following types of relationships: association — between actor and use case; generalization between actors; synthesis between use cases; inclusion between use cases.

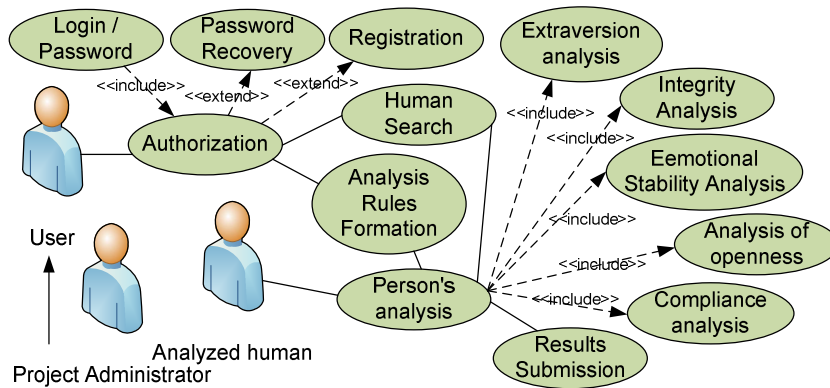


Fig. 1. Use case diagram

There are generalization relationships between project administrator and user. There are association relationships between user and use cases, as well as between investigated person and use cases (i.e. authorization, search, and analysis of individual and results representation). There are generalization relationships between use cases. Inclusion relationships are presented between authorization, password recovery and registration; between person and specific analysis. On the Fig. 2, there is an example of annotated UML-diagram of packets with dependency relationship, which reflects a typical architecture of WEB-based software system to work with the database and decision-making system logic.

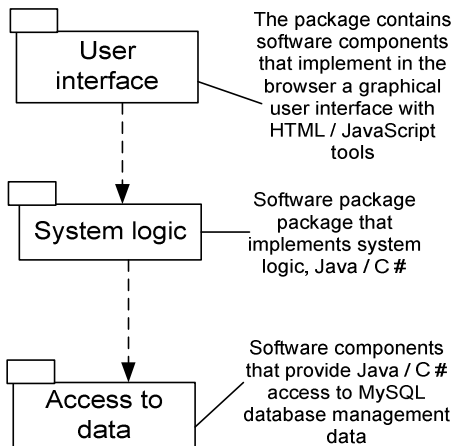


Fig. 2. Package diagram

On sequence diagram (Fig. 3), there is a description of psychological analysis of personality, which is triggered by a particular user of the system [26-37].

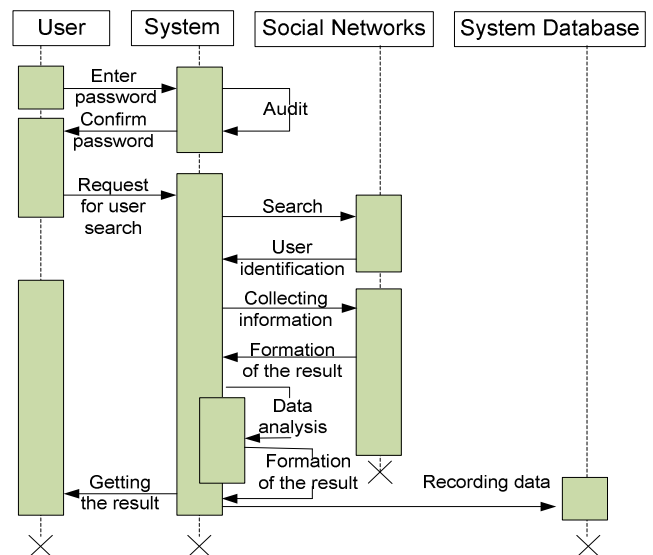


Fig. 3. Sequence diagram

He/she passes authorization and submits a request to search a necessary person in the database. The system finds this person, works out his/her data, and returns the desired result [38-44]. Before shutting, the system makes a record of the given session to the database.

The following diagram needed is activity diagram (Fig. 4) with the following sequence of actions [45-64]:

1. User authorization/authentication. Program completion in case of authentication error.
2. New research start.
3. Search for user. If the user is not found, it is necessary to carry out re-search or complete process.
4. Search for information. If the access is closed, it is necessary to carry out re-search of another user or complete process.
5. Information gathering.
6. Analysis of the psychological state of a person simultaneously in six categories.
7. Results representation on the screen.
8. Obtaining findings and recommendations.
9. Work completion.

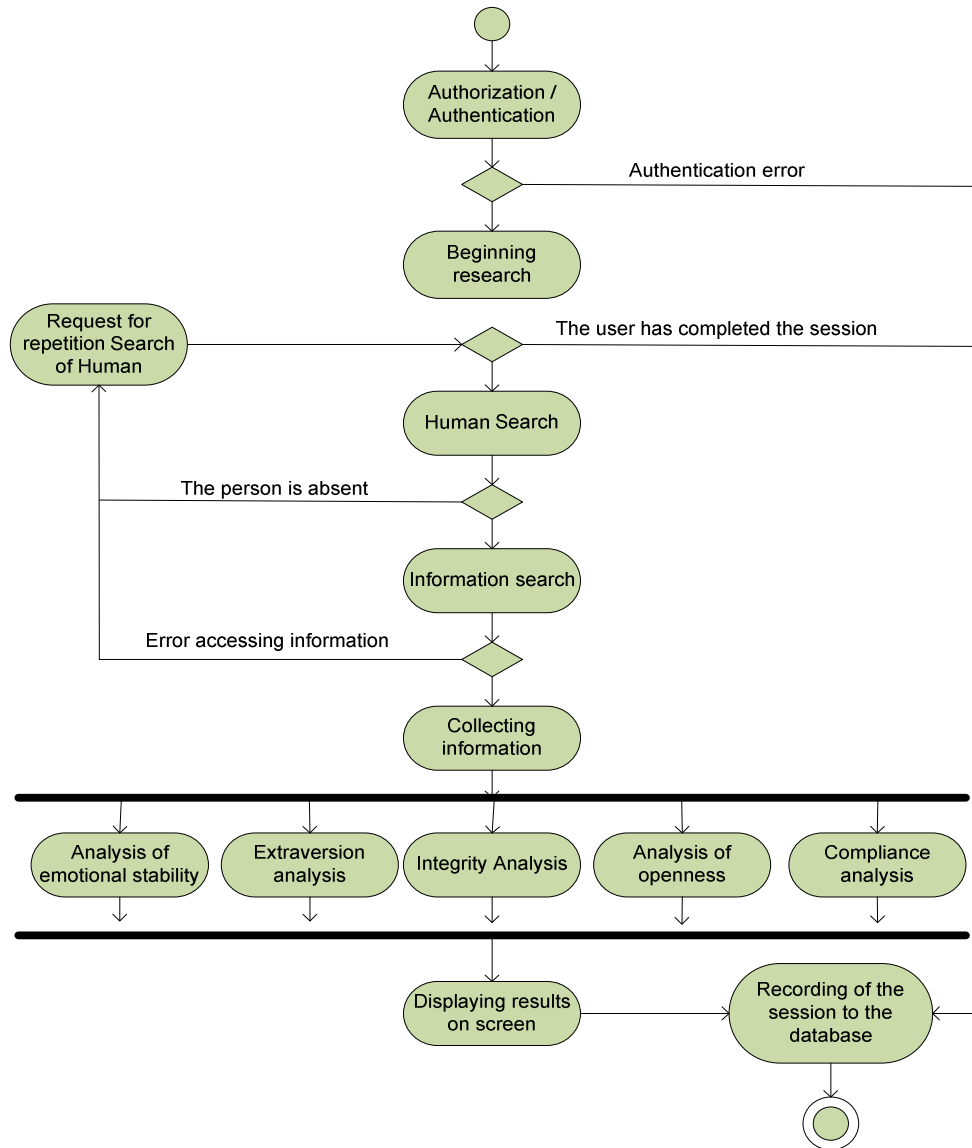


Fig. 4. Activity diagram

This IS is focused on the use of the Internet. Hence, when choosing instruments and means of implementation attention was paid to those technologies that will allow necessary software implementation. During the development of IS, one can face some problems, since for a long time there was no single and a recognized standard for all the Internet technologies. Sure, with the advent of HTML5 and CSS3, as well as MS Edge browser release by Microsoft situation has changed. However, the problem of backward compatibility of sites still remains. When selecting architectural solutions, it is necessary to use two parts: server and client. Server side

is a computer or server that contains a database which stores the necessary information and software components that will implement certain features of the system.

Client side is the environment (browser) that allows displaying the requested information. In this case, it allows displaying website, its pages, content, etc. Both parts are closely related to each other and function as one.

In this case, server hardware will constitute a remote server. In this case, there is no need to create separate server for the project, as there are no data in the program working

process, which could be handled on the server side. Just in future, its implementation is necessary. The client side will be included in the user's device (PC, laptop, tablet, or other device) and will appear in the browser. When developing this service, it is necessary to consider how it will behave on different devices with different screen size and different browsers. Recently, tablets and smartphones have gained special popularity as the most mobile devices. This led to the fact that for viewing content on the site one had to use the zoom, which does not often reflect the content correctly. Therefore, to solve these problems, they apply so-called cross-platform and cross-browser compatibility.

Cross-platform ability allowed to adapt the website to the small screens with diagonals and concentrated at the most desired content only. This reduces the load time of web pages, which became one of the key points. Cross-platform ability may be achieved through using media queries and use of "rubber" layouts.

Cross-browser compatibility is designed to make it possible for site to appear the same on all popular browsers, since each browser had its own certain functions implementation technologies. Therefore, there are different individual scripts and queries in order to make site look the same in any browser.

III. CONCLUSIONS

The actual task of research and development of methods and means for processing data in intellectual information systems forming content with the use of the classification, mathematical and software means and a generalized system architecture. The necessity for development of methods and means of processing data in intelligent information systems forming content by improving system architecture in order to automate processes of formation, management and marketing of content has been justified. Terminology has been analyzed and a classification of intelligent information systems forming content has been created for determining characteristic patterns, trends, processes of system design and modeling as well as to determine the shortcomings of existing methods and tools for content management. A formal model of an intelligent information system for forming content has been developed, allowing us to develop content lifecycle, generalized typical system architecture and standardized methods for processing information resources. General architecture of an intelligent information system for forming content has been improved when compared to existing ones by adding modules for processing of information resources. A complex method of forming content has been developed as well as an operational method for content management and a complex method for content marketing in order to achieve a working effect at the level of a system developer. General recommendations for the design of system architecture have been developed differing from the existing ones with more details of stages and presence of modules for processing information resources. This allows an easy implementation of effective information resources processing at the level of a system developer. The architecture of system modules has been developed to implement the content lifecycle. Applied software means for formation, management and marketing of content has been developed and implemented in order to achieve effect from working at the level of an owner and user of an intellectual information system for forming content.

REFERENCES

- [1] J. Bennett, Visualization Critique, 2012. Retrieved from: <http://vizthinker.com/visualization-critique/>.
- [2] M. Dzheffri, Recruiting 5.0: Psikhologicheskyye profili v sotsial'nykh setyakh, <http://www.hr-portal.ru/blog/rekruting-50-psihologicheskii-profile-v-socialnyh-setyah>.
- [3] V. Lytvyn, P. Pukach, I. Bobyk, and V. Vysotska, "The method of formation of the status of personality understanding based on the content analysis," *Eastern-European Journal of Enterprise Technologies*, 5/2(83), pp. 4–12, 2016.
- [4] Global Web Index, Social Web Involvement, Retrieved from: <http://www.pamorama.net/wp-content/uploads/2010/12/Global-Map-of-Social-Web-Involvement-Global-Web-Index-2009.pdf>.
- [5] D. Kluemper, P. Rosen, and K. Mossholder, "Social Networking Websites, Personality Ratings, and the Organizational Context: More Than Meets the Eye?," *Journal of Applied Social Psychology*, 42(5), pp. 1143–1172, 2012.
- [6] M. Kosinski, D. Stillwell, and T. Graepel, "Private traits and attributes are predictable from digital records of human behavior," *Proceedings of the National Academy of Sciences*, vol. 110(15), 5802–5805, 2013.
- [7] H. Schwartz, J. Eichstaedt, and M. Kern, "Personality, Gender, and Age in the Language of Social Media: The Open-Vocabulary Approach," *Plos One*. 8(9), 2013, <http://journals.plos.org/plosone/article?id=10.1371/journal.pone.007379>.
- [8] O. Kanishcheva, V. Vysotska, L. Chyrun, and A. Gozhnyj, "Method of Integration and Content Management of the Information Resources Network," *Advances in Intelligent Systems and Computing*, 689, Springer, pp. 204–216, 2017.
- [9] J. Su, V. Vysotska, A. Sachenko, V. Lytvyn, and Y. Burov, "Information resources processing using linguistic analysis of textual content," *Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications*, 9th IEEE International Conference, Bucharest, Romania, pp. 573–578, 2017.
- [10] O. Naum, L. Chyrun, O. Kanishcheva, and V. Vysotska, "Intellectual System Design for Content Formation," *Computer Science and Information Technologies*, CSIT, pp. 131–138, 2017.
- [11] B. Shuotian, Z. Tingshao, and C. Li, Big-Five Personality Prediction Based on User Behaviors at Social Network Sites, 2012, <http://arxiv.org/pdf/1204.4809v1.pdf>.
- [12] V. Lytvyn, V. Vysotska, P. Pukach, I. Bobyk, and D. Uhryn, "Development of a method for the recognition of author's style in the Ukrainian language texts based on linguometry, stylemetry and glottochronology," *Eastern-European Journal of Enterprise Technologies*, 4/2(88), pp. 10–18, 2017.
- [13] A. Mishler, E.S. Crabb, S. Paletz, B. Hefright, and E. Golonka, "Using Structural Topic Modeling to Detect Events and Cluster Twitter Users in the Ukrainian Crisis," *Communications in Computer and Information Science*, vol. 528, Springer, pp. 639–644, 2015.
- [14] Jivani Anjali Ganesh, "A Comparative Study of Stemming Algorithms," *Int. J. Comp. Tech. Appl.*, vol. 2(6), 1930–1938, 2011.
- [15] B. Mobasher, Data mining for web personalization. The adaptive web, Springer Berlin Heidelberg, 2007, pp. 90–135.
- [16] C. E. Dinuca, and D. Ciobanu, "Web Content Mining," *University of Petroşani, Economics*, pp. 85–92, 2012.
- [17] G. Xu, Y. Zhang, and L. Li. "Web content mining," *Web Mining and Social Networking*, Springer US, pp. 71–87, 2011.
- [18] K. Kowalska, Cai Di, and S. Wade. "Sentiment analysis of polish texts," *Inter. J. of Computer and Communication Engineering* 1.1, pp. 39–41, 2012.
- [19] N. Kotsyba. "The current state of work on the Polish-Ukrainian Parallel Corpus," *Organization and Development of Digital Lexical Resources*, pp. 55–60, 2009.
- [20] P. Zhezhnych, and O. Markiv, "Linguistic Comparison Quality Evaluation of Web-Site Content with Tourism Documentation Objects," *Advances in Intelligent Systems and Computing*, vol. 689, pp. 656–667, 2018.
- [21] V. Lytvyn, "The similarity metric of scientific papers summaries on the basis of adaptive ontologies," *Perspective Technologies and Methods in MEMS Design*, p. 162, 2011.
- [22] V. Vysotska, I. Rishnyak, and L. Chyrun, "Analysis and evaluation of risks in electronic commerce," *CAD Systems in Microelectronics*, 9th International Conference, Lviv, Ukraine, pp. 332–333, 2007.
- [23] I. Khomytska, and V. Teslyuk, "The Method of Statistical Analysis of the Scientific, Colloquial, Belles-Lettres and Newspaper Styles on the

- Phonological Level,” *Advances in Intelligent Systems and Computing*, vol. 512, pp. 149–163, 2017.
- [24] V. Lytvyn, V. Vysotska, P. Pukach, O. Brodyak, and D. Ugryn, “Development of a method for determining the keywords in the slavic language texts based on the technology of web mining,” *Eastern European Journal of Enterprise Technologies*, 2/2(86), pp. 4–12, 2017.
- [25] I. Khomytska, and V. Teslyuk, “Specifics of Phonostatistical Structure of the Scientific Style in English Style System,” *Computer Science and Information Technologies, CSIT*, pp. 129-131, 2016.
- [26] J. Chen, D. Dosyn, V. Lytvyn, and A. Sachenko, “Smart Data Integration by Goal Driven Ontology Learning,” *Advances in Big Data*, pp. 283–292, 2016.
- [27] V. Lytvyn, V. Vysotska, L. Chyrun, and L. Chyrun, “Distance Learning Method for Modern Youth Promotion and Involvement in Independent Scientific Researches,” *IEEE First International Conference on Data Stream Mining & Processing, Lviv, Ukraine*, pp. 269-274, August 23-27, 2016.
- [28] K. Mykich, and Y. Burov, “Algebraic model for knowledge representation in situational awareness systems,” *Computer Sciences and Information Technologies, CSIT*, pp.165-167, 2016.
- [29] K. Mykich, and Y. Burov, “Uncertainty in situational awareness systems,” *Modern Problems of Radio Engineering, Telecommunications and Computer Science*, pp. 729-732, 2016.
- [30] K. Mykich, and Y. Burov, “Algebraic Framework for Knowledge Processing in Systems with Situational Awareness,” *Advances in Intelligent Systems and Computing*, Springer Verlag, 217-228, 2016.
- [31] K. Mykich, and Y. Burov, “Research of uncertainties in situational awareness systems and methods of their processing,” *Eastern European Journal of Enterprise Technologies*, vol. 1(79), 19-26, 2016.
- [32] V. Lytvyn, V. Vysotska, O. Veres, I. Rishnyak, and H. Rishnyak, “The Risk Management Modelling in Multi Project Environment,” *Computer Science and Information Technologies, Proc. of XII-th Int. Conf. CSIT*, pp. 32–35, 2017.
- [33] Lytvyn V., Vysotska V., Burov Y., Veres O., Rishnyak I. “The Contextual Search Method Based on Domain Thesaurus,” *Advances in Intelligent Systems and Computing*, 689, pp. 310–319, 2017.
- [34] M. Davydov, and O. Lozynska, “Information System for Translation into Ukrainian Sign Language on Mobile Devices,” *Computer Science and Information Technologies, CSIT*, pp. 48-51, 2017.
- [35] P. Kravets, “The control agent with fuzzy logic”, *Perspective Technologies and Methods in MEMS Design*, pp. 40-41, 2010.
- [36] P. Kravets, “The game method for orthonormal systems construction”, *The Experience of Designing and Application of CAD Systems in Microelectronics, CADSM 2007*.
- [37] P. Kravets, “Game Model of Dragonfly Animat Self-Learning”, *Perspective Technologies and Methods in MEMS Design*, pp. 195-201, 2016.
- [38] M. Davydov, and O. Lozynska, “Linguistic Models of Assistive Computer Technologies for Cognition and Communication, *Computer Science and Information Technologies*,” XI-th Int. Conf. CSIT’2017, pp. 171-175, 2017.
- [39] V. Vysotska, “Linguistic Analysis of Textual Commercial Content for Information Resources Processing. *Modern Problems of Radio Engineering*,” *Telecommunications and Computer Science, TCSET’2016*, pp. 709–713, 2016.
- [40] V. Vysotska, L. Chyrun, and L. Chyrun, “Information Technology of Processing Information Resources in Electronic Content Commerce Systems,” *Computer Science and Information Technologies, CSIT’2016*, pp. 212–222, 2016.
- [41] V. Vysotska, L. Chyrun, and L. Chyrun, “The Commercial Content Digest Formation and Distributional Process,” *Computer Science and Information Technologies, Proc. of the XI-th Int. Conf. CSIT’2016*, pp. 186–189, 2016.
- [42] V. Lytvyn, V. Vysotska, O. Veres, I. Rishnyak, and H. Rishnyak, “Content linguistic analysis methods for textual documents classification,” *Computer Science and Information Technologies, CSIT’2016*, pp. 190–192, 2016.
- [43] V. Lytvyn, and V. Vysotska, “Designing architecture of electronic content commerce system,” *Computer Science and Information Technologies, proc. of the X-th Int. Conf. CSIT’2015*, 115–119, 2015.
- [44] V. Vysotska, and L. Chyrun, “Analysis features of information resources processing,” *Computer Science and Information Technologies*, pp. 124–128, 2015.
- [45] V. Lytvyn, V. Vysotska, D. Dosyn, R. Holoschuk, and Z. Rybchak, “Application of Sentence Parsing for Determining Keywords in Ukrainian Texts,” *Computer Science and Information Technologies*, pp. 326–331, 2017.
- [46] V. Lytvyn, O. Tsmots, “The process of managerial decision making support within the early warning system”, *Actual Problems of Economics*, Vol. 11(149), pp. 222-229, 2013.
- [47] O. Maksymiv, T. Rak, and D. Peleshko, “Video-based Flame Detection using LBP-based Descriptor: Influences of Classifiers Variety on Detection Efficiency,” *International Journal of Intelligent Systems and Applications*, vol. 9(2), pp. 42–48, 2017.
- [48] D. Peleshko, T. Rak, and I. Izonin, “Image Superresolution via Divergence Matrix and Automatic Detection of Crossover,” *International Journal of Intelligent Systems and Application*, vol. 8(12), pp. 1–8, 2016.
- [49] A. Pérez, M. Enrech, “Virtual Library Services for a Virtual University: User-Oriented Virtual Sites in an Open Library”, *EADTU, Paris*, 1999.
- [50] A. Pérez, M Enrech, “Defining library services for a virtual community”, *Libraries Without Walls Conference, Lesvos, Grecia, Centre for research in Library and Information Management*, 1999.
- [51] V. Riznyk, “Multi-modular Optimum Coding Systems Based on Remarkable Geometric Properties of Space,” *Advances in Intelligent Systems and Computing*, 512, pp. 129–148, 2017.
- [52] V. Teslyuk, V. Beregovskiy, P. Denysyuk, T. Teslyuk, and A. Lozynskiy, “Development and Implementation of the Technical Accident Prevention Subsystem for the Smart Home System,” *International Journal of Intelligent Systems and Applications*, vol. 10(1), pp. 1–8, 2018.
- [53] T. Basyuk, “The main reasons of attendance falling of internet resource,” *Computer Science and Information Technologies*, pp. 91–93, 2015.
- [54] M. Korobchinsky, V. Vysotska, L. Chyrun, and L. Chyrun, “Peculiarities of Content Forming and Analysis in Internet Newspaper Covering Music News,” *Computer Science and Information Technologies, CSIT’2017*, pp. 52–57, 2017.
- [55] M.A. Gonçalves, E.A. Fox, L.T.nWatson, and N.A. Kipp, “Streams, Structures, Spaces, Scenarios, Societies (5S): A Formal Model for Digital Libraries”, *ACM Transactions on Information Systems (TOIS)*, vol. 22(2), pp. 270-312, 2004.
- [56] N. Shakhovska, V. Vysotska, and L. Chyrun, “Features of E-Learning Realization Using Virtual Research Laboratory,” *Computer Science and Information Technologies, CSIT’2016*, pp.143-148, 2016.
- [57] R. Tkachenko, P. Tkachenko, I. Izonin, Y. Tsymbal, “Learning-based image scaling using neural-like structure of geometric transformation paradigm”, In *Studies in Computational Intelligence*, vol. 730, Springer Verlag, pp. 537–565, 2018.
- [58] Y. Rashkevych, D. Peleshko, O. Vynokurova, I. Izonin and N. Lotoshynska, "Single-frame image super-resolution based on singular square matrix operator," 2017 IEEE First Ukraine Conference on Electrical and Computer Engineering (UKRCON), pp. 944-948, 2017.
- [59] O. Chernukha, and Y. Bilushchak, “Mathematical modeling of random concentration field and its second moments in a semispace with erlangian distribution of layered inclusions”, *Task Quarterly*, vol. 20(3), pp. 295-334, 2016.
- [60] B. Rusyn, O. Lutsyk, O. Lysak, A. Lukeniuk, and L. Pohreliuk, “Lossless Image Compression in the Remote Sensing Applications”, *Data stream mining & processing, DSMP*, pp.195-198, 2016.
- [61] Kowalik Dagmara, “Polish vocational competence standards for the needs of adult education and the European labour market”, *International Conference on Advanced Information Engineering and Education Science, ICAEES*, pp. 95-98, 2013.
- [62] I. Khomytska, V. Teslyuk, “Modelling of Phonostatistical Structures of English Backlingual Phoneme Group in Style System”, *Proceedings of CADMS*, pp. 324-327, 2017.
- [63] I. Khomytska, V. Teslyuk, “Modelling of Phonostatistical Structures of the Colloquial and Newspaper Styles in English Sonorant Phoneme Group”, *Computer Science and Information Technologies, CSIT*, pp. 67-70, 2017.
- [64] V. Pasichnyk, and T. Shestakevych, “The model of data analysis of the psychophysiological survey results,” *Advances in Intelligent Systems and Computing*, 512, pp. 271–281, 2017.

Methods of Building Intelligent Decision Support Systems Based on Adaptive Ontology

Vasyl Lytvyn
Information Systems and Network
Department
Lviv Polytechnic National University
Lviv, Ukraine
Vasyl.V.Lytvyn@lpnu.ua

Victoria Vysotska
Information Systems and Network
Department
Lviv Polytechnic National University
Lviv, Ukraine
Victoria.A.Vysotska@lpnu.ua

Dmytro Dosyn
Systems Analysis Laboratory
Karpenko Physico-Mechanical Institute
of the NAS of Ukraine
Lviv, Ukraine
Dmytro.Dosyn@gmail.com

Olga Lozynska
Information Systems and Network
Department
Lviv Polytechnic National University
Lviv, Ukraine
Olha.V.Lozynska@lpnu.ua

Oksana Oborska
Information Systems and Network
Department
Lviv Polytechnic National University
Lviv, Ukraine
oksana949@gmail.com

Abstract—The approach to the development of intelligent decision support systems using ontology knowledge bases consisting of such systems in the article is considered. The classification of such systems in terms of their operation based on ontologies is carried out. The mathematical functioning of intelligent decision support systems and intelligent search system based on ontology is developed. The notion of adaptive ontologies is introduced. An adaptive ontology is proposed to define as an ontology with concepts and relations weighted according to its importance for a given subject domain. This model of ontology defines not only explicit, but also implicit knowledge. The mathematical functioning of intelligent decision support systems using adaptive ontologies formalizes decision making of such a system. Semantic metric based on adaptive ontologies was built, unlike other metrics, takes into account the causal relation between the concepts, not just their taxonomy. For feature problems mathematical software based on automated determination of properties of set concepts was developed, according to the values which the process of decision support was carried out.

Keywords—adaptive ontologies; semantic metric; intelligent decision support systems.

I. INTRODUCTION

Scientific investigations of the using an ontology for information systems as intelligent search system (ICS), including intelligent decision support systems (IDSS) began at the end of the previous century and is rapidly developing. The basic theoretical principles of formal mathematical models of ontologies were developed in the works by T. Hrubera [1], who proposed to consider ontology as a three-dimensional tuple; in the works by N. Huarino [2]. The method of constructing ontologies and possible ways of development; G.Sova introduced the concept of conceptual graphs [3] and M. Montes-Gomez used it to present the ontologies [4]. The use of ontologies applied during the operation of information systems was described in the works of R. Knapp, A. Kelli, A. Galopina [5-6]; the problem of construction of intellectual systems as ICS based on ontologies was considered in the works by R. Fensli, J. Pettersen, A. Ugon, A. Galopin, M. Rospocher [7-11]. Analyzing the work as a whole, it can be concluded that

investigations of using an ontology in the construction of applied information systems as ICS is being carried out. These facts demonstrate the relevance of the problems of building IDSS based on ontologies as a subject of research.

II. METHODOLOGY

Formally the ontology consists of concepts (terms, concepts), organized in taxonomy, relations between concepts, as well as related axioms and rules of inference. Therefore, model of an ontology O is understood as:

$$O = \langle C, R, F \rangle, \quad (1)$$

where element of model C is a finite set of terms (concepts) software, which sets the ontology O ; $R: C \rightarrow C$ is a finite set of relations between concepts (terms, concepts) of the given software; F is a finite set of functions of interpretation (axiomatization, restrictions) defined on the concepts or ways of ontology O . Model Ontology (1) specifies only explicit knowledge. The theory of IDSS shows that its efficiency depends on both - explicitly and implicitly represented knowledge. So a model of an ontology must be developed to represent its internal knowledge formally so clear to be available to the user of the system and/or to the domain expert. In turn, this model will provide to IDSS a high quality of functioning because ontology is a core of IDSS knowledge base (KB). For the decision-making process a query language for ontology, such as SPARQL is used. However, ontologies consist of tens of thousands of concepts and connections that can not be memorized. An alternative to the query language is metric. It was suggested that the metrics based on ontologies should be built. Modern research of IDSS construction is underway in two ways: 1) IDSS classification (Case-Based Reasoning); 2) IDSS of planning activities (search for goals stated in the space of states). IDSS choice depends on the type of problem. The method for removing precedents is most effective for extensive KB, not based on a deductive theory; where the solutions are doubled in an each particular situation, and can be reused in similar cases; the goal of solving is - to get the right solution which is not guaranteed, but if is found will be the best possible.

The reasoning using a precedents looking approach is a method of IDSS design, which allows to solve a problem or recognize a situation by searching for a similarity among stored solutions in a form of classes in a database (DB). Formally given situation S belongs to $Class_k$ from the set of $\mathbf{Class} = \{Class_1, \dots, Class_N\}$, if between S and $Class_k$ ($k = \overline{1, N}$) there is lowest distance:

$$Class_k = \arg \min_i d(Class_i, S), \quad i = \overline{1, N}. \quad (2)$$

Normally IDSS activity should reach a goal state. The IDSS first step consists of developing a plan to achieve this state using all possible alternatives. The planning process is laid down on decomposition principle. This task ZP consists of three main components: a set of current states St , set of actions A , and a set of goal states $Goal$:

$$ZP = \langle St, A, Goal \rangle. \quad (3)$$

A performance and a condition should be assessed by the IDSS for effective planning. It can be seen that in any cases an IDSS requires metrics. In the first case, it is essential to evaluate the vicinity class, in the second to determine relevant conditions and actions. The method of construction of the metric depends on the efficiency of IDSS [12]. After analyzing the kinds of ontology-solving problems, it is concluded that they consist of two groups. The one of them has the feature of special concepts properties and their values. These include problems of diagnosing diseases, image recognition, classification and clustering using the machine learning techniques and more. Such problems can be called Features. For other types of problems concepts and their names rather than their meaning or statistics of their using are essential. These problems can be attributed to problem of texts clustering, AI search services, textual documents semantic parsing and automated annotation. Problems of this kind are called semantic. For the effective functioning of IDSS it is necessary to build the metrics on which it is possible to check the relevance of classes or grades. Design of this metrics is defined by the problem kind: whether it is symbolic or semantic. Efficiency of adaptation of ontology to the peculiarities of given domain determines its components and a process of self-learning during the operation. Effective solving the task of creating such mechanism is an automated weighing up of an ontology concepts and relations while KB population. This role is performed by the concepts and relations importance. Weight importance for the concept (communication) is a numerical measure which characterizes the certain concepts (communication) importance in particular software and changes dynamically according to certain rules while this IDSS operating. It was suggested that the ontology model expand ontology (1) by entering its formal concepts description and the weight ratios importance [13, 14]. This ontology is defined as formula:

$$\hat{O} = \langle \hat{C}, \hat{R}, F \rangle, \quad (4)$$

where value $\hat{C} = \langle C, W \rangle$, $\hat{R} = \langle R, L \rangle$, turn W is the C concepts importance weight, L is the R relations importance weight.

The defined in this way ontology is called adaptive, i.e. the one that adapts to software modification for using default of the concepts and relations importance weights between them. This ontology is uniquely defined as a weighted KB. Therefore, the metrics was built on such graphs. Advantages of model (4) over (1) are: 1) the ability to build metrics based on ontology; 2) the ability to adapt to the KB of IDSS user needs; 3) ability to specify the knowledge importance in expert software terms; 4) adaptive ontologies (AO) unlike conventional ontology reflects not only the explicit knowledge but also implicit (hidden); 5) data mining methods (k -nearest neighbors, Bayesian networks, decision trees) is a special AO case according to the setting weight rules for concepts and relations importance. In building KB terms for IDSS such an approach is obtained or expert system users are provided with ready DB, the core of which is the ontology, and their only task is to set up this KB by setting weights for its elements importance. IDSS functioning process for classification problems lies in the fact that some current situation belongs to $ZClass: S \rightarrow Class$. For doing this, the distance between the current situation and individual classes $d_i = d(S, Class_i)$ is found. The situation S belongs to the class to which the distance is the smallest. It is proposed to carry out a decision that corresponds to this class. Basically, methods classification are reduced to the induction of decision trees (DT) or to the nearest neighbor algorithm, supplemented with the software knowledge. As to the found solution adaptation and use, this problem is not formalized yet and is significantly dependent on software. It was suggested that AO be used for classification, i.e. to project classes and the current situation onto the software ontology; to enter within ontology IN the metrics by which to seek the necessary distance [12, 15].

For semantic tasks it was proposed to determine the distance between the class and the situation as the distance between the class and the current situation 'critical' concept. Since AO serves as a weighted KB, the concept is called the weights centre of the corresponding weighted center CG. If C_{class}^j is class weights center, and C_s^k is the current situation center weights, then the distance between this class and the current situation is defined as $d(Class, S) = d(C_{class}^j, C_s^k)$. From a mathematical point of view, the weights KB center is a concept by which the average distance to all other concepts is the smallest. Obviously, defined in this way the distance will depend on how we ask the distance between 2 adjacent peaks of KB. It was suggested that the distance between the peaks connected by communication be determined as

$$d_{ij} = \frac{Q}{L_{ij}(W_i + W_j)}, \quad (5)$$

where W_i and W_j are weight importance peaks C_i and C_j in accordance; L_{ij} is the communication importance weight between the peaks; Q is constant, that depends on ontology. Assumed if $L_{ii} = \infty$ then $d_{ii} = 0$. Later, the centers of appropriate weights of KB were found. The weights center is

the KB pinnacle, for which the average distance \bar{d}_i is the smallest: $\bar{d}_i^* = \min_i \bar{d}_i$. The average distance \bar{d}_i to top C_i is calculated using the formula:

$$\bar{d}_i = \frac{\sum_{j=1, j \neq i}^n d_{ij}^*}{n-1}, \quad (6)$$

where value n is the vertices number; d_{ij}^* is shortest path between vertices C_i and C_j , that is found the known algorithms using, such as Floyd-Warshall, Ford or Dijkstra. The proposed distance satisfies of the three axioms for metrics. Then this metrics for problems of classification within feature space was built. The set of classes $\mathbf{Class} = \{Class_1, Class_2, \dots, Class_N\}$ describes the characteristics (properties) $X = \{x_1, x_2, \dots, x_M\}$. D_i is domain properties x_i ; W_i are weight importance properties x_i class $Class_i$. The value of properties x_i denotes $z_i = z(x_i)$. So,

$$Class_i \leftrightarrow X_i = \{x_{i_1} = z_{i_1}, x_{i_2} = z_{i_2}, \dots, x_{i_k} = z_{i_k}\},$$

where $z_{i_j} \in D_{i_j}$. Then distance between $Class_i$ and the current situation S is defined as:

$$d_i = \sum_{i_j \in I_i} \varphi(z_{i_j}, z_{i_j}^S), \quad (7)$$

where z_{i_j} – is the property value x_{i_j} in $Class_i$; $z_{i_j}^S$ is the property value x_{i_j} of the current situation S ; \bar{I}_i is indexes sets of the major properties $Class_i$, $\bar{I}_i = \bar{I}_{i1} \cup \bar{I}_{i2} \cup \dots \cup \bar{I}_{iN_i}$, N_i is the properties number that should be considered to make a decision regarding the appropriateness S of $Class_i$,

$$\bar{I}_{i1} = \left\{ i_{s1} \mid i_{s1} = \arg \max_{i_j \in I_i} w_{i_j} \right\},$$

$$\bar{I}_{i2} = \left\{ i_{s2} \mid i_{s2} = \arg \max_{i_j \in I_i / I_{s1}} w_{i_j} \right\}.$$

The function $\varphi(\xi, \eta)$ can select any known metric (e.g., Lemming, Euclidean, Zhuravlev, Manhattan, etc.) within a single solution to the problem, depending on the data used (quantitative, qualitative, and mixed) [16].

For IDSS planning activities, the following model is described. The function $v(St(i))$ is state assessment of $St(i)$. Then a_{ij}^k is the transition from state $St(i)$ in a state $St(j)$, using alternative α_k ; $v(a_{ij}^k)$ is rating action a_{ij}^k .

Condition *Goal* determines that attributes subset X has to reach $z(x, Goal) \forall x \in X$.

Any state $St(i)$ is given a certain set of attributes Y_i , which gain value $z(y, St(i)) \forall y \in Y_i$. To assess the state $St(i)$ a display Ψ set of signs and their meanings of set state $St(i)$ to state attributes and values *Goal* by rules of ontology KB (SWRL), i.e. $\Psi : Y_i \xrightarrow{O} X$ should be made. Then assessment $v(St(i))$ is calculated

$$v(St(i)) = d(St(i), Goal) = \sum_{x \in X_W} \varphi(z(\Psi(x), St(i)), z(x, Goal))$$

where value X_W is attributes set with largest weights in AO, φ is the function of same as in (7). Obviously, when lower the assessment is, then better the condition is. Power set $|X_W|$ of user-defined system.

For IDSS actions select, the rationality of the behavior of the user is relied on, i.e. the effort to minimize the resources cost for the goal attainment. Every action a_{ij}^k is determined by the cost of resources g_{ij}^k (price move from state to state), where $k = 1, 2, \dots, n_i$. Value n_i is the number of alternatives α_k for the transition a_{ij} . For example, the problem of the pipelines of each alternative is characterized by cost resources and service life. Information about alternatives and resources is stored in this ontology. The signs and benefit importance from the transition to the state (its use dates, etc.) contained in the DB are learnt. Obviously, new alternatives may appear as IDSS replenishment module includes this ontology. Action assessment is directly proportional to resources consumption, i.e. $v(a_{ij}^k) = E \cdot g_{ij}^k$, where E is scalar value. In general, the decision on the selection of actions based alternatives is perform according to the formula: $o_i(a_{ij}^k) = \delta(v(a_{ij}^k), v(St(j)))$. After evaluating the actions and states the task of choosing the path in the state space is reduced to the dynamic programming problem:

$$St(j) = a(St(i), o_i), \Theta(St(0), \bar{o}) \Rightarrow \max(\min).$$

The solution as transition from the initial to the final state was found using methods suitable for solving such problems, [17-23]. For semantic problems of activities planning it is hard to say anything about *Goal* state [24-28]. For example, problem in text documents quasi-referencing as the ultimate goal is quasi-reference, but it can only be assumed what it should look like. The assessment of the state in this problem coincides with the semantic units importance assessment (sentences, token, word), depending on this problem. It was proposed to build metrics based on weight measure TF-IDF ontology software for such problems [29-37]. That is function value $v(St) = (TF-IDF) \cdot W$. This assessment has significant advantages over the others, because it is counted as both a frequency analysis for the terms use in the textual content (TF-IDF), and the specificity IN, to which the subject

matter of the text belongs [38-49]. The new status for problems quasi-referencing is to add new sentences to the quasi-reference [50-63].

III. RESULTS

IDSS consists of the following elements: DB, the core of which is JSC; DB, in which, depending on the type of problem, a set of classes and decisions relevant to them, weight of importance of concepts, types and weight ratios of importance, mentioned features, the history of the values of attributes (for tasks planning); control module solving (using built metrics depending on the task); replenishment module knowledge (builds, optimizes and teaches ontology) are kept [46-51]. To implement these components, such software is selected, to build ontologies - ontology editor Protégé OWL API; to record rules in KB - SWRL, which is included as a separate module Protégé; to build a database - a database management system MySQL; to build control module by

1. *The correlation between diffractometric investigations and calculations, based on the model of rigid spheres, allowed making prediction of the change of the surface tension and evaluating the steel wettability by extremum of a continuous function of structural melt factor. The influence of stainless steel elements of the laser doped into the surface on structural factors of melts Pb and Li Pb was investigated.*

2. *The damaging of power plant equipment, made of stainless austenitic steels is considered. It has been found that initiation of intergranular stress corrosion cracks in the weld region of the welded joints made of this steel is caused by interaction of 3 factors – the determined degree of basic metal sensitization, high service stress, that is higher than the material yield strength and the increased oxygen concentration in the heat carrier.*

The value of the weights of concepts and relations taken from the developed ontology are the basis of Material frequency method. Using equation (5), where $Q = 50$,

solving the problem, and knowledge replenishment module such programming languages as PHP, Python, Java, C #, depending on the purpose of IDSS are used. IDSS for semantic tasks such as ICS and IDSS of textual content classification are developed. Search engine is called intelligent methods if it carries out search based on the context. Such a system can be considered IDSS classification [52-63]. Indeed, the text unit, under which the searched (phrases, phrase, sentence, etc.) is the current situation, will be called standard. Text documents found in this way are classes. They are ranked according to the distance of the standard. For weight used when finding the distance, weight of ontology concepts, relating to the subject to which the standard belongs, is taken. The effectiveness of the functioning of the IRS will be shown on the analysis of abstract of scientific articles. Two abstracts of articles from the magazine “Physics and mechanics of materials” will be considered.

obtained weighted KG of these annotations, that shown in Fig. 1.

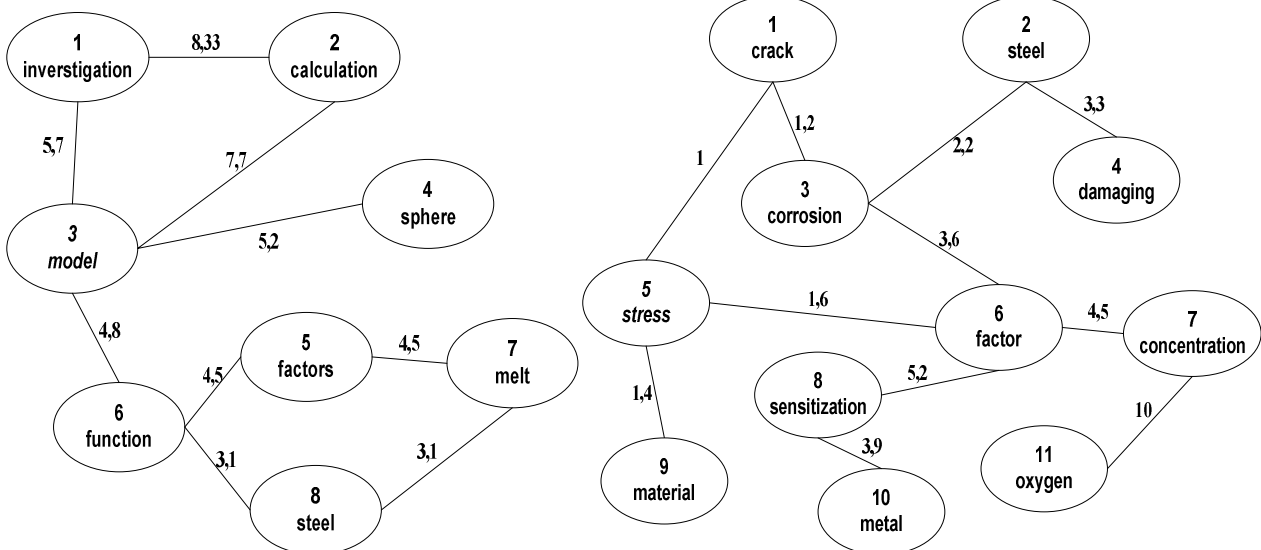


Fig. 1. Weighted conceptual graphs of the two annotations

The above concepts are their indexes. Using the algorithm and Floyd-Uorshalla formula (6), the centers of weights of the corresponding KG are: $C^1 = \{3\} = \{\text{'model'}\}$ for which $\bar{d}_3 = 7,37$ and element $C^2 = \{5\} = \{\text{'stress'}\}$ for value $\bar{d}_5 = 5,8$. A search is carried out for 'corrosion', i.e. this concept center is considered the weights centre and the found centers of annotations weights are the relevant centers

weights in these classes. Since the 'corrosion' is in the second annotation. The distance to this annotation is the distance between the 'corrosion' concept weights and annotations center: $d(\text{Pr}_2, S) = d(C^3, C^5) = 2,2$.

For the first annotation, the distance should be looked for by ontology. According to ontology of materials, the way from 'corrosion' concept to 'model' concept is the following: 'corrosion' → 'physical_process' → 'process' → 'model'.

Given the concepts and relations weight (the 1st 2 are hierarchical, the 3rd is functional), we get value $d(\text{Pr}_1, S) = 4, 6$. Similarly, the distance to other annotations from the 'corrosion' keyword is obtained.

This method is not the alternative to finding relevant information by keyword but its complement. If a keyword search does not provide the desired solution, the developed search-based contextual ontology software is used. The ontology defined scientific knowledge. That way a search only makes sense for scientific information. In this case, the search results for the 'corrosion' keyword would be only the 2nd abstract. The using developed method also offers users to see an article which meets the first annotation. The distance to the reduced first annotation is the smallest of all abstracts of the issue.

IV. CONCLUSION

The method for IDSS construction based on ontological approach was developed. And the efficiency of such systems through the using mathematical and the developed software, which is based on the using ontologies in these systems, ontologies adaptation to the specific problems in ontology domain is achieved. Traditional ontology structure by introducing into their structure the concepts and relations importance weights was modified. This was made possible by setting up these measures, adapted to the specific problems in ontology domain and the system's user needs. This model for ontologies defines not only explicit, but also implicit knowledge. The mathematical model of IDSS functioning is based on AO, which helped to formalize such a system decision making. Semantic metric based on AO was built, unlike other metrics, takes into account the causal relation between the concepts, not just their taxonomy. Mathematical software for feature problems based on automated determination of concepts properties set is developed, according to the values which the decision support process is carried out. The software operation of IDSS, based on the constructed models, methods and algorithms, makes it possible to implement individual components and functional modules of IDSS, the core KB of which is ontology.

REFERENCES

- [1] T. Gruber, "A translation approach to portable ontologies," *Knowledge Acquisition*, vol. 5, pp. 199-220, 1993.
- [2] N. Guarino, "Formal Ontology, Conceptual Analysis and Knowledge Representation," *Human-Computer Studies*, Vol. 43, pp. 625-640, 1995.
- [3] J. Sowa, "Conceptual Graphs as a universal knowledge representation. Semantic Networks in Artificial Intelligence," *Computers & Mathematics with Applications*, part 1, pp. 75-95, 1992.
- [4] M. Montes-y-Gómez, A. Gelbukh, and A. López-López, Comparison of Conceptual Graphs, <http://ccc.inaoep.mx/~mmontes/publicaciones/2000/ComparisonCG>.
- [5] H. Bulskov, R. Knappe, and R. Andreasen, On Querying Ontologies and Databases. *FQAS*, 2004, pp. 191-202.
- [6] A. Calli, G. Gottlob, "A. Pieris, Advanced processing for ontological queries," *Very Large Databases*, pp. 554-565, 2010.
- [7] Rune Fensli, and Jan Pettersen Nytnun, An Ontology-Based Decision Support System for Interventions based on Monitoring Medical Conditions on Patients in Hospital Wards. *Information and Communication Technology*, Spring, 2014.
- [8] A. Ugon, K. Sedki, A. Kotti, B. Seroussi, C. Philippe, J. Ganascia, P. Garda, J. Bouaud, and A. Pinna, "Decision System Integrating Preferences to Support Sleep Staging," *Studies in health technology and informatics*, vol. 228, pp. 514-518, 2016.
- [9] A. Galopin, J. Bouaud, S. Pereira, and B. Seroussi, "An Ontology-Based Clinical Decision Support System for the Management of Patients with Multiple Chronic Disorders," *Studies in health technology and informatics*, vol. 216, pp. 275-279, 2015.
- [10] M. Rospocher, and L. Serafini, "An Ontological Framework for Decision Support Joint," *International Semantic Technology Conference*, pp. 239-254, 2012.
- [11] M. Rospocher, and L. Serafini, "Ontology-centric decision support. Proceedings," *Semantic Technologies Meet Recommender Systems & Big Data*, vol. 919, pp. 61-72, 2012.
- [12] V. Lytvyn, V. Vysotska, I. Peleshchak, I. Rishnyak, and R. Peleshchak, "Time Dependence of the Output Signal Morphology for Nonlinear Oscillator Neuron Based on Van der Pol Model," *International Journal of Intelligent Systems and Applications (IJISA)*, vol. 10, no.4, pp. 8-17, 2018.
- [13] R. Tkachenko, P. Tkachenko, I. Izonin, Y. Tsymbal, "Learning-based image scaling using neural-like structure of geometric transformation paradigm," In *Studies in Computational Intelligence*, vol. 730, Springer Verlag, pp. 537-565, 2018.
- [14] Y. Rashkevych, D. Peleshko, O. Vynokurova, I. Izonin and N. Lotoshynska, "Single-frame image super-resolution based on singular square matrix operator," 2017 IEEE First Ukraine Conference on Electrical and Computer Engineering (UKRCON), pp. 944-948, 2017.
- [15] V. Lytvyn, V. Vysotska, P. Pukach, M. Vovk, and D. Ugryn, "Method of functioning of intelligent agents, designed to solve action planning problems based on ontological approach," *Eastern-European Journal of Enterprise Technologies*, vol. 3/2(87), pp. 11-17, 2017.
- [16] V. Lytvyn, V. Vysotska, O. Veres, I. Rishnyak, and H. Rishnyak, "The Risk Management Modelling in Multi Project Environment," *Computer Science and Information Technologies*, pp. 32-35, 2017.
- [17] V. Lytvyn, V. Vysotska, O. Veres, I. Rishnyak, and H. Rishnyak. "Content linguistic analysis methods for textual documents classification," *Computer Science and Information Technologies, CSIT'2016*, pp. 190-192, 2016.
- [18] I. Khomytska, and V. Teslyuk, Modelling of Phonostatistical Structures of English Backlingual Phoneme Group in Style System," *Proceedings of the 14th International Conference, CADMS, Polyana, Ukraine*, pp. 324-327, 2017.
- [19] I. Khomytska, V. Teslyuk, "Modelling of Phonostatistical Structures of English Backlingual Phoneme Group in Style System", *Proceedings of CADMS*, pp. 324-327, 2017.
- [20] I. Khomytska, V. Teslyuk, "Modelling of Phonostatistical Structures of the Colloquial and Newspaper Styles in English Sonorant Phoneme Group", *Computer Science and Information Technologies, CSIT*, pp. 67-70, 2017.
- [21] I. Khomytska, and V. Teslyuk, "Modelling of Phonostatistical Structures of the Colloquial and Newspaper Styles in English Sonorant Phoneme Group," *XIIth Scientific and Technical Conference, CSIT, Lviv*, pp.67-70, 2017.
- [22] V. Lytvyn, V. Vysotska, P. Pukach, I. Bobyk, and D. Uhryn, "Development of a method for the recognition of author's style in the Ukrainian language texts based on linguometry, stylemetry and glottochronology," *Eastern-European Journal of Enterprise Technologies*, 88, pp. 10-18, 2017.
- [23] V. Lytvyn, V. Vysotska, P. Pukach, O. Brodyak, and D. Ugryn, "Development of a method for determining the keywords in the slavic language texts based on the technology of web mining," *Eastern-European Journal of Enterprise Technologies*, Vol. 2/2(86), 2017, pp. 4-12.
- [24] T. Basyuk, "The main reasons of attendance falling of internet resource," *Computer Science and Information Technologies*, pp. 91-93, 2015.
- [25] P. Zhezhnych, and O. Markiv, "Linguistic Comparison Quality Evaluation of Web-Site Content with Tourism Documentation Objects," *Advances in Intelligent Systems and Computing*, Vol. 689, pp. 656-667, 2018.
- [26] K. Mykich, and Y. Burov, "Algebraic model for knowledge representation in situational awareness systems," *Int. Conf. Computer Sciences and Information Technologies*, pp.165-167, 2016.
- [27] K. Mykich, and Y. Burov, "Uncertainty in situational awareness systems," *Int. Conf. Modern Problems of Radio Engineering, Telecommunications and Computer Science*, pp. 729-732, 2016.
- [28] K. Mykich, and Y. Burov, "Algebraic Framework for Knowledge Processing in Systems with Situational Awareness," *Advances in Intelligent Systems and Computing*, Springer Verlag, pp. 217-228.

- [29] K. Mykich, Y. Burov, Research of uncertainties in situational awareness systems and methods of their processing, *Eastern European Journal of Enterprise Technologies*, vol. 1(79), pp.19-26, 2016.
- [30] O. Kanishcheva, V. Vysotska, L. Chyrun, and A. Gozhij, "Method of Integration and Content Management of the Information Resources Network," *Advances in Intelligent Systems and Computing*, 689, Springer, pp. 204–216, 2017.
- [31] J. Su, V. Vysotska, A. Sachenko, V. Lytvyn, and Y. Burov, "Information resources processing using linguistic analysis of textual content," *Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications 9th IEEE Internat. Conf.*, Bucharest, Romania, pp. 573–578, 2017.
- [32] M. Korobchinsky, V. Vysotska, L. Chyrun, and L. Chyrun, "Peculiarities of Content Forming and Analysis in Internet Newspaper Covering Music News," *Proceedings of the Scientific and Technical Conference, CSIT, Computer Science and Information Technologies*, pp. 52–57, 2017.
- [33] O. Naum, L. Chyrun, O. Kanishcheva, and V. Vysotska. "Intellectual System Design for Content Formation." *Proceedings of the Scientific and Technical Conference, CSIT, Computer Science and Information Technologies*, pp. 131–138, 2017.
- [34] V. Lytvyn, V. Vysotska, Y. Burov, O. Veres, and I. Rishnyak, "The Contextual Search Method Based on Domain Thesaurus," *Advances in Intelligent Systems and Computing*, 689, pp. 310–319, 2017.
- [35] M. Davydov, and O. Lozynska, "Information System for Translation into Ukrainian Sign Language on Mobile Devices," *Int. Conf. Computer Science and Information Technologies*, pp. 48-51, 2017.
- [36] M. Davydov, and O. Lozynska, "Linguistic Models of Assistive Computer Technologies for Cognition and Communication," *Computer Science and Information Technologies*, 171-175, 2017.
- [37] M. Davydov, and O. Lozynska, "Mathematical Method of Translation into Ukrainian Sign Language Based on Ontologies," *Advances in Intelligent Systems and Computing*, vol. 689, pp. 89-100, 2017.
- [38] P. Kravets. "The control agent with fuzzy logic," *Perspective Technologies and Methods in MEMS Design, MEMSTECH'2010*, pp. 40-41, 2010.
- [39] B. Mobasher, "Data mining for web personalization," *The adaptive web*, Springer, 2007, pp. 90–135.
- [40] C.E. Dinucă, D. Ciobanu. "Web Content Mining." University of Petroșani, Economics, pp.85-92, 2012.
- [41] G. Xu, Y. Zhang, and L. Li. "Web content mining." *Web Mining and Social Networking*, Springer, pp. 71–87, 2011.
- [42] Jivani Anjali Ganesh, "A Comparative Study of Stemming Algorithms," *Int. J. Comp. Tech. Appl.*, vol. 6, pp. 1930–1938, 2011.
- [43] A. Mishler, E.S. Crabb, S. Paletz, B. Hefright, and E. Golonka. "Using Structural Topic Modeling to Detect Events and Cluster Twitter Users in the Ukrainian Crisis," *Communications in Computer and Information Science*, vol. 528, pp. 639–644, 2015.
- [44] K. Kowalska, Cai Di, and S. Wade, "Sentiment analysis of polish texts," *Computer and Communication Engineering*, pp. 39-41, 2012.
- [45] N. Kotsyba, "The current state of work on the Polish-Ukrainian Parallel Corpus," *Organization and Development of Digital Lexical Resources*, pp. 55–60, 2009.
- [46] V. Vysotska, "Linguistic Analysis of Textual Commercial Content for Information Resources Processing," *Modern Problems of Radio Engineering, Telecommunications and Computer Science*, Proc. of the Int. Conf. TCSET, 2016, pp. 709–713, 2016.
- [47] V. Vysotska, L. Chyrun, and L. Chyrun, "Information Technology of Processing Information Resources in Electronic Content Commerce Systems," *Computer Science and Information Technologies, XI-th Int. Conf. CSIT*, pp. 212–222, 2016.
- [48] V. Vysotska, L. Chyrun, and L. Chyrun, "The Commercial Content Digest Formation and Distributional Process," *Computer Science and Information Technologies, XI-th Int. Conf. CSIT*, pp. 186–189, 2016.
- [49] V. Lytvyn, I. Bobyk, and V. Vysotska, "Application of algorithmic algebra system for grammatical analysis of symbolic computation expressions of propositional logic," *Radio Electronics Computer Science Control*, vol. 4(39), pp. 77–89, 2016.
- [50] Vasyly Lytvyn, Victoria Vysotska, Dmytro Dosyn, Roman Holoschuk, and Zoriana Rybchak, "Application of Sentence Parsing for Determining Keywords in Ukrainian Texts," *Computer Science and Information Technologies*, pp. 326–331, 2017.
- [51] V. Lytvyn, V. Vysotska, L. Chyrun, and L. Chyrun, "Distance Learning Method for Modern Youth Promotion and Involvement in Independent Scientific Researches," *Proc. of the IEEE First Int. Conf. on Data Stream Mining & Processing (DSMP)*, pp. 269-274, 2016.
- [52] V. Lytvyn, V. Vysotska, D. Uhryn, M. Hrendus, O. Naum, "Analysis of statistical methods for stable combinations determination of keywords identification," *Eastern-European Journal of Enterprise Technologies*, vol. 2/2(92), pp. 23-37, 2018.
- [53] O. Chernukha, and Y. Bilushchak, "Mathematical modeling of random concentration field and its second moments in a semispace with erlangian distribution of layered inclusions", *Task Quarterly*, vol. 20(3), pp. 295-334, 2016.
- [54] B. Rusyn, O. Lutsyk, O. Lysak, A. Lukeniuk, and L. Pohreliuk, "Lossless Image Compression in the Remote Sensing Applications", *Data stream mining & processing, DSMP*, pp.195-198, 2016.
- [55] Kowalik Dagmara, "Polish vocational competence standards for the needs of adult education and the European labour market", *International Conference on Advanced Information Engineering and Education Science, ICAEES*, pp. 95-98, 2013.
- [56] P. Kravets, "The game method for orthonormal systems construction", *The Experience of Designing and Application of CAD Systems in Microelectronics, CADSM 2007*.
- [57] P. Kravets, "Game Model of Dragonfly Animat Self-Learning", *Perspective Technologies and Methods in MEMS Design*, pp. 195-201, 2016.
- [58] V. Pasichnyk, and T. Shestakevych, "The model of data analysis of the psychophysiological survey results," *Advances in Intelligent Systems and Computing*, 512, pp. 271–281, 2017.
- [59] O. Maksymiv, T. Rak, and D. Peleshko, "Video-based Flame Detection using LBP-based Descriptor: Influences of Classifiers Variety on Detection Efficiency," *IJISA*, vol. 9(2), pp. 42–48, 2017.
- [60] D. Peleshko, T. Rak, and I. Izonin, "Image Superresolution via Divergence Matrix and Automatic Detection of Crossover," *IJISA*, vol. 8(12), pp. 1–8, 2016.
- [61] O. Bazylyk, P. Taradaha, O. Nadobko, L. Chyrun, and T. Shestakevych, "The results of software complex OPTAN use for modeling and optimization of standard engineering processes of printed circuit boards manufacturing." *TCSET'2012*, pp. 107–108, 2012.
- [62] A. Bondariev, M. Kiselychuk, O. Nadobko, L. Nedostup, L. Chyrun, and T. Shestakevych, "The software complex development for modeling and optimizing of processes of radio-engineering equipment quality providing at the stage of manufacture," *Proceedings of International Conference TCSET'2012*, pp. 159, 2012.
- [63] V. Teslyuk, V. Beregovskiy, P. Denysyuk, T. Teslyuk, and A. Lozynskiy, "Development and Implementation of the Technical Accident Prevention Subsystem for the Smart Home System," *International Journal of Intelligent Systems and Applications*, vol. 10(1), pp. 1–8, 2018.

About Kernel Structure Construction of the Generalized Neural Functions

Fedir Geche

Department of Cybernetics and Applied Mathematics
Uzhhorod National University
Uzhhorod, Ukraine
fgeche@hotmail.com

Oksana Mulesa

Department of Cybernetics and Applied Mathematics
Uzhhorod National University
Uzhhorod, Ukraine
Oksana.Mulesa@uzhnu.edu.ua,

Anatoliy Batyuk

ACS Department
Lviv Polytechnic National University
abatyuk@gmail.com

Veronika Voloshchuk

Uzhhorod National University
Uzhhorod, Ukraine
veronika.smolanka@uzhnu.edu.ua

Abstract— The paper introduces concept of a modified kernel of the Boolean functions. Applying such a concept, the criteria for the implementation of the Boolean functions by one generalized neural element are obtained. The effective and necessary conditions to check whether the algebra of logic functions belong to the class of the generalized neural functions are given. A sufficient condition for the implementation of the Boolean functions is obtained by one generalized neural element on the basis of which it is possible to develop effective methods for the synthesis of the generalized integer neural elements with a large number of inputs.

Keywords— spectrum of function, generalized neural element, structure vector, kernel of function, convex hull, character of group, synthesis, metric, matrix of tolerance.

I. INTRODUCTION

In recent years, there has been an increased interest in neural-like structures, which are widely used in image recognition, compression of discrete signals and images, forecasting, business, medicine and engineering.

Widespread usage of neural networks to effectively solve applied problems will become possible if effective methods of synthesis (training) of neural elements with different functions of activation and synthesis of logical circuits of them are developed.

It is required to synthesize reliable (integer) neural elements with a large number of inputs concerning such issues as image recognition, compression and transmission of discrete signals. Classical methods of approximation and various iterative methods for the synthesis of neural elements are almost not suitable for finding the vectors of neural element structures to implement discrete functions with a large number of inputs (several hundreds, thousands).

Artificial neural networks and neural-like structures are effectively used for the classification and recognition of images [1] and for improving their quality [2]. Intelligent blocks of various systems for controlling chemical processes [3], for prediction of economic [4], biological [5] processes are developed on their basis. Neural network methods have been successfully applied to compress signals and images [6-9], in the banking sector to assess credit risk [9,11], in automated control systems of technological processes [12], in the field of intellectual data processing [13] and for

constructing the logical blocks of various safety systems [14].

When selecting mathematical models of neural elements for the construction of neural-like structures, the functionality of these elements is of importance. It is crucial for optimizing the number of elements in the corresponding logical structures. The generalized neural elements allowing to extend the functionality of ordinary neural elements with threshold activation functions are considered. The properties of the kernels of functions of the algebra of logic that are implemented on these elements are discovered.

II. THE KERNEL PROPERTIES OF THE GENERALIZED NEURAL FUNCTIONS

Let $H_2 = \{-1, 1\}$ – a cyclic group of the 2nd order, $G_n = H_2 \otimes \dots \otimes H_2$ – a direct product n of the cyclic groups H_2 i $\chi(G_n)$ – a group of characters [15] of the group G_n over the field of R . Define on the set $R \setminus \{0\}$ a function as follows:

$$\text{Rsign}x = \begin{cases} 1, & \text{if } x > 0, \\ -1, & \text{if } x < 0. \end{cases} \quad (1)$$

Let $Z_2 = \{0, 1\}$, $i \in \{0, 1, 2, \dots, 2^n - 1\}$ and (i_1, \dots, i_n) – binary code of i , i.e. $i = i_1 2^{n-1} + i_2 2^{n-2} + \dots + i_n$, $i_j \in \{0, 1\}$. The value of the character χ_i on the element $\mathbf{g} = ((-1)^{\alpha_1}, \dots, (-1)^{\alpha_n}) \in G_n$ $((\alpha_1, \dots, \alpha_n) \in Z_2^n$ – n -fold Cartesian product $Z_2 = \{0, 1\}$) is determined as follows:

$$\chi_i(\mathbf{g}) = (-1)^{\alpha_1 i_1 + \alpha_2 i_2 + \dots + \alpha_n i_n}. \quad (2)$$

Considering the orthogonality of the characters [15] the group of characters $X(G_n)$ forms an orthogonal basis of the space $V_R = \{\phi | \phi: G_n \rightarrow R\}$. Since the Boolean function of n variables in $\{-1, 1\}$ sets unambiguous mapping of the

form $f : G_n \rightarrow H_2$, to $f \in V_R$ and it means that an arbitrary Boolean function f can be written unambiguously:

$$f(\mathbf{g}) = s_0 \chi_0(\mathbf{g}) + s_1 \chi_1(\mathbf{g}) + \dots + s_{2^n-1} \chi_{2^n-1}(\mathbf{g}). \quad (3)$$

A vector $\mathbf{s}_f = (s_0, s_1, \dots, s_{2^n-1})$ is called the spectrum of the Boolean function f in the system of characters $\chi(G_n)$ (in the system of Walsh-Hadamard basic functions [16]).

With different characters $\chi(G_n)$, in addition to the main one, make up m -element set $\chi = \{\chi_{i_1}, \dots, \chi_{i_m}\}$ and concerning the chosen system of characters, according to [17], consider the mathematical model of a neural element with a generalized threshold activation function (of a generalized neural element):

$$f(x_1(\mathbf{g}), \dots, x_n(\mathbf{g})) = \text{Rsign}\left(\sum_{j=1}^m \omega_j \chi_{i_j}(\mathbf{g}) + \omega_0\right), \quad (4)$$

where $\mathbf{w} = (\omega_1, \dots, \omega_m; \omega_0)$ is called a vector of the structure of the generalized neural element (GNE) regarding the system of characters $\chi \in \chi(G_n)$ i $\mathbf{g} \in G_n$.

Let $\mathbf{w}(\mathbf{g}) = \omega_1 \chi_{i_1}(\mathbf{g}) + \dots + \omega_m \chi_{i_m}(\mathbf{g}) + \omega_0$. If $\mathbf{w} = (\omega_1, \dots, \omega_m; \omega_0)$ is a vector of the GNE structure regarding the system of characters $\chi = \{\chi_{i_1}, \dots, \chi_{i_m}\}$ of the group $\chi(G_n)$, that realizes the Boolean function $f : G_n \rightarrow H_2$, from (1) and (4) we have the following result:

$$\forall \mathbf{g} \in G_n \quad \mathbf{w}(\mathbf{g}) \neq 0. \quad (5)$$

Further we only consider such neural elements whose vectors of structure satisfy the condition (5). A set of all such $m+1$ -dimensional real vectors satisfying the condition (5) regarding the system χ we will denote as $W_{m+1}(\chi) = W_{m+1}(\chi_{i_1}, \dots, \chi_{i_m})$.

Definition 1. The Boolean function $f : G_n \rightarrow H_2$ is called a generalized neural function concerning the system of characters $\chi = \{\chi_{i_1}, \dots, \chi_{i_m}\} \subset \chi(G_n)$ if $\mathbf{w} = (\omega_1, \dots, \omega_m; \omega_0) \in W_{m+1}(\chi)$ that is applied to as an equation (4)

To introduce the concept of kernels of the generalized neural functions and to study their basic properties, the Boolean functions will be considered as in $H_2 = \{-1, 1\}$, as in $Z_2 = \{0, 1\}$. Let $f(x_1, \dots, x_n)$ the Boolean function in $\{-1, 1\}$, i.e. $f : G_n \rightarrow H_2$. We will consider the problem on the implementation of the Boolean function $f(x_1, \dots, x_n)$ by one GNE regarding the system of characters $\chi = \{\chi_{i_1}, \chi_{i_2}, \dots, \chi_{i_m}\} \subset X(G_n)$ in $\{0, 1\}$. Using the

transformation $\mathbf{x}' = \frac{1}{2}(\mathbf{x}+1)$ we define the mapping of the form $\{-1, 1\} \rightarrow \{0, 1\}$ and build the system $\chi' = \left\{ \chi'_{i_1} = \frac{1}{2}(\chi_{i_1} + 1), \chi'_{i_2} = \frac{1}{2}(\chi_{i_2} + 1), \dots, \chi'_{i_m} = \frac{1}{2}(\chi_{i_m} + 1) \right\}$

Let $f^{-1}(-1) = \{\mathbf{g} \in G_n \mid f(\mathbf{g}) = -1\}$ and $f^{-1}(1) = \{\mathbf{g} \in G_n \mid f(\mathbf{g}) = 1\}$. Using the system χ' we will determine:

$$f_{\chi'}^{-1}(0) = \bigcup_{\mathbf{g} \in f^{-1}(-1)} \left\{ \chi'_{i_1}(\mathbf{g}), \dots, \chi'_{i_m}(\mathbf{g}) \right\},$$

$$f_{\chi'}^{-1}(1) = \bigcup_{\mathbf{g} \in f^{-1}(1)} \left\{ \chi'_{i_1}(\mathbf{g}), \dots, \chi'_{i_m}(\mathbf{g}) \right\}.$$

Definition 2. The kernel of the Boolean function $f : G_n \rightarrow H_2$ regarding the system of characters $\chi = \{\chi_{i_1}, \chi_{i_2}, \dots, \chi_{i_m}\} \subset X(G_n)$ in $\{0, 1\}$ is called a set $K(f_{\chi})$, which is determined as follows:

$$K(f_{\chi}) = \begin{cases} f_{\chi}^{-1}(1), & \text{if } |f_{\chi}^{-1}(1)| \leq |f_{\chi}^{-1}(0)|, \\ f_{\chi}^{-1}(0), & \text{if } |f_{\chi}^{-1}(1)| > |f_{\chi}^{-1}(0)|, \end{cases}$$

if $f_{\chi}^{-1}(1) \cap f_{\chi}^{-1}(0) = \emptyset$, where $|f_{\chi}^{-1}(i)|$ – a number of set elements $f_{\chi}^{-1}(i)$ ($i \in \{0, 1\}$).

If $f_{\chi}^{-1}(1) \cap f_{\chi}^{-1}(0) \neq \emptyset$, the kernel $K(f_{\chi})$ does not exist and it means that a function f is not realized by one GNE regarding the system χ .

Let Z_2^m – m th Cartesian degree of the set $Z_2 = \{0, 1\}$. Assume, that the function $f : G_n \rightarrow H_2$ has the kernel $K(f_{\chi})$ regarding the system $\chi = \{\chi_{i_1}, \dots, \chi_{i_m}\}$, i.e. $f_{\chi}^{-1}(1) \cap f_{\chi}^{-1}(0) = \emptyset$. The sets $f_{\chi}^{-1}(1)$, $f_{\chi}^{-1}(0)$ satisfy one of the conditions:

1) $Z_2^m = f_{\chi}^{-1}(1) \cup f_{\chi}^{-1}(0)$ – a function f_{χ} completely defined on the set Z_2^m ;

2) $Z_2^m \neq f_{\chi}^{-1}(1) \cup f_{\chi}^{-1}(0)$ – a function f_{χ} partially defined on the set Z_2^m .

In the first case, the kernel $K(f_{\chi})$ is defined unambiguously and $K(f_{\chi}, \emptyset) = K(f_{\chi})$.

In the second case, we introduce the concept of the generalized kernel concerning the system of characters χ .

Let $K(f_{\chi}) = \{\mathbf{a}_1, \dots, \mathbf{a}_q\}$ – a kernel of the Boolean function f_{χ} regarding the system $\chi = \{\chi_{i_1}, \dots, \chi_{i_m}\}$ and

$f_{\chi}^{-1}(\ast)$ is a set of those combinations with Z_2^m on which the function is not defined, then under the extended kernel function f_{χ} regarding the system χ we imply $K(f_{\chi}, A) = \{\mathbf{a}_1, \dots, \mathbf{a}_q, \mathbf{a}_{q+1}, \dots, \mathbf{a}_{q+s}\}$, where $\mathbf{a}_{q+1}, \dots, \mathbf{a}_{q+s}$ – various arbitrary elements of the set from $f_{\chi}^{-1}(\ast)$, $q + s \leq 2^{m-1}$ and $A = \{\mathbf{a}_{q+1}, \dots, \mathbf{a}_{q+s}\}$. Note, that a set A may be empty.

We introduce a concept of the modified kernel $K(f_{\chi}, M)$ of the Boolean function $f : G_n \rightarrow H_2$ concerning the system of characters $\chi = \{\chi_{i_1}, \chi_{i_2}, \dots, \chi_{i_m}\} \subset X(G_n)$ as follows:

$$K(f_{\chi}, M) = \begin{cases} K(f_{\chi}), & \text{if } A = \emptyset; \\ K(f_{\chi}, A), & \text{if } A \neq \emptyset. \end{cases}$$

Definition 3. The Boolean function $f : G_n \rightarrow H_2$ is realized by one generalized neural element with the structure vector $\mathbf{w} = (\omega_1, \dots, \omega_m; \omega_0)$ over R concerning the system of characters $\chi = \{\chi_{i_1}, \chi_{i_2}, \dots, \chi_{i_m}\} \subset X(G_n)$ in $\{0, 1\}$, if there exists such a modified kernel $K(f_{\chi}, M)$. Let $f : G_n \rightarrow H_2$, the system $\chi = \{\chi_{i_1}, \chi_{i_2}, \dots, \chi_{i_m}\} \subset X(G_n)$ and the modified kernel $K(f_{\chi}, M)$. We will make for a function f concerning the system χ and the modified kernel $K(f_{\chi}, M)$ the Boolean function $f_{\chi}^* : Z_2^m \rightarrow Z_2$ as follows:

$$\forall \mathbf{a} \in K(f_{\chi}, M) \quad f_{\chi}^*(\mathbf{a}) = f_{\chi}(\mathbf{a}),$$

$$\forall \mathbf{a} \in Z_2^m \setminus K(f_{\chi}, M) \quad f_{\chi}^*(\mathbf{a}) = \overline{f_{\chi}(\mathbf{a})},$$

where the vinculum means a logical denial operation.

By definition, function f_{χ}^* and $K(f_{\chi}) \subset K(f_{\chi}, M)$ we obtain

$$f_{\chi}^{-1}(\alpha) \subset f_{\chi}^{*-1}(\alpha),$$

where $\alpha \in \{0, 1\}$.

We define the kernel of function f_{χ}^* as a modified kernel of function f_{χ} , i.e.

$$K(f_{\chi}^*) = K(f_{\chi}, M).$$

We will determine a convex linear hull $\text{conv}K(f_{\chi}^*)$ of the kernel elements $K(f_{\chi}^*)$ as follows:

$$\text{conv}K(f_{\chi}^*) = \left\{ \mathbf{x} \in [0, 1]^m \mid \mathbf{x} = \sum_{i=1}^l \lambda_i \mathbf{a}_i, \sum_{i=1}^l \lambda_i = 1, \lambda_i \geq 0, \dots, \lambda_l \geq 0; \mathbf{a}_1, \dots, \mathbf{a}_l \in K(f_{\chi}^*) \right\}.$$

Similarly, we define $\text{conv}K(f_{\chi}^*)^*$ for $K(f_{\chi}^*)^* = Z_2^m \setminus K(f_{\chi}^*)$.

Theorem 1. The Boolean function $f : G_n \rightarrow H_2$ ($f \neq \text{const}$) is realized by one generalized neural element regarding the system of characters $\chi = \{\chi_{i_1}, \chi_{i_2}, \dots, \chi_{i_m}\} \subset X(G_n)$ only when there exists such a modified kernel $K(f_{\chi}, M)$, that

$$\text{conv}K(f_{\chi}^*) \cap \text{conv}K(f_{\chi}^*)^* = \emptyset.$$

Let \mathbf{a}, \mathbf{b} – arbitrary kernel elements $K(f_{\chi})$ ($\mathbf{a} \neq \mathbf{b}$) of the Boolean function $f : G_n \rightarrow H_2$ regarding the system of characters $\chi = \{\chi_{i_1}, \chi_{i_2}, \dots, \chi_{i_m}\} \subset X(G_n)$ i $O(\mathbf{a}, \mathbf{b})$ – a set of such ortho-vectors $\mathbf{e}_{i_1}, \dots, \mathbf{e}_{i_s}$, that $\mathbf{a} \oplus \mathbf{b} = \mathbf{e}_{i_1} + \mathbf{e}_{i_2} + \dots + \mathbf{e}_{i_s}$, where \oplus – a coordinate-wise sum of vectors by module 2, $i_r \neq i_k$, if $r \neq k$. Denote $H(\mathbf{a}, \mathbf{b})$ to be the subset of the group Z_2^m (Z_2^m forms a group regarding the operation \oplus), which is generated by the elements $O(\mathbf{a}, \mathbf{b})$, i.e.

$$H(\mathbf{a}, \mathbf{b}) = \langle \mathbf{e}_{i_1}, \dots, \mathbf{e}_{i_s} \mid \mathbf{e}_{i_1}, \dots, \mathbf{e}_{i_s} \in O(\mathbf{a}, \mathbf{b}) \rangle.$$

Let $\mathbf{a} = (\alpha_1, \dots, \alpha_m)$, $\mathbf{b} = (\beta_1, \dots, \beta_m) \in Z_2^m$. A coordinate-wise conjunction of vectors \mathbf{a} and \mathbf{b} we will define as $\mathbf{a} \& \mathbf{b} = (\alpha_1 \& \beta_1, \dots, \alpha_m \& \beta_m)$ as well as $H(\mathbf{a} \& \mathbf{b})$ we will define an adjacent group class Z_2^m by the subset $H(\mathbf{a}, \mathbf{b})$, that is defined by the element $\mathbf{a} \& \mathbf{b}$, i.e. $H(\mathbf{a} \& \mathbf{b}) = \mathbf{a} \& \mathbf{b} \oplus H(\mathbf{a}, \mathbf{b})$.

Set the metric $\rho(\mathbf{a}, \mathbf{b})$ on Z_2^m as follows:

$$\rho(\mathbf{a}, \mathbf{b}) = \sum_{i=1}^m (\alpha_i \oplus \beta_i),$$

where $\mathbf{a} = (\alpha_1, \dots, \alpha_m)$, $\mathbf{b} = (\beta_1, \dots, \beta_m) \in Z_2^m$.

Theorem 2. If the Boolean function $f : G_n \rightarrow H_2$ ($f \neq \text{const}$) is realized by one generalized neural element regarding the system of characters $\chi = \{\chi_{i_1}, \chi_{i_2}, \dots, \chi_{i_m}\} \subset X(G_n)$ with the modified kernel $K(f_{\chi}, M)$, for any two different elements \mathbf{a}, \mathbf{b} with $K(f_{\chi}^*)$, for which $|H(\mathbf{a} \& \mathbf{b}) \cap K(f_{\chi}^*)^*| \geq 2$ and for any two different elements \mathbf{g}, \mathbf{h} with $H(\mathbf{a} \& \mathbf{b}) \cap K(f_{\chi}^*)^*$, inequality is realized $\rho(\mathbf{g}, \mathbf{h}) < \rho(\mathbf{a}, \mathbf{b})$.

Let $K(f_\chi^*) = \{\mathbf{a}_1, \dots, \mathbf{a}_t\}$ – a kernel of the Boolean function $f: G_n \rightarrow H_2$ ($f \neq const$) regarding the system of characters $\chi = \{\chi_{i_1}, \chi_{i_2}, \dots, \chi_{i_m}\} \subset X(G_n)$ with the modified kernel $K(f_\chi, M)$ and $K(f_\chi^*)_i = \{\mathbf{a}_i \oplus \mathbf{a}_1, \dots, \mathbf{a}_i \oplus \mathbf{a}_t\}$ – the consolidated kernel [18] of function f_χ^* concerning the element $\mathbf{a}_i \in K(f_\chi^*)$. A set of all consolidated kernels of the Boolean function f_χ^* we will defined as $T(f_\chi^*) = \{K(f_\chi^*)_i = \mathbf{a}_i, K(f_\chi^*)_i | i = 1, 2, \dots, t\}$.

It is believed that the vector $\mathbf{a} = (\alpha_1, \dots, \alpha_m) \in Z_2^m$ precedes the vector $\mathbf{b} = (\beta_1, \dots, \beta_m) \in Z_2^m$ $\mathbf{a} < \mathbf{b}$, if $\alpha_i \leq \beta_i$ ($i = 1, 2, \dots, m$). We will denote $N_{\mathbf{a}}$ to be a set of all such vectors with Z_2^m , which precedes the vector \mathbf{a} .

Theorem 3. If the Boolean function $f: G_n \rightarrow H_2$ is realized by one generalized neural element regarding the system of characters $\chi = \{\chi_{i_1}, \chi_{i_2}, \dots, \chi_{i_m}\} \subset X(G_n)$ with the modified kernel $K(f_\chi, M)$, the kernel $K(f_\chi^*)$ of function f_χ^* satisfies the condition

$$\mathbf{a} = (\alpha_1, \dots, \alpha_m) \in K(f_\chi^*) \Rightarrow \bar{\mathbf{a}} = (\bar{\alpha}_1, \dots, \bar{\alpha}_m) \notin K(f_\chi^*),$$

where $\bar{\alpha}_i$ – an inverted value α_i .

Theorem 4. If the Boolean function $f: G_n \rightarrow H_2$ is realized by one generalized neural element regarding the system of characters in $\chi = \{\chi_{i_1}, \chi_{i_2}, \dots, \chi_{i_m}\} \subset X(G_n)$ with the modified kernel $K(f_\chi, M)$, a set of the consolidated kernel $T(f_\chi^*)$ contains the element $K(f_\chi^*)_i$, that

$$\forall \mathbf{a} \in K(f_\chi^*)_i \Rightarrow N_{\mathbf{a}} \subset K(f_\chi^*)_i.$$

Consider a set of matrices of tolerance [17]

$$F_m = \left\{ L_1 = (0_1), L_2 = \begin{pmatrix} L_1 & 0_1 \\ L_1^* & 0_1 \end{pmatrix}, \dots, L_m = \begin{pmatrix} L_{m-1} & 0_{m-1} \\ L_{m-1}^* & 0_{m-1} \end{pmatrix} \right\},$$

where 0_r – zero column of $2^{r-1} \times 1$.

Using the kernel elements $K(f_\chi^*)$ we will construct the matrix $K_\xi(f_\chi^*)$ as follows: the first line of the matrix $K_\xi(f_\chi^*)$ will be $\mathbf{a}_{\xi(1)} = (\alpha_{\xi(1)1}, \dots, \alpha_{\xi(1)m})$ out of $K(f_\chi)$, the second line of the matrix will be $\mathbf{a}_{\xi(2)} = (\alpha_{\xi(2)1}, \dots, \alpha_{\xi(2)m})$, the last line $K_\xi(f_\chi^*)$ will be $\mathbf{a}_{\xi(q)} = (\alpha_{\xi(q)1}, \dots, \alpha_{\xi(q)m})$, where $\xi(i)$ – an effect of substitution $\xi \in S_q$ for i (a symmetric group of degree q). Let, then (a symmetric group of degree m).

Theorem 5. If a set of the consolidated kernels $T(f_\chi^*)$ of the Boolean function $f: G_n \rightarrow H_2$ ($f \neq const$) regarding the system of characters $\chi = \{\chi_{i_1}, \dots, \chi_{i_m}\} \subset G_n$ with the modified kernel $K(f_\chi, M)$ contains the element $K(f_\chi^*)_i$ for which there exist the elements $\xi \in S_t$, $\sigma \in S_m$ and such a matrix of tolerance $L_{j_i} \in F_m$, that

$$K_\xi^\sigma(f_\chi^*)_i = (L_{j_i} \underbrace{0_{m-j_i}}_{m-j_i}) \nabla (L_{j_i}^*(q_0^i) \underbrace{0_{m-j_i}}_{m-j_i}) \nabla (L_{j_i+1}^*(q_1^i) \underbrace{0_{m-j_i-1}}_{m-j_i-1}) \nabla \dots \nabla (L_{j_i+r}^*(q_r^i) \underbrace{0_{m-j_i-r}}_{m-j_i-r})$$

where $q_0^i \geq q_1^i \geq \dots \geq q_r^i$, then the function $f: G_n \rightarrow H_2$ is realized by one generalized neural element regarding the system of characters χ . Based on the above mentioned theorems, an effective algorithm for synthesizing optimal generalized integer neural elements with a large number of inputs can be constructed.

III. THE EXAMPLES OF PRACTICAL APPLICATION OF RESEARCH FINDINGS

We will demonstrate the effectiveness and practical feasibility of using the research results obtained for the synthesis of the generalized neural elements with a large number of inputs (several hundred, several thousand) in the following examples.

1. Let $\mathbf{e}_0 = (0, 0, \dots, 0)$, $\mathbf{e}_1 = (1, 0, \dots, 0)$, $\mathbf{e}_2 = (0, 1, \dots, 0), \dots$, $\mathbf{e}_m = (0, 0, \dots, 1)$ – m - dimensional Boolean vectors and a complete prototype of Boolean function units $f: G_n \rightarrow H_2$ regarding the system of characters $\chi = \{\chi_{i_1}, \dots, \chi_{i_m}\} \subset \chi(G_n)$ ($2 \leq m \leq n$) is set as follows:

$$f_\chi^{-1}(1) = \{\mathbf{e}_1, \mathbf{e}_2, \dots, \mathbf{e}_m\} \text{ i } f_\chi^{-1}(1) \cup f_\chi^{-1}(0) = Z_2^m.$$

Then by definition $K(f_\chi^*) = K(f_\chi) = f_\chi^{-1}(1)$. We will construct a set of the consolidated kernels

$$T(f_\chi^*) = \left\{ K(f_\chi^*)_1 = \mathbf{e}_1 K(f_\chi^*), K(f_\chi^*)_2 = \mathbf{e}_2 K(f_\chi^*), \dots, K(f_\chi^*)_m = \mathbf{e}_m K(f_\chi^*) \right\},$$

where

$$K(f_\chi^*)_1 = \{\mathbf{e}_0, \mathbf{e}_1 \oplus \mathbf{e}_2, \mathbf{e}_1 \oplus \mathbf{e}_3, \dots, \mathbf{e}_1 \oplus \mathbf{e}_m\},$$

...

$$K(f_\chi^*)_m = \{\mathbf{e}_m \oplus \mathbf{e}_1, \mathbf{e}_m \oplus \mathbf{e}_2, \mathbf{e}_m \oplus \mathbf{e}_3, \dots, \mathbf{e}_0\}$$

Each consolidated kernel $K(f_\chi^*)_k$ has an element $\mathbf{e}_k \oplus \mathbf{e}_i$ ($k \neq i; i \neq 0$), but does not contain the elements \mathbf{e}_k i \mathbf{e}_i , which precede this element. So, none of the consolidated kernels out of $T(f_\chi^*)$ does not satisfy the conditions of Theorem 4 and it means that the function f is not realized by one generalized neural element concerning the system χ .

2. Let $\mathbf{e}_0 = (0,0,\dots,0)$, $\mathbf{e}_1 = (1,0,\dots,0)$, $\mathbf{e}_2 = (0,1,\dots,0),\dots$, $\mathbf{e}_m = (0,0,\dots,1)$ – m - dimensional Boolean vectors. We will consider the problem of the Boolean function implementation $f: G_n \rightarrow H_2$ regarding the system of characters $\chi = \{\chi_{i_1}, \dots, \chi_{i_m}\} \subset \chi(G_n) (3 \leq t < m \leq n)$, if

$$f_\chi^{-1}(1) = \{\mathbf{e}_1, \mathbf{e}_2, \dots, \mathbf{e}_t, \mathbf{e}_0, \mathbf{e}_1 \oplus \mathbf{e}_2, \dots, \mathbf{e}_1 \oplus \mathbf{e}_m\} \text{ and}$$

$$f_\chi^{-1}(1) \cup f_\chi^{-1}(0) = Z_2^m.$$

The restriction $t \geq 3$ is imposed in terms of unambiguousness of the definition $K(f_\chi^*)$. If $t \geq 3$, then any $m(t < m \leq n) m + t \leq 2^{m-1}$ and in this case $K(f_\chi^*) = K(f_\chi) = f_\chi^{-1}(1)$, and in the opposite case ($t < 3$) $K(f_\chi^*) = f_\chi^{-1}(0)$.

In this case $K(f_\chi^*) = f_\chi^{-1}(1)$. We will construct a set of the consolidated kernels

$$T(f_\chi^*) = \left\{ \begin{array}{l} K(f_\chi^*)_1 = \mathbf{e}_1 K(f_\chi^*), K(f_\chi^*)_2 = \mathbf{e}_2 K(f_\chi^*), \dots \\ \dots, K(f_\chi^*)_{t+m} = (\mathbf{e}_1 \oplus \mathbf{e}_m) K(f_\chi^*) \end{array} \right\}.$$

We will write out the elements of the consolidated kernels $K(f_\chi^*)_i$:

$$K(f_\chi^*)_1 = \{\mathbf{e}_0, \mathbf{e}_1 \oplus \mathbf{e}_2, \dots, \mathbf{e}_1 \oplus \mathbf{e}_t, \mathbf{e}_1, \mathbf{e}_2, \dots, \mathbf{e}_m\},$$

$$K(f_\chi^*)_2 = \left\{ \begin{array}{l} \mathbf{e}_1 \oplus \mathbf{e}_2, \mathbf{e}_0, \dots, \mathbf{e}_2 \oplus \mathbf{e}_t, \mathbf{e}_2, \mathbf{e}_1, \mathbf{e}_1 \oplus \mathbf{e}_2 \oplus \mathbf{e}_3, \dots \\ \dots, \mathbf{e}_1 \oplus \mathbf{e}_2 \oplus \mathbf{e}_m \end{array} \right\},$$

.....

$$K(f_\chi^*)_{t+m} = \left\{ \begin{array}{l} \mathbf{e}_m, \mathbf{e}_1 \oplus \mathbf{e}_2 \oplus \mathbf{e}_m, \dots, \mathbf{e}_1 \oplus \mathbf{e}_t \oplus \mathbf{e}_m, \mathbf{e}_1 \oplus \mathbf{e}_m, \\ \mathbf{e}_2, \dots, \mathbf{e}_{m-1}, \mathbf{e}_0 \end{array} \right\}.$$

Out of the elements $K(f_\chi^*)_1$ we can construct a matrix

$$K_\xi(f_\chi^*)_1 = (L_3 \underbrace{0 \dots 0}_{m-3}) \nabla (L_3 \underbrace{2 \dots 0}_{m-3}) \nabla \dots$$

$$\nabla (L_t \underbrace{2 \dots 0}_{m-t}) \nabla (L_{t+1} \underbrace{1 \dots 0}_{m-t-1}) \nabla \dots \nabla (L_m \underbrace{1 \dots 1}_{m-t-1}), \quad (6)$$

where the element $\xi \in S_{t+m}$ defines the order of the elements $3 K(f_\chi^*)_1$ within the matrix.

On the basis of Theorem 5 and equality (6) we will construct a vector $\mathbf{w} = (\omega_1, \omega_2, \dots, \omega_m)$:

$$\omega_1 = -1, \omega_2 = \omega_1 - 1 = -2, \omega_3 = \omega_4 = \dots = \omega_t = -4, \omega_{t+1} = \dots, = \omega_m = -5.$$

Then

$$\mathbf{w}_1 = \mathbf{e}_1 \mathbf{w}^{\sigma^{-1}} \text{ i } \omega_0^{(1*)} = \min\{\langle \mathbf{w}_1, \mathbf{x} \rangle \mid \mathbf{x} \in f_\chi^{-1}(1)\} - \varepsilon,$$

$$\text{where } \varepsilon \in \left(0, \frac{1}{2}\right).$$

Taking into account that σ is a single element of the group S_m and setting $\varepsilon = \frac{1}{4}$ we will obtain the following

vector structure $(\mathbf{w}_1 = (\omega_1^{(1)}, \dots, \omega_m^{(1)}), \omega_0^{(1*)})$ GNE regarding the system $\chi = \{\chi_{i_1}, \dots, \chi_{i_m}\} \subset \chi(G_n)$:

$$\omega_1^{(1)} = 1, \omega_2^{(1)} = -2, \omega_3^{(1)} = \omega_4^{(1)} = \dots = \omega_t^{(1)} = -4, \omega_{t+1}^{(1)} = \dots$$

$$= \omega_m^{(1)} = -5 \text{ i } \omega_0^{(1*)} = -\frac{17}{4},$$

that realizes the function f_χ in $\{0,1\}$.

$$\omega_0^{(1)} = \sum_{j=1}^m \omega_j^{(1)} - 2\omega_0^{(1*)} = 1 - 2 + (-4)(t-2) +$$

$$+ (-5)(m-t) - 2\left(-\frac{17}{4}\right) = t - 5m + 15,5$$

We reveal and a generalized neural element with vector structure $(\mathbf{w}_1; \omega_0^{(1)})$ realizes the function $f: G_n \rightarrow H_2$ in $\{-1,1\}$

3. We will consider the case where a number of characters m within the system of characters $\chi = \{\chi_{i_1}, \dots, \chi_{i_m}\} \subset \chi(G_n)$ satisfies inequality

$n < m < 2^n$. Let $f: G_n \rightarrow H_2$ and $f_\chi^{-1}(1) = \{\mathbf{e}_1, \mathbf{e}_2, \dots, \mathbf{e}_m\}$, $f_\chi^{-1}(0) = \{\bar{\mathbf{e}}_1, \bar{\mathbf{e}}_2, \dots, \bar{\mathbf{e}}_m\}$, $f_\chi^{-1}(*) = Z_2^m \setminus (f_\chi^{-1}(1) \cup f_\chi^{-1}(0))$, where the vinculum signifies a logical denial operation of the Boolean vector coordinates. In case of $m > n$ we always have $Z_2^m \neq f_\chi^{-1}(1) \cup f_\chi^{-1}(0)$. Since $f_\chi^{-1}(1) \cap f_\chi^{-1}(0) = \emptyset$,

the kernel $K(f_\chi)$ the following exists $K(f_\chi) = f_\chi^{-1}(1)$. It was shown above (point 1) that when $K(f_\chi) = f_\chi^{-1}(1) = \{\mathbf{e}_1, \mathbf{e}_2, \dots, \mathbf{e}_m\}$, $Z_2^m = f_\chi^{-1}(1) \cup f_\chi^{-1}(0)$, the function f is not realized by one GNE regarding the system χ . In this case $Z_2^m \neq f_\chi^{-1}(1) \cup f_\chi^{-1}(0)$ and we will demonstrate, that it is possible to construct such an extended kernel that $K(f_\chi, A)$, that a function corresponding to this

kernel f_χ^* will be realized by one generalized neural element regarding the system χ , which means that the function f is also realized by one GNE concerning χ . Actually, if one of the elements $f_\chi^{-1}(*)$ we will construct a set $A = \{\mathbf{e}_0, \mathbf{e}_1 \oplus \mathbf{e}_2, \dots, \mathbf{e}_1 \oplus \mathbf{e}_m\}$, then

$$K(f_\chi, A) = \{\mathbf{e}_1, \mathbf{e}_2, \dots, \mathbf{e}_t, \mathbf{e}_0, \mathbf{e}_1 \oplus \mathbf{e}_2, \dots, \mathbf{e}_1 \oplus \mathbf{e}_m\} \text{ and}$$

$$K(f_\chi^*) = K(f_\chi, A).$$

Out of the elements of the consolidated kernel

$$K(f_\chi^*)_1 = \mathbf{e}_1 K(f_\chi^*) = \{\mathbf{e}_0, \mathbf{e}_1 \oplus \mathbf{e}_2, \dots, \mathbf{e}_1 \oplus \mathbf{e}_m, \mathbf{e}_1, \mathbf{e}_2, \dots, \mathbf{e}_m\}$$

The following matrix can be constructed

$$K_{\xi}(f_{\chi}^*)_1 = (L_3(0, \dots, 0) \nabla_{m-3}^* (L_3(2) 0, \dots, 0) \nabla_{m-3}^* (L_4(2) 0, \dots, 0) \nabla_{m-4}^* \dots \nabla_{m-4}^* (L_m^*(2))) \quad (7)$$

in which the order of the elements of $K(f_{\chi}^*)_1$ into the matrix lines is determined by the element $\xi \in S_{2m}$.

Then, by Theorem 5 and equation (7), the coordinates of the vector $\mathbf{w} = (\omega_1, \omega_2, \dots, \omega_m)$ are determined as follows:

$$\omega_1 = -1, \omega_2 = -2, \omega_3 = \omega_4 = \omega_m = -4.$$

Similar to the previous case we obtain

$$\omega_1^{(1)} = 1, \omega_2^{(1)} = -2, \omega_3^{(1)} = \omega_4^{(1)} = \dots = \omega_m^{(1)} = -4 \quad \text{i} \quad \omega_0^{(1*)} = -\frac{17}{4}.$$

Hence GNE with the structure vector $(\mathbf{w}_1 = (1, -2, -4, \dots, -4); \omega_0^{(1*)} = -\frac{17}{4})$ implements a function f_{χ} in $\{0,1\}$. We will calculate

$$\omega_0^{(1)} = \sum_{j=1}^m \omega_j^{(1)} - 2\omega_0^{(1*)} = 1 - 2 + (-4)(m-2) - 2\left(-\frac{17}{4}\right) = -4m + 15,5.$$

To conclude, a generalized neural element with vector structure $(\mathbf{w}_1; \omega_0^{(1)})$ implement a function $f: G_n \rightarrow H_2$ concerning the system $\chi = \{\chi_i, \dots, \chi_{i_m}\} \subset \chi(G_n)$.

IV. CONCLUSION

Concerning the research findings, the effective methods of synthesis of neural networks on the basis of the generalized neural elements for the processing, compression, classification and recognition of discrete signals are developed, as well as when the approximation and iteration methods of synthesis of neural elements can not practically be applied due to a large number of inputs of neural elements.

REFERENCES

[1] M. Azarbad, S. Hakimi, and A. Ebrahimzadeh, "Automatic Recognition of Digital Communication Signal," *International Journal of Energy, Information and Communications* vol.3, is. 4, pp. 21–33, 2012.

[2] I. V. Isonin, R. O. Tkachenko, D. D. Peleshko, and D. A. Batuk, "Neural network method for changing the resolution of images," *Systems of information processing*, is. 9(134 pp. 30–34, 2015.

[3] F. Amato, J. L. González-Hernández, and J. Havel, "Artificial neural networks combined with experimental desing: a "soft" approach for chemical kinetics," *Talanta*, vol. 93, pp. 72–78. 2012.

[4] F. Geche, O. Mulesa, S. Geche, and M. Vashkeba, "Development of the method of synthesis of the prediction scheme on the basis of basic

forecasting models." *Technological audit and production reserves*, no. 3/2(23), pp. 36–41, 2015. Mode of access : DOI : [10.15587/2312-8372.2015.44932](https://doi.org/10.15587/2312-8372.2015.44932).

[5] P. Dey, A. Lamba, S. Kumary, and N. Marwaha, "Application of an artificial neural network in the prognosis of chronic myeloid leukemia," *Analytical and quantitative cytology and histology, International Academy of Cytology and American Society of Cytology*, vol. 33 (6), pp. 335–339, 2011.

[6] A. S. Liu, and Q. Zhu, "Automatic modulation classification based on the combination of clustering and neural network," *The Journal of China Universities of Ports and Telecommunication*, vol. 18, no.4, pp. 13–19, 2011.

[7] A. Pathok, and A. K. Wadhvani, "Data Compression of ECG Signals Using Error Back Propagation (EBP) Algorithm," *International Journal of Engineering and Advence Technology (IJEAT)*, vol. 1, iss. 4, pp. 256–260, 2012.

[8] Ye. Bodyansky, P. Grimm, S. Mashtalir, and V. Vinarski, "Fast training of neural networks for image compression," *Lecture Notes in Computer Science. Berlin-Heidelberg-New York, Springer*, vol. 6171, pp. 165–173, 2010.

[9] Ye. Bodyanskiy I. Pliss., D. Peleshko, Yu. Rashkevych, and O. Vynokurova, "Hybrid Generalized Additive Wavelet-Neuro-Fuzzy-System and its Adaptive Learning," *Dependability Engineering and Complex Systems: Proceedings of the Eleventh International Conference on Dependability and Complex Systems DepCoS-RELCOMEX.*, Brunow, Poland, pp.51-61, June 27-July 1, 2016

[10] N. V. Shovgun, "Analysis of the effectiveness of fuzzy neural networks concerning the credit risk assessment," *Information technologies & knowledge. ITHEA IBS ISC*, vol.7, pp. 286–293, 2013.

[11] Y. Bodyanskiy, G. Setlak, D. Peleshko, and O. Vynokurova, "Hybrid Generalized Additive Neuro-Fuzzy System and its Adaptive Learning Algorithms," *8th IEEE International Conference on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications*, Warsaw, Poland, pp. 328-333, 24-26 September 2015

[12] T. Teslyuk, I. Tsmots., V. Teslyuk, M. Medykovskyy, and Y. Opotyak, "Architecture and Models for System-Level Computer-Aided Design of the Management System of Energy Efficiency of Technological Processes at the Enterprise," *Advances in Intelligent Systems and Computing*, vol. 689, Springer, Cham. pp. 538 – 557, 2018.

[13] I. Tsmots, V. Teslyuk, T. Teslyuk, and I. Ihnatyev, "Basic Components of Neuronetworks with Parallel Vertical Group Data Real-Time Processing," *Advances in Intelligent Systems and Computing*, vol. 689, Springer, Cham. pp. 558 – 576, 2018.

[14] Vasyl Teslyuk, Vasyl Beregovskiy, Pavlo Denysyuk, Taras Teslyuk, and Andrii Lozynskiy, "Development and Implementation of the Technical Accident Prevention Subsystem for the Smart Home System," *International Journal of Intelligent Systems and Applications(IJISA)*, vol.10, no.1, pp.1-8, 2018. DOI:10.5815/ijisa.2018.01.01.

[15] C. Curtis, I. Reiner, *Representation theory of finite groups and associative algebras*. M.: Nauka, 1969.

[16] B. I. Golubov, A. V. Efimov, and V. A. Skvortsov, *Walsh series and transformations. Theory and applications*. M.: Nauka, 1987.

[17] F. Geche, O. Mulesa, and V. Buchok, "Synthesis of generalized neural elements by means of the tolerance matrices," *Eastern European Journal of Advanced Technology*, vol. 4 / 4(88), pp.50-62, 2017.

[18] N.N. Aisenberg, A.A. Bovdi, E.J. Gergo, and F.E. Geche, "Some algebraic aspects of threshold logic," *Cybernetics*, no. 2, pp. 26-30, 1980.

Implementation of Information Technologies in the organization of Forest Fire Suppression Process

Olga Smotr

Department of Project Management, Information Technologies and Telecommunications
Lviv State University of Life Safety
Lviv, Ukraine
olgasmotr@gmail.com

Nazarii Burak

Department of Project Management, Information Technologies and Telecommunications
Lviv State University of Life Safety
Lviv, Ukraine
nazar.burak@ukr.net

Yuriy Borzov

Department of Project Management, Information Technologies and Telecommunications
Lviv State University of Life Safety
Lviv, Ukraine
uob1968@gmail.com

Solomija Ljaskovska

Department of designing and operation of machines
Lviv Polytechnic National University
Lviv, Ukraine
solomiam@gmail.com

Abstract— the article deals with information and technical support of the forest fire suppression process. The status update on the problem is considered. The necessity of using modern information decision-making systems in the process of forest fires suppression is substantiated. Improvement of the method of a decision support system constructing using MVC coding patterns is proposed.

Keywords — *information technologies, decision-making systems, MVC coding patterns, organization of the forest fire suppression process.*

I. INTRODUCTION

Forest fires are essential natural and anthropogenic factors that excitedly change functions and conditions of the forests. The analysis of the statistical data on the number of forest fires and the area of forests destroyed by fires in the European countries and in Ukraine over the past twenty years has shown that the number of fires as well as the damage caused by them increase gradually [1,2].

The range of urgent tasks aimed at the rapid and effective suppression of forest fires is increasing significantly, and the conditions for their solution are continuously complicated, both by the scale of destructive actions and by various random factors. This requires the development of measures aimed at optimal control of forces and means for the suppression of such fires. This involves such peculiarities of work as operational data collection, analysis and rapid decision-making. The use of information technologies aimed at minimizing the time of managerial decision-making and the cost of its implementation is necessary in such conditions. Implementation of the modern information decision-making systems in the work of the State Emergency Service of Ukraine (SES of Ukraine) gives the possibility to provide rapid and qualitative processing of the incoming information on the nature and characteristics of the forest fire, to predict its parameters and to generate acceptable scenarios for its development. This, in turn, allows the forest fireground commander (FFC) to make scientifically-substantiated and effective management decisions for its suppression.

The world practice of effective management in various sectors, including control in emergency situations [3-6], proves the need for application of informational decision-making systems (DMS), in which the knowledge and experience of specialists working in the relevant spheres of human activity are used.

Today, the departments of the State Emergency Service of Ukraine use the "Government Information and Analytical System for the Suppression of Emergencies" (GIASSE) to solve operational and tactical tasks. In particular, in the Lviv region, the Supervisory Control And Data Acquisition (SCADA), the territorial subsystem of the GIASSE [7], is used.

II. PROBLEM STATEMENT

However, the Supervisory Control and Data Acquisition do not solve such tasks as:

- evaluating of the potential for emergency response throughout the region;
- modelling of forest fire spread;
- generating, analysis and selection of the best fire-fighting tactics, taking into account the operational environment, technical capabilities and resource constraints;
- visualisation of the chosen scenario;
- support of the chosen scenario implementation in real time mode with the possibility of its correction due to changes in the operational environment.

SCADA is mainly intended for supervisory monitoring and does not help the forest fireground commander to make a decision.

The above circumstances determine the need for further improvement of the existing and implementation of new modern information DMS in the operational activities of SES of Ukraine.

According to the main provisions of the general management theory, managing forces and means during the

forest fires suppression can be considered in time as the functioning of supervisory control system. In particular, it can be represented as a dynamic system [8] (Fig. 1), where: the input parameters P_i are determined by terrain features, forest vegetation, etc.; perturbing actions P_j – weather conditions (speed and direction of wind, humidity, etc.); ΔP_i , ΔP_j – leading factors, the values of which depend on the parameters of the purpose of fireground command and control C_k ; O_l – set of output parameters (forest fire area, its intensity and rate of fire spread, etc.). All these conditions can vary: topographic - in space, weather - in space and in time, the purpose of fireground command and control depends on the strategy and tactics of forest fire suppression. The values of the leading factors (actions) depend on the intensity of the forest fire spread and the efforts made to eliminate it. The dynamics of forest fires and their suppression is usually rapid, and meteorological conditions are often unpredictable. The information on the conditions of the forest fire is usually incomplete, uncertain and, in some cases, erroneous, which complicates the process of its suppression. The human factor (e.g. professional training and skills) and the material base of fire and rescue units are also important, because these factors effect on the speed of the forest fire suppression.

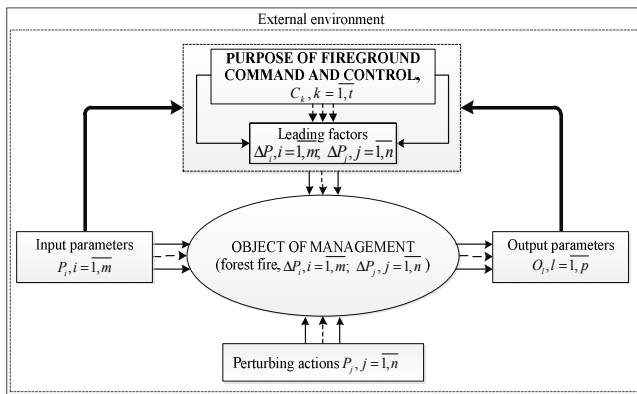


Fig. 1. Management of a dynamic forest fire suppression system

In this context, the decision-making process, that deals with the strategy of forest fire suppression, in its essence, is identified with the activities of the decision maker, that is, the forest fireground commander.

When operating fire-rescue units (FRU), DMS-based activities of the forest fireground commander can be shown by using such block diagrams (Fig. 2), on which it is possible to study the main stages of the management decision-making process.

Obviously, modelling the current state of the control object (forest fire), generating a set of variants of its interaction with the objects of action (FRU's forces and means), predicting the expected effects of the decisions, etc. are impossible without prompt and qualitative modelling of the forest fire behavior.

III. THE PROPOSED METHOD OF CONSTRUCTING A DECISION-MAKING SYSTEM IN ORGANIZING THE FOREST FIRE SUPPRESSION PROCESS

To simulate the forest fire behaviour, a scheme based on classic Model-View-Controller coding patterns [9, 10] is widely used nowadays. It consists of three levels:

- Model level – Provides data (usually for view), and responds to queries (usually from the controller), changing its status;
- View level – is responsible for displaying both the input information and the information obtained during the process of modelling (user interface);
- Controller level - interprets data entered by the user and informs the model about the need for the corresponding response (control logic).

The improved structure of the decision-making system for forest fire elimination is shown in Fig. 3

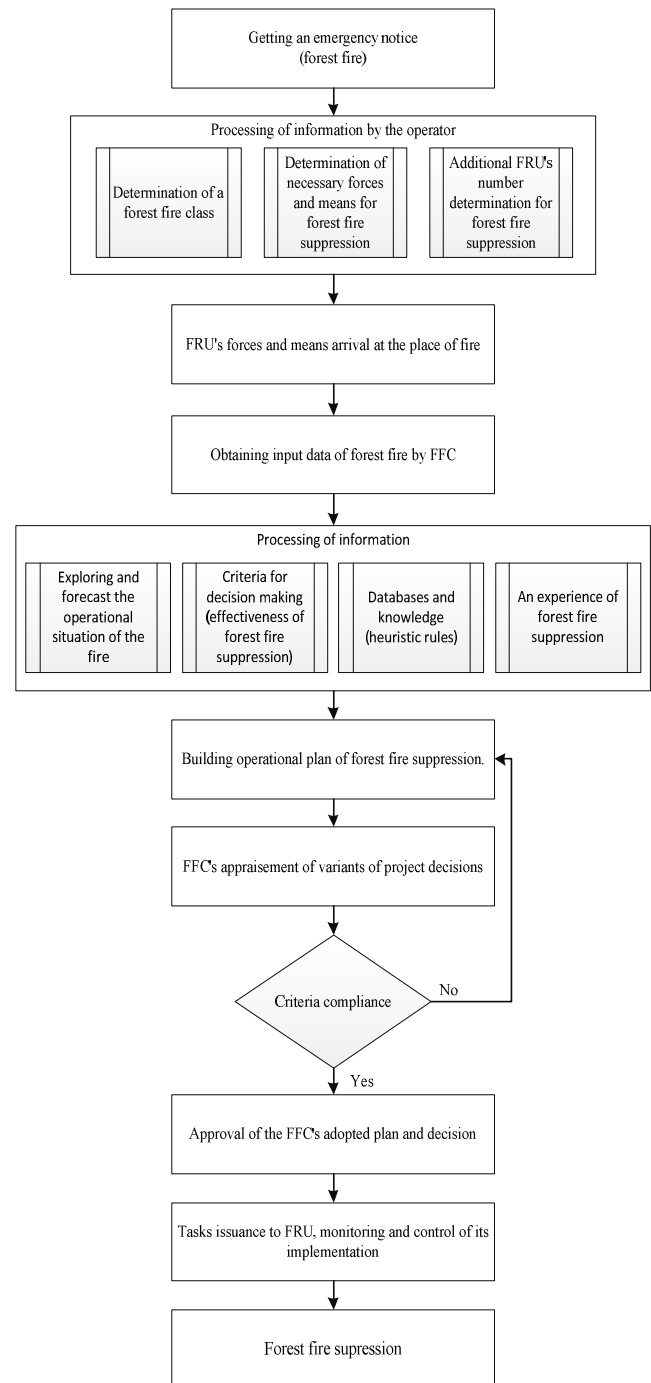


Fig. 2. FFC's activities

In such scheme, both View and Controller levels are directly dependent on the Model level, but the Model level does not depend either on the View or on the Controller. This is one of the key benefits of this delineation, which allows

building a control object model regardless of its visual representation, as well as creating several different views for one model.

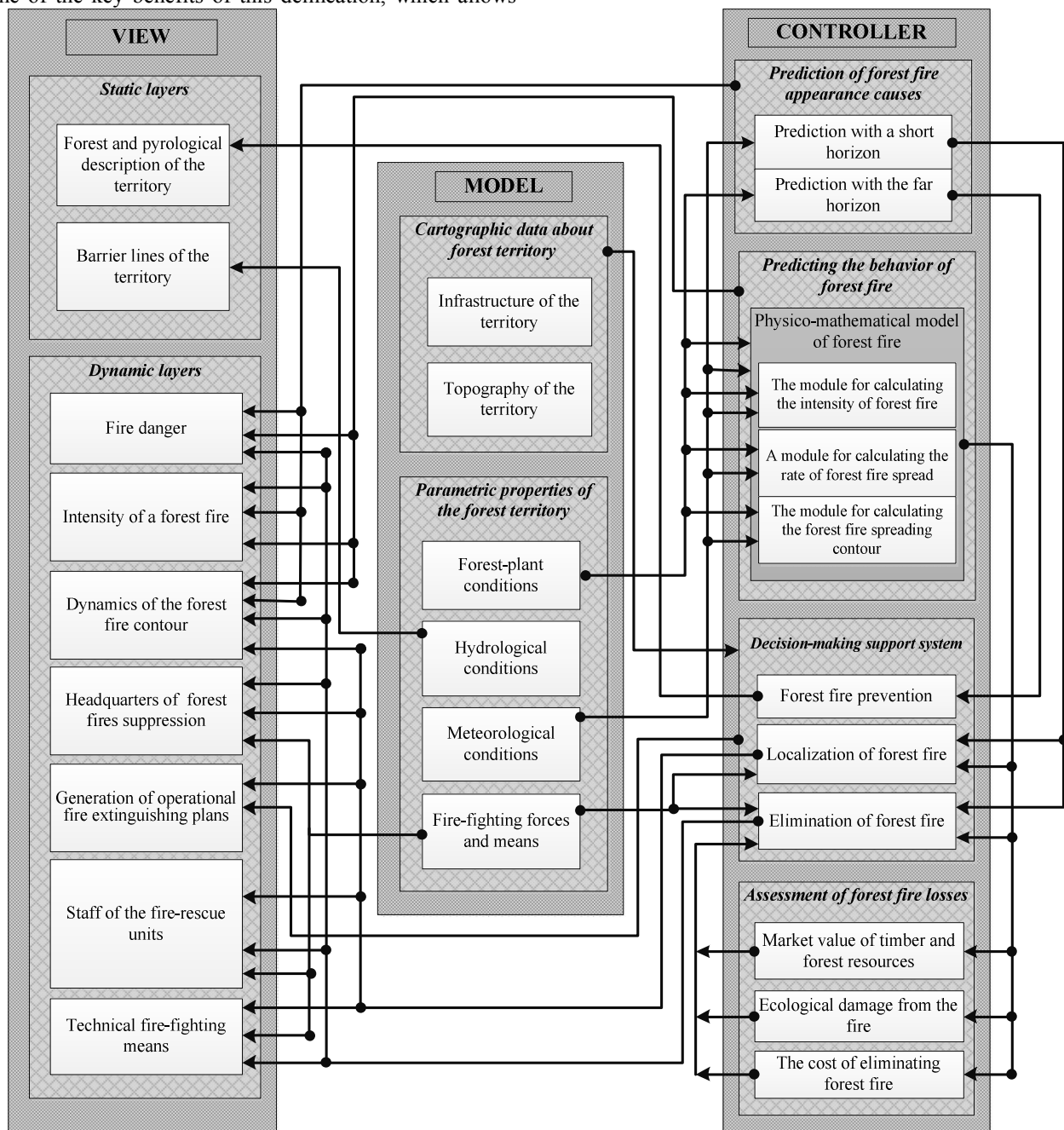


Fig. 3. Structure of the decision-making system for forest fire elimination (based on [11])

The View level of the MVC scheme is an electronic map of the forest area given to the forest fireground commander for a visual familiarization with its features. The map has a layered structure and consists of the main (static) and additional (dynamic) layers. Static layers usually reflect forest vegetation and pyrological characteristics of forest areas that may be subject to fires. Such layers include forest areas (which represent layers of deciduous and coniferous species, young stands or litter, etc.), soils, water bodies (showing lakes, swamps, rivers, etc.), access roads and routs (the type of road surface, width, height/descent are taken into account), etc. Dynamic layers display data that undergoes pre-processing at the Controller level. These are layers such

as a mapping scheme for the distribution of fire danger (pre-calculated in the forecasting block for the reasons for the occurrence of a forest fire), forest fire behaviour (determined in the block of forest fire spread forecasting), and decision-making systems. In addition, the forest fire command center, the personnel of the fire and rescue units, as well as the location of fire extinguishing equipment (the number and composition of which are also determined in the appropriate block) are displayed visually on the dynamic layer.

The Model is one of the three components of the MVC coding pattern that describes the state of spatial data at the time of the forest fire detection. It consists of two interrelated parts – cartographic and attributive. Cartographic data

describes the positional characteristics of terrain (topography and infrastructure of vector objects). Attributive data gives the characteristics of vector objects – the dynamics of forces and means transfer, as well as meteorological and hydrological conditions.

The logical part of MVC coding pattern (Controller level) consists of three main blocks, each of which is aimed at solving one of the tasks:

- prediction of forest fire spread;
- assessment of fire losses;
- decision-making system for forest fires preventing and suppression.

We offer to add the fourth block to the list of tasks that are solved by the MVC controller: analysis of the causes of forest fires and calculation of burning index class. This will give an opportunity to evaluate the preparedness of forest vegetation areas for combustion and predict fire risk. At the same time, we offer to implement two options for predicting the causes of forest fires. The first of these is prediction with a short horizon and is used in the development of operational tactics for the forest fire suppression. The second – with the far horizon (used for prediction of the causes of forest fires and development of a strategy for their prevention).

Considered above proposed method of constructing a decision-making system in organizing the forest fire suppression process makes it possible not only to model fire behaviour but also to calculate fire risks for taking preventive measures.

IV. PRACTICAL SOLUTION OF FOREST FIRE BEHAVIOUR PREDICTION PROBLEM

An important stage in development of conceptual strategies and choosing effective forest fire suppression tactics is prediction of forest fire global characteristics. They are determined by the influence of its main factors on the spreading geometry at any given time.

When developing the process of forest fire liquidation management strategy is necessary to anticipate and take into account the dangerous tendencies of its further development and also potential threats to settlements and security objects. Therefore, such prediction is extremely important.

Calculation of the probable forest fire spread rate and the intensity of the heat production during combustion process is made within a defined stage of time prediction: first in the direction of spreading the front of the fire, then the flanks and the rear. We used an empirical formula for prediction the speed of fire limit spread on each forest section. This formula contains variable coefficients of relative influence of various factors, namely:

$$\tilde{V}^h(\alpha_v, \omega_a, v_c) = \left\{ V_i^h(\alpha_v, \omega_a, v_c) = V_0(r_n, p_\kappa, u_g) K_i(\alpha_v) K_i(\omega_a) K_i^h(v_c), i = \overline{1, N^d} \right\} (1)$$

where $V_0 = V_0(r_n, p_\kappa, u_g)$ – basic (calm) rate of fire spread, m/min; r_n – type of forest vegetation specified in the area's pyrological description; p_κ – complex meteorological fire danger index at the time of fire appearance; u_g – conditions for drying of forest vegetation in area; $K_i(\alpha_v)$, $K_i(\omega_a)$, $K_i^h(v_c)$ – coefficients of influence on the rate of fire spread

in accordance with the terrain inclination α_v , relative air humidity ω_a and wind speed v_c ; N^d – number of forest areas; index 'h' defines wind direction $h \in \{\text{fr}, \text{fl}, \text{r}\}$: fr – frontal, fl – flank, r – rear.

For all coefficients of influence with using the method of least squares approximation of experimental data empirical dependencies were obtained for next calculation of their values.

Coefficient of the terrain inclination influence $K_i(\alpha_v)$ at i – forest area, specified in area's pyrological description of forest vegetation. It depends on magnitude of the angle α_v and slope exposure towards the sides of horizon and direction of combustion process spreading (up the slope – angles are positive, down – are negative, across the slope – zero):

$$K(\alpha_v) = 0,0142e^{0,1635\alpha_v} + 1,0776e^{0,0169\alpha_v}. \quad (2)$$

The value of relative air humidity influence coefficient $K_i(\omega_a)$ on i -area of forest should correspond to the value predicted its relative humidity $\omega_a = \omega_a(t)$ during the course of fire at t - moment of its occurrence.

$$K(\omega_a) = 2,028e^{-0,0898\omega_a} + 1,9982e^{-0,0186\omega_a}. \quad (3)$$

Coefficient of wind power impact $K_i^h(v_c)$ on i – forest area is determined using quadratic form that based on speed v_c and wind direction relative to the spread of fire:

$$K^h(v_c) = 1,0 + \begin{pmatrix} 0,17877 \\ 0,16192 \\ 0,03659 \end{pmatrix} v_c + \begin{pmatrix} 0,32705 \\ 0,10822 \\ 0,02609 \end{pmatrix} v_c^2 : h \in \{\text{fr}, \text{fl}, \text{r}\}, \quad (4)$$

where $K^h(v_c) = (K^{\text{fr}}(v_c), K^{\text{fl}}(v_c), K^{\text{r}}(v_c))$.

Wind speed v_c under the forest mass is normalized by the following formula:

$$v_c(v_m, p_d) = v_m K_d(p_d), \quad (5)$$

where v_m – wind speed according to the weather station, m/s; $K_d(p_d)$ – coefficient of normalising to the fullness of the tree p_d , which is determined by the ternary algebraic form.

$$K_d(p_d) = 0,7818 - 0,9452p_d - 0,2527p_d^2 + 0,46995p_d^3. \quad (6)$$

Calculation of the values for each forest area and each tactical part of the forest fire is carried out by the end of the first stage of prediction. For determining saturation point of last section we should take into account the duration of fire spreading to this area, namely:

$$L_i^h = V_i^h(\dots) \cdot (t'_i - t'_{i-1}), i = N^d : h \in \{\text{fr}, \text{fl}, \text{r}\}, \quad (7)$$

where L_i^h – the distance that the edge of fire must pass on $i = N^d$ section to the end of determined forecasting stage, m; $V_i^h(\dots)$ – speed of fire spread on the area in the last sector,

m/min; t' – time of the forecasting stage completion, min; t'_{i-1} – time of the fire spread completion in penultimate section, min.

For confirmation of formulas (3) and (4) was conducted full-factorial experiment of determining actual forest fires spread speed. The relative error of calculated fire spread rate by dependence (1) in relation to the actual value which was obtained experimentally, is by about 5.5%. This value of the error indicates that the simulation results which are obtained on the basis of observational statistical data processing are close to the actual ones. Similar researches were conducted for predicting the intensity of fire heat release at its boundaries, for computing speeds of fire perimeter increase, which depends on the speed of fire frontal edge spread, and for determining forests soot height.

The developed mathematical models of forest fire spread dynamics allow us to determine the speed of forest fires edge spreading and to predict the intensity of heat release from it. All calculations depend on the characteristics of main combustion conductor's types and complex meteorological indicator of fire hazard.

Practical calculations of forest fires edge spread speed and predicting the intensity of heat release were performed in special software. This software uses detailed maps of forest vegetation and weather forecast data. This makes it possible to take into account dangerous trends in forest fire spreading and to develop a strategy for its elimination. Also it helps to determine methods of stopping and calculating the necessary amount of forces and means for eliminating the fire (see Fig.4.)

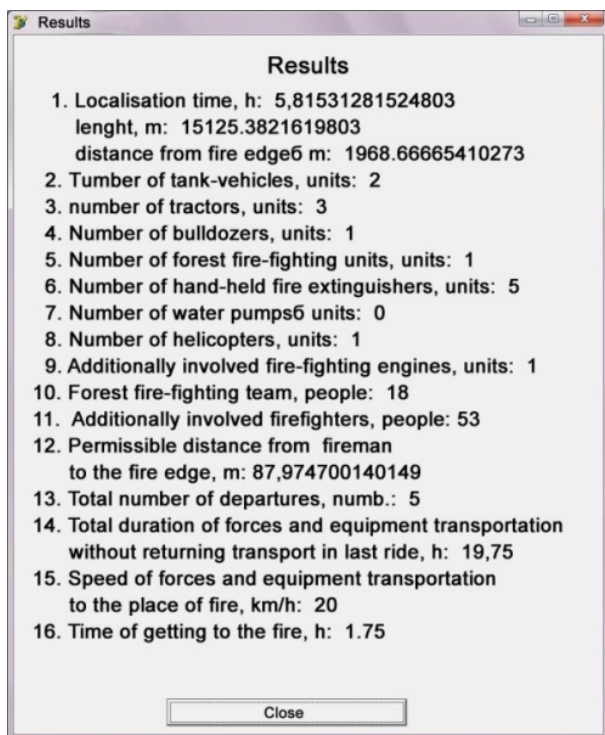


Fig. 4. Optimal amount of means and forces for the forest fire elimination calculated in program

V. CONCLUSION

Forest fireground commander in emergency situation has to make operational decisions in such conditions: incompleteness, unreliability, inaccuracy of incoming information, rapid flow of events, limited time for a comprehensive analysis of the situation, a significant number of participants involved in the forest fire suppression process, etc. Therefore, probability of making an effective decision without the use of decision-making systems is rather low. In order to reduce the probability of wrong choice in determining the forest fire suppression strategy, and to reduce the time for making management decisions, the FCC must apply proper decision-making systems that accurately describe the processes occurring during the forest fires and are use the special knowledge of the best specialists with extensive experience in this field. Unfortunately, despite the significant number of existing decision-making systems, today the SES of Ukraine does not have effective decision-making systems suitable directly for the forest fireground commanders.

One of the ways used for improvement of such systems is introducing MVC coding patterns. Such patterns enable not only to model the forest fire behaviour but also to calculate fire risks and to assess the material, environmental and socio-organizational consequences of potential fires in order to taking for taking preventive measures.

REFERENCES

- [1] National report on the state of technogenic and natural safety in Ukraine, 2017. Retrieved from www.dsns.gov.ua/files/2017/8/18/Analit%20dopovid/2%20statistit.pdf.
- [2] Official web-page of National Interagency Fire Center (USA). Retrieved from www.nifc.gov.
- [3] S. D. Bushuyev, N. S. Bushueva, I. A. Babaev, et al., Creative technologies in managing projects and program. Kiev: Summit Book. 2010.
- [4] V. Titova, "Informational and analytical decision support for operational duty services," *Artificial intelligence*, no. 4, pp. 504-509, 2006.
- [5] C. W. Holsapple, and A. B. Whinston, *Decision Support Systems: A Knowledge-Based Approach*. West St. Paul, MN, 1996. Retrieved from <http://www.uky.edu/BusinesEconomics/dssakba/inmat.htm>.
- [6] O. O. Smotr, "Use of information systems to support decision-making in management of fire and rescue units of the Emergency Ministry of Ukraine," In *The problems of information technologies application, special technical means in the activity of executive authorities, educational process, interaction with other services*, Lviv: Lviv State University of Internal Affairs, pp. 270-273, 2011
- [7] Official web-page of The State Emergency Service of Ukraine in Lviv region. Retrieved from <http://lviv.dsns.gov.ua/ua/Sistema-operativno-dispatcherskogo-upravlinnya-SODU.html>.
- [8] G. A. Dorrer, *Mathematical models of forest fire dynamics*. Moscow: Forest industry, 1979.
- [9] G. D. Glavatsky, and V. M. Grumans, "Information model and tasks of optimizing the process for forest fires combating," *Forestry*, no.1, pp. 36-41. 2002
- [10] V. E. Khodakov, and M. V. Zharikova, "Architecture of information technology for decision support for prevention and suppression of forest fires," *Problems of Information Technology*. no. 02(006), pp.116-122. 2009.
- [11] O. O. Smotr, and Yu. I. Grytsyuk, "Models and methods of management of extinguishing forest fires," *Bulletin of the Lviv State University of Life Safety*, no. 5, pp. 123-129, 2011.

Transformation of Information Based on Noisy Codes

Oleg Riznyk
PIT Department
Lviv Polytechnic National University
Lviv, UKRAINE
riznykoleg@gmail.ua

Yurii Noga
CMP Department
Lviv Polytechnic National University
Lviv, UKRAINE
yra.noga@gmail.com

Olexandr Povshuk
PIT Department
Lviv Polytechnic National University
Lviv, UKRAINE
alefox.pov@gmail.com

Yurii Kynash
PIT Department
Lviv Polytechnic National University
Lviv, UKRAINE
yuk.itvs@gmail.com

Abstract—In the article transformations of information are examined on the basis of noise codes for realization of code of information. The worked out methods of construction of code combinations of numbers are on the basis of theory of numerical bundles, which enables presentation of code combinations of numbers as a noise code.

Keywords—bundle, code, Golomb ruler, noise code.

I. INTRODUCTION

The main problem of modern theory and technology of communication and radio control is to increase the noise immunity of telecommunication systems and, in particular, command radio control lines, in the influence of both natural obstacles and created by the opponent artificial obstacles. One of the main concepts of increasing noise immunity, developed in this paper, is to quickly carry out a change in the working ensembles of codes, thereby increasing the immunity, energy and parametric secrecy of the communication system, as well as protection of information from unauthorized access [1, 2, 3, 5].

In this paper, in order to increase the noise immunity of modern information systems, a regular and constructive method of synthesis of full classes of linear and nonlinear systems of noisy codes was developed, which allowed to significantly increase the parametric secrecy and protect information from unauthorized access [4, 6, 21, 23].

The scientific and technical basis for the emergence of modern telecommunication networks is ensuring the maximum transmission capacity of transmission systems C with the available bandwidth of the communication line ΔF according to the Nyquist formula obtained using the theorem of V.O. Kotelnikov:

$$C = 2\Delta F \log M, \text{ [bits/sec]}, \quad (1)$$

where M - the number of discrete signal values.

This formula is valid in the absence of noise in the communication line. Practically there are interferences (noises) in the line of communication, which leads to errors in the transmission of information. Then the maximum throughput of the transmission system will be determined by Shannon's formula:

$$C = \Delta F \log \left(1 + \frac{P_p}{P_n} \right), \text{ [bits/sec]}, \quad (2)$$

where P_p , P_n - average code power and noise.

An important requirement for digital transmission systems is to provide maximum noise immunity to systems, which also depend on the type of modulation being used (the best noise immunity provides multi-positional frequency manipulation). In order to increase the noise immunity of transmission systems, noise-proof encoding is used, in which an unnecessary checking bits are added to the information message to correct errors, which, however, leads to the expansion of the signal spectrum. The most commonly used block codes are BCH (Bose–Chaudhuri–Hocquenghem), Reed-Solomon codes, and continuous convolutional codes decoded by the Viterbi algorithm [7, 8, 20, 22].

II. FORMULATION OF THE PROBLEM

Especially interesting is the combination of methods of encoding and encryption. It can be argued that, encoding is an elementary encryption, and encryption is an elementary noise-free encoding. The development and implementation of such universal methods - the perspective of modern information systems.

The feature of the noisy codes is that they create a noise-like spectrum of the code sequence (most closely related to the pseudorandom sequence), and their mutual correlation is minimal. The best code for creating a noisy sequence is the Barker code, but it has a lot of redundancy. To reduce it, we will build the noisy codes based on the Golomb rulers [13, 14].

A useful feature of systems with a noisy signal is their high confidentiality and noise immunity, especially to narrowband noise. The basis of the technology of noisy codes is the use in the communication channel for transferring of information of several implementations of these codes, the separation of which at reception is carried out by selecting their sequence.

At the same time, the confident identification of such codes can be obtained by introducing redundancy, that is,

when using for sending messages of significantly excessive sequence than the transmitted message takes.

The advantage of a noisy code is the ability to apply a new kind of selection - by sequencing. An interesting feature of systems with noise-coded codes is its adaptive properties - with a decrease in the number of interferences, noise immunity increases.

The disadvantage is the transition to a more complex carrier of information, which naturally leads to the known complication of messaging systems.

III. SOLVING THE PROBLEM

The weak place of many coding systems is the statistical weakness of the code, that is, by analyzing statistics for a certain period, one can think of what the system is and then act more directionally. That is, the key search time is sharply reduced. This system operates with noisy codes, which by their properties, including statistical, is almost identical to Gaussian white noise.

Properties of these sequences:

- in each period of the sequence number 1 and 0 differs by no more than one;
- among groups of consecutive 1 and 0 in each period, the half has a duration of one character, the fourth part has a duration of two characters, the eighth part has a duration of four characters, and so on.
- the correlation function of the sequence has a single significant peak of amplitude 1 and with all shifts is equal $1 / m$ (m - the length of the sequence).
- the correlation between vectors is calculated by the formula:

$$\rho(x, y) = \frac{A - B}{A + B}, \quad (3)$$

where A - the number of positions in which the characters of the sequences x and y coincide; B - number of positions in which the characters of the sequences x and y are different.

In mathematics, the optimal ruler or ruler of Golomb is called a set of non-negative integers, arranged in the form of divisions on an imaginary line so that the distance between any two divisions is unique. In other words, throughout the line, it is impossible to find two numbers, the difference between which would be repeated twice [1, 2].

The number of divisions on the Golomb ruler is called its order, and the greatest distance between the two divisions is its length. Sometimes the Golomb rulers are described by the distances between adjacent divisions, and not by the absolute coordinates of the divisions.

The maximum number of pairs that can be made from divisions of a ruler of order n is:

$$\binom{n}{2} = \frac{n(n-1)}{2}. \quad (4)$$

Therefore, in the canonical representation of the Golomb ruler, the least division corresponds to the zero coordinate, and the subsequent division is based on the least of two

possible distances. It is not necessary that the Golomb ruler is able to measure all distances within its length, but if so, then such a line is called perfect. However, perfect rulers exist only for orders of less than five.

Golomb ruler is called optimal if there are no shorter rulers of the same order. In other words, the ruler is called optimal if the value of its last division is minimal possible [14].

When conducting research on a sequence of elements each j -ordered pairs of numbers (p_j, q_j) ; $p_j, q_j \in \{1, 2, \dots, N\}$ the amount is matched $L_j = L(p_j, q_j)$ on a sequence of entire positive numbers (Table I):

$$L_j = L(p_j, q_j) = \sum_{i=p_j}^{q_j} k_i, \quad p_j \leq q_j \quad (5)$$

TABLE I. VALUE OF POSSIBLE AMOUNTS FOR ELEMENTS OF THE GOLOMB RULER

	q_j							
p_j	1	2	...	$l-1$	l	...	$N-1$	N
1	k_1	k_2	...	k_{l-1}	k_l	...	k_{N-1}	k_N
2		$\sum_{i=1}^2 k_i$...	$\sum_{i=1}^{l-1} k_i$	$\sum_{i=1}^l k_i$...	$\sum_{i=1}^{N-1} k_i$	$\sum_{i=1}^N k_i$
...			
$l-1$				$\sum_{i=1}^{l-1} k_i$	$\sum_{i=1}^l k_i$...	$\sum_{i=1}^{N-1} k_i$	$\sum_{i=1}^N k_i$
l					$\sum_{i=1}^l k_i$...	$\sum_{i=1}^{N-1} k_i$	$\sum_{i=1}^N k_i$
...								...
N								$\sum_{i=1}^N k_i$

Maximum number of possible L_N sums in a sequence of numbers whose values differ, is determined by the trivial dependence:

$$L_N = \frac{N(N+1)}{2}. \quad (6)$$

In the general case, the simple Golomb ruler of order N in sequence of N numbers, is sequence called $K_N = (k_1, k_2, \dots, k_l, \dots, k_N)$, on which the sum of the values of all L_N numbers are gathered starting from a given number. In a simpler version, these amounts exhaust the values of numbers of the natural series $1, 2, \dots, L_N$.

One of the practical uses of the Golomb ruler is the use of it in phase antenna arrays of radio antennas, for example in radio telescopes. Antennas with configuration [0 1 4 6] can be found in the base stations of cellular communication of the standard CDMA [11, 12].

We will use the Golomb rulers to generate noise codes, since the Golomb ruler should, by definition, have all the different readings, and in large quantities its length, it becomes similar to the sequence of noisy codes by their definition [14, 16].

The proposed method of constructing noise-like codes, based on the transformation of the Golomb rulers.

To construct the noisy codes using the Golomb ruler of order N multiplicity R select line with L_N numbered in increasing order cells of a one-dimensional array and fill in the informational "ones" cells whose numbers coincide with the numbers determined from the Golomb ruler. In the remaining cells, we will add "zeros".

The generated sequence of units and zeros is L_N - dimensional noise code, the cyclic shift of which the rest of the permitted combinations can be obtained.

An example of such a code is a table of code combinations, compiled using the Golomb ruler of order $N = 7$ and multiplicity $R = 1$ (Table II):

0 1 4 10 18 23 25.

Any of L_N different code combinations of the noisy code contains exactly N single characters in the same digits, which follows from the properties of the Golomb ruler. Rest $L_N - N$ code combinations of the noisy code contain zeros [14, 15].

The minimum coding distance for this noisy code is defined as:

$$d_{\min} = 2(N-2). \quad (7)$$

The number of errors that can be detected t_1 , and the number of errors that can be corrected t_2 with the help of a noisy code, is determined by the minimum code distance:

$$t_1 \leq d_{\min} - 1, \quad (8)$$

$$t_2 \leq (t_1 - 1) / 2. \quad (9)$$

TABLE II. NOISE-CODED CODES BASED ON THE GOLOMB RULER $N = 7$ AND $R = 1$

1	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	1	0	1	
1	1	1	0	0	1	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0
0	1	1	1	0	0	1	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0
1	0	1	1	1	0	0	1	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0
0	1	0	1	1	1	0	0	1	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0
0	0	1	0	1	1	1	0	0	1	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0
0	0	0	1	0	1	1	1	0	0	1	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0	0	0	0	1
1	0	0	0	0	1	0	1	1	1	0	0	0	1	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0
0	1	0	0	0	0	1	0	1	1	1	0	0	1	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0
0	0	1	0	0	0	0	1	0	1	1	1	0	0	1	0	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0
0	0	0	0	1	0	0	0	0	1	0	1	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0
0	0	0	0	0	1	0	0	0	0	1	0	1	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1
0	0	0	0	0	0	1	0	0	0	0	1	0	1	1	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	1
1	0	0	0	0	0	0	1	0	0	0	0	1	0	1	1	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0
0	1	0	0	0	0	0	0	1	0	0	0	0	1	0	1	1	1	0	0	1	0	0	0	0	0	0	0	0	0	0
0	0	1	0	0	0	0	0	0	1	0	0	0	0	1	0	1	1	0	0	1	0	0	0	0	0	0	0	0	0	0
0	0	0	0	1	0	0	0	0	0	0	0	1	0	0	0	0	1	0	0	0	0	0	0	0	0	1	0	0	0	0
0	0	0	0	0	1	0	0	0	0	0	0	0	1	0	0	0	0	0	0	1	0	1	1	0	0	0	0	0	1	0
1	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	1	0	1	1	0
0	1	0	0	0	0	0	1	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	1	0	1	1	0
0	0	1	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	1	0	1
0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	1	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Formulas to determine the number of errors that can be corrected t_2 or discovered t_1 using the described noisy code [14]:

$$t_1 \leq 2N - 5, \quad (10)$$

$$t_2 \leq N - 2. \quad (11)$$

In these cases, the values of the parameters L_N and N not connected with each other by any analytical dependence and can be chosen arbitrarily.

This raises the question of establishing the optimal ratio between L_N and N , for which the considered noise code gets additional benefits. The noise immunity of the noisy code increases with increasing N on condition of minimizing the length of the Golomb ruler L_N [13].

Built using the Golomb rulers, the noise-coded codes allow to detect before $2N - 5$ or fix up to $N - 2$ mistakes

The block diagram of the system of reception-transmission of information using noise-free codes is shown in Fig. 1

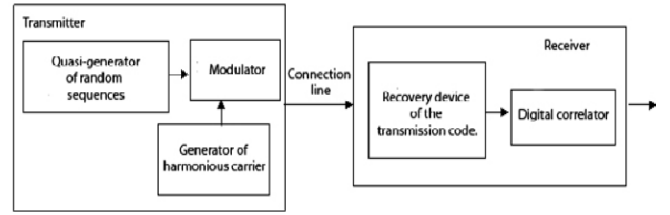


Fig. 1. Structural scheme of the system of reception and transmission of information using noisy codes

The main hardware parts of the receiver-transmitting system, which allow you to reproduce the advantages of noisy codes, is a quasi-generator of random sequences (in our case, Golomb rulers) and a digital correlator.

The quasi-generator of random sequences defines the structure of the noisy code, and the digital correlator carries out reception of an agreed pattern [9, 10, 18].

Generators of the quasi-generator of random sequences are simple in hardware execution. We can say that the generators of the noisy code do not cause difficulties during hardware implementation, and the noisy codes themselves have good potential opportunities for improving the transmission and transmitting paths [19, 20].

A program for encoding and decoding software has been developed with error correction by means of noise sequences where it is necessary to specify:

- input data (elements of a noisy sequence);
- number of errors that are found and corrected;
- path to a file that needs to be coded and decode based on a noisy sequence.

IV. CONCLUSIONS

Noisy codes relate to a set of extremely irregular branched structures. The basic concepts of theory are still in the process of formation and development, but the field of their application is continuously expanding.

The great interest in these codes is due to the fact that their analogs, such as Barker quasi-codes, the Golomb ruler, numerical ring-bundles are used in real tasks, and in typical, and not in exotic situations.

The research of various types of noisy code sequences indicates the benefits of those that are synthesized based on the Golomb rulers, which allows to achieve greater crypto stability and noise immunity when converting information compared to classical noisy code sequences.

The algorithm and program of the simplified synthesis of noise immune noisy code sequence based on the Golomb rulers and the creation of an efficient coding and decoding algorithm are developed.

Studies show that the use of noisy code sequences based on the Golomb rulers in information conversion tasks ensures simplicity of hardware application.

REFERENCES

- [1] channels for UAV based on the generalized binary Barker sequences," 2013 IEEE 2nd International Conference Actual Problems of Unmanned Air Vehicles Developments Proceedings (APUAVD), Kiev, pp. 99-103, 2013. doi: 10.1109/APUAVD.2013.6705296.
- [2] C. R. Lakshmi, D. Trivikramarao, S. Subhani and V. S. Ghali, "Barker coded thermal wave imaging for anomaly detection," 2018 Conference on Signal Processing And Communication Engineering Systems (SPACES), Vijayawada, India, pp. 198-201, 2018. doi: 10.1109/SPACES.2018.8316345.
- [3] D. Zerbino and Y. Kynash, "Usage the parabola function instead of sine accelerates signal processing," 2015 Xth International Scientific and Technical Conference "Computer Sciences and Information Technologies" (CSIT), Lviv, pp. 192-194, 2015. doi: 10.1109/STC-CSIT.2015.7325463.
- [4] I. Dronyuk, M. Nazarkevych, O. Fedevych, "Synthesis of Noise-Like Signal Based on Ateb-Functions," In: Vishnevsky V., Kozyrev D. (eds) Distributed Computer and Communication Networks. DCCN 2015. Communications in Computer and Information Science, vol 601. Springer, Cham., pp. 132-140, 2016. doi: 10.1007/978-3-319-30843-2_14.
- [5] G. Dua and R. Mulaveesala, "Applications of barker coded infrared imaging method for characterisation of glass fibre reinforced plastic materials," in Electronics Letters, vol. 49, no. 17, pp. 1071-1073, August 15 2013. doi: 10.1049/el.2013.1661.
- [6] H. Bae, J. Kim and J. Burm, "The enhancement of Signal-to-Noise Ratio of SAW tags using 5-bit Barker code sequence," The 40th European Microwave Conference, Paris, pp. 49-52, 2010. doi: 10.23919/EUMC.2010.5616316.
- [7] J. Fu, G. Wei and Q. Huang, "Barker coded excitation using LFM carrier for improving axial resolution in ultrasound imaging," 2013 ICME International Conference on Complex Medical Engineering, Beijing, pp. 150-153, 2013. doi: 10.1109/ICME.2013.6548229.
- [8] J. S. Jeong, "Suppression of therapeutic interference by using Barker code and adaptive notch filtering for real-time HIFU surgery and ultrasound imaging," in Electronics Letters, vol. 49, no. 14, pp. 871-873, July 4 2013. doi: 10.1049/el.2013.0418.
- [9] J. Soba, A. Munir and A. B. Suksmono, "Barker code radar simulation for target range detection using software defined radio," 2013 International Conference on Information Technology and Electrical Engineering (ICITEE), Yogyakarta, pp. 271-276, 2013. doi: 10.1109/ICITEE.2013.6676251.
- [10] J. Zhu et al., "Detection of scatters motion induced by mechanical vibrator using 7-chip barker-coded excitation," 2014 7th International Conference on Biomedical Engineering and Informatics, Dalian, pp. 51-55, 2014. doi: 10.1109/BMEI.2014.7002741.
- [11] M. Kellman, F. Rivest, A. Pechacek, L. Sohn and M. Lustig, "Barker-Coded node-pore resistive pulse sensing with built-in coincidence correction," 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), New Orleans, LA, pp. 1053-1057, 2017. doi: 10.1109/ICASSP.2017.7952317.
- [12] N. Liu and C. Peng, "Barker and m-Sequence Auto-Correlation Reception," 2009 First International Conference on Information Science and Engineering, Nanjing, pp. 2563-2565, 2009. doi: 10.1109/ICISE.2009.359.
- [13] O. Riznik, I. Yurchak, E. Vdovenko and A. Korchagina, "Model of stegosystem images on the basis of pseudonoise codes," 2010 Proceedings of Vth International Conference on Perspective Technologies and Methods in MEMS Design, Lviv, pp. 51-52, 2010.
- [14] O. Riznyk, B. Balych and I. Yurchak, "A synthesis of barker sequences is by means of numerical bundles," 2017 14th International Conference The Experience of Designing and Application of CAD Systems in Microelectronics (CADSM), Lviv, 2017, pp. 82-84. doi: 10.1109/CADSM.2017.7916090.
- [15] O. Riznyk, I. Yurchak and O. Povshuk, "Synthesis of optimal recovery systems in distributed computing using ideal ring bundles," 2016 XII International Conference on Perspective Technologies and Methods in MEMS Design (MEMSTECH), Lviv, 2016, pp. 220-222. doi: 10.1109/MEMSTECH.2016.7507545.
- [16] Pilsu Kim, Eunji Jung, Sua Bae, Kangsik Kim and Tai-kyong Song, "Barker-sequence-modulated golay coded excitation technique for ultrasound imaging," IEEE International Ultrasonics Symposium (IUS), Tours, pp. 1-4, 2016. doi: 10.1109/ULTSYM.2016.7728737.
- [17] R. C. Nilawar and D. M. Bhalerao, "Reduction of SFD bits of WiFi OFDM frame using wobulation echo signal and barker code," 2015 International Conference on Pervasive Computing (ICPC), Pune, pp. 1-3, 2015. doi: 10.1109/PERVASIVE.2015.7087095.
- [18] S. König, M. Schmidt and C. Hoene, "Precise time of flight measurements in IEEE 802.11 networks by cross-correlating the sampled signal with a continuous Barker code," 7th IEEE International Conference on Mobile Ad-hoc and Sensor Systems (IEEE MASS 2010), San Francisco, CA, pp. 642-649, 2010. doi: 10.1109/MASS.2010.5663785.
- [19] S. M. Omar, F. Kassem, R. Mitri, H. Hijazi and M. Saleh, "A novel barker code algorithm for resolving range ambiguity in high PRF radars," 2015 European Radar Conference (EuRAD), Paris, pp. 81-84, 2015. doi: 10.1109/EuRAD.2015.7346242.
- [20] Siti Julia Rosli1, Hasliza Rahim, Ruzelita Ngadiran, K. N. Abdul Rani, Muhammad Imran Ahmad, and F.h Wee, "Design of Binary Coded Pulse Trains with Good Autocorrelation Properties for Radar Communications," 2018 MATEC Web of Conferences, 2018. doi: 10.1051/mateconf/201815006016.
- [21] R. Tkachenko, P. Tkachenko, I. Izonin, and Y. Tsymbal, "Learning-based image scaling using neural-like structure of geometric transformation paradigm," In Studies in Computational Intelligence, vol. 730, pp. 537-565, 2018. Springer Verlag. doi: 10.1007/978-3-319-63754-9_25.
- [22] Y. Chunhong and L. Zengli, "The Superiority Analysis of Linear Frequency Modulation and Barker Code Composite Radar Signal," 2013 Ninth International Conference on Computational Intelligence and Security, Leshan, pp. 182-184, 2013. doi: 10.1109/CIS.2013.45.
- [23] Y. Tsymbal and R. Tkachenko, "A digital watermarking scheme based on autoassociative neural networks of the geometric transformations model," IEEE First International Conference on Data Stream Mining & Processing (DSMP), Lviv, pp. 231-234, 2016. doi: 10.1109/DSMP.2016.7583547.

Unsupervised Real-Time Stream-Based Novelty Detection Technique

An Approach in a Corporate Cloud

Anna Vergeles
Cloud Operations
Oracle
Kharkiv, Ukraine
anna.vergeles@oracle.com

Dmytro Prokopenko
Cloud Operations
Oracle
Kharkiv, Ukraine
dmytro.p.prokopenko@oracle.com

Alexander Khaya
Cloud Operations
Oracle
Kharkiv, Ukraine
alexander.khaya@oracle.com

Nataliia Manakova
Cloud Operations
Oracle
Kharkiv, Ukraine
nataliia.manakova@oracle.com

Abstract—A highly loaded cloud application environment requires the highest stability and operability, generates large telemetry data streams. These are obvious and actual prerequisites to develop a workload shift detector for the failures prevention aim. Having studied the previous works, the authors developed an approach to the detection of changepoints based on the specific conditions of the streaming telemetry data. The simulation of data center workload has allowed us to generate telemetry data under specific workload, thus we can evaluate the performance of the detector under various conditions. The conducted experiment has shown the viability of the proposed approach as well as directions for further study and improvement.

Keywords — *high-load, cloud, SaaS, telemetry, sensors, logs, real-time, monitoring, streaming data, changepoint, novelty, anomaly detection, unsupervised*

I. INTRODUCTION

Oracle Field Service Cloud (OFSC) is a high-load cloud-based mobile workforce management application distributed as Software-as-a-Service with strict Service License Agreement. SLA imposes restrictions to mean time to repair, which is the main reason why early detection of abnormal behavior is crucial for us.

OFSC's environment consists of hundreds of servers with different roles in several data centers across the world. Most common roles are front ends, back ends, databases, storages, etc. Servers within the same cluster with a certain role have similar operating mode while their belonging to different roles almost surely implies working in different modes.

Actually, the cloud architecture is not constant. Increasing number of customers results in scaling and allocating a larger number of hosts in some clusters. Meanwhile, a number of clusters and their role may vary. OFSC as application is being continuously changing and improving.

Whereas OFSC is a cloud-based application, Cloud Operations department performs continuous service monitoring at all the levels of the cloud hierarchy (IaaS,

PaaS, SaaS), as well as monitoring of client experience for a variety of functional tasks of the service. Operations produce a significant volume of data: continuous telemetry and application logs both “as-is” and aggregated.

One promising approach to high accessibility and reliability in cloud technology is an analysis of this collected technical and workload parameters in order to identify novel patterns. Once identified, this novel abnormal working behavior should be reported to operations teams that subsequently perform appropriate procedures. In that meaning, abnormal events are events that do not conform to expected patterns. Each discrepancy may be considered an anomaly.

Types and specifics of anomalies do not usually repeat, considering support system of our application working around the clock – they fix and solve emerging incidents. Accordingly, almost each time we deal with a novelty (anomaly), or unseen before operating mode. Supervised learning techniques are not applicable - there is no need in the spot-on detection of expected or scheduled situations like virtual machine restart, and at the same time unplanned situations usually occur only once – each issue results in a set of preventive measures.

The main task of our research is to propose and implement an approach to early detection of novelties and anomalies in a real-working cloud service based on unsupervised learning as a mechanism to watch for shifts in regular workload.

II. REVIEW (RELATED WORKS)

The cloud technologies growth and the necessity of remote hardware monitoring on each of the hierarchical levels as IaaS, PaaS, SaaS, are producing a huge amount of data generated endlessly and in real-time. Processing of such data can be carried out in streaming mode (often called online processing) or in batch mode, which both have its pros and cons [1-5].

One of the important issues of data processing in the streaming mode is the necessity of high-speed computations, especially for machine learning tasks in general and for

detecting anomalies in particular, where model training requires several passes and it takes considerable time. Taking into account the peculiarities of our conditions, in order to reduce time delays for learning models in real-time, we focus on partially real-time anomaly detection algorithms, which have initial non-real-time learning phase [6, 7].

Training some models for anomaly detection tasks can be performed in supervised, unsupervised, semi-supervised modes. Since we cannot afford pre-labeling the data, moreover, it makes no sense in constantly changing cloud environment, for our approach, we choose an unsupervised mode.

In scientific papers, we considered the widest range of models suitable for this mode of machine learning, which have specific features for a different type of data stream [8–10].

Besides, it should be highlighted, in many cases, the behavior of the system can change over time, due to well-known problem named concept drift. For example, reconfiguration and upgrading may influence system behavior, so models must adapt to a new definition of “normal” in an unsupervised, automatic way [11, 12].

Individual anomaly rate computed by individual models can be combined in some ensemble technique (e.g., averaging or voting) to form a final score. Such approach presented in many studies [13–17]

Thus, after studying related works, the solution of the studied problem was carried out in the direction of real-time processing of streaming data through formation of an ensemble of models based on non-real-time unsupervised learning techniques.

III. METHODOLOGY

This research deals with the data represented as either time series from multiple univariate sensors or aggregated panel data. The main approach is to split the whole process into historic learning part (non-real-time) and real-time pipeline. We mine historical data for expected patterns or states using historic learning module and right after that, in real-time we compare observed mode with the computed expected mode.

A. Data source

Our study is conducted in the field of delivering cloud-based SaaS application, so we have two main types of time series data sources:

- Analogue signals from multiple sensors assembled as snapshots – telemetry, or monitoring data, collected at regular intervals for IaaS, PaaS and SaaS levels.
- Discrete signals without consistent spacing, i.e. data from system and/or application logs. This type of data requires performing pre-processing aggregations.

Those data sources are, in fact, nothing more than infinite data streams of large volume and variety that arrive at the rapid rate – a typical streaming data with its inherent aspects of the concept drift and real-time delivery.

B. Prehandling data

Monitoring checks have different data acquisition intervals depending on a check type. At the historic processing stage, we apply last observation carried forward conversion for these non-uniformly spaced telemetry time series, assuming the last observed value to be valid within the whole time interval. This approach results in alignment of observations on a time grid with step $\Delta\tau$. Generally, at each timestamp in a time grid, τ_i , every metric can be represented as a vector of numeric and factor values:

$$V_{\tau_i} = \langle \text{TIME}; \text{OBJECT}; \text{GROUP}; \text{METRIC}; \text{VALUE}; \text{TAGE} \rangle$$

Furthermore, whereas we have historical data forcibly structured by a time grid, it makes sense to apply the same spacing to future observations to get equidistant points of making decision.

C. Ensemble of models and novelty detector

Our cloud-based service is rapidly changing in response to changes in customers’ business processes, implementing new application features, etc. Concept drift is a common phenomenon when data distribution changes over time. Any trained model has limited time when it makes accurate predictions. Soon enough, a previously trained model cannot accurately represent data behavior and requires retraining.

The best approach to overcome concept drift is to timely adjust expectations to a current state of a constantly changing system mode. It is important to emphasize that in anomaly detection being more adaptive and flexible should not mean being less sensitive.

In order to adapt to changes in concepts and at the same time to keep an accurate model, we use an ensemble of models to implement novelty detector. With this approach, we calculate anomaly rates independently by each model for each metric for each observation. Among other things, ensemble gives the advantage of uncorrelated errors.

Since we are only discussing a pilot approach to detecting anomalies, our goal is not to select the optimal set of detectors or to find the optimal weights for selected detectors, but rather to focus on a process of detection under certain conditions. We choose several models as inputs to the ensemble, listed below and illustrated in Fig 1:

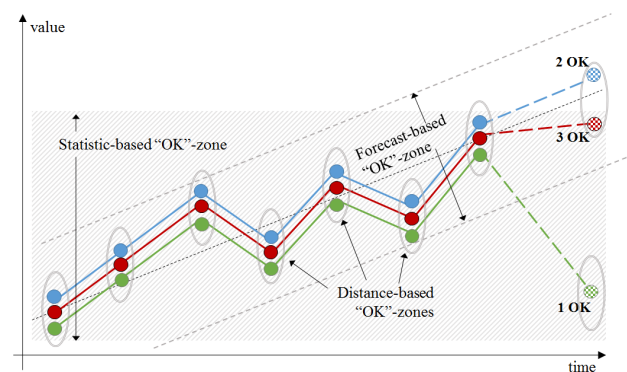


Fig. 1. Basic principle of the proposed novelty detection ensemble

1. Forecast-based (Fb) model. It estimates how big the normalized distance is between the observed data and the value predicted by some forecast model in the non-real-time module for each metric. For our task, we use SSA (Singular spectrum analysis [18]) prediction that is quite accurate for

our type of time series. In this model, alert rate rises as soon as an observed value significantly deviates from its prediction.

2. Statistic-based (Sb) model supplements forecast-based model but works independently. The model collects basic statistics from known historical modes and works well for situations when for some metrics graph of the regular has a persistent slope. This model detects trend’s permanent growth or recession, which can lead to absolute resource consumption.

3. Distance-based (Db) model takes into account distance from each observed value x_i to every other value in a group for a particular metric in a current time slot. It will alert when there are no other values besides x_i within the certain neighborhood.

D General approach

During the processing of historical data, all models that are part of the ensemble undergo the automatic parameters tuning phase. This set of mutable parameters plays a significant role in the real-time phase. Each parameter is fed in a separate stream and is matched with corresponding observation to calculate anomaly rate. Parallel processing makes our system highly configurable and, what is especially important, fast – there are no complex computations in the real-time phase.

Total anomaly score is calculated in two steps. At first, for each metric we aggregate $S(x)$ as a weighted combination of Fb rate, Sb rate and Db rate. Afterwards, we combine all $S(x)$ into single total anomaly score for the whole system as a harmonic mean. Pre-trained set of parameters expires over a certain period. In our experiment, this period is set to one hour. After expiration, parameters undergo new automatic tuning phase. The general simplified algorithm is presented in Fig. 2.

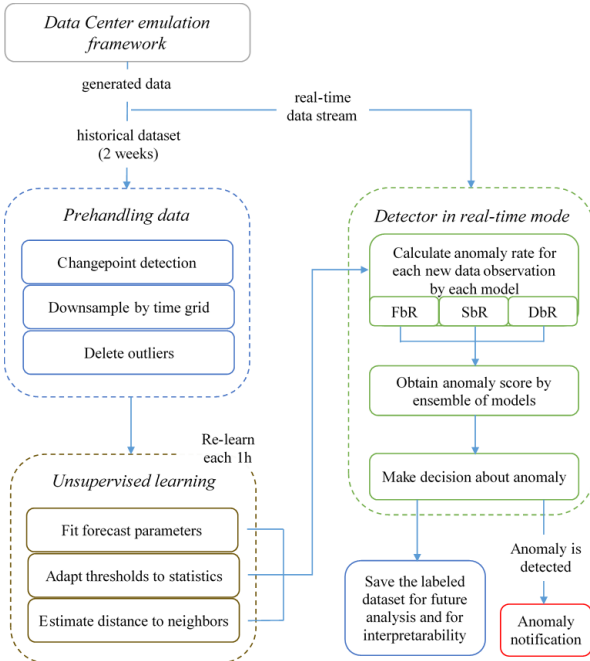


Fig. 2. Simplified algorithm

IV. EXPERIMENTAL STUDY

This section explains the experimental framework for simulating of workload and to approbating proposed novelty detection approach. During the experiment, we generated several data streams of historic and monitoring.

A Experimental framework.

We conducted an experiment within an internal local environment that was created specifically for this study. The environment consists of several virtual machines (VM) to simulate part of data center environment: two front ends (VM1 and VM2) that share single balancer node, several nodes that produce synthetically generated load, and three additional nodes, detailed below, that form an anomaly detection cluster assembled with several open source technologies.

Nodes in the anomaly detection cluster are: queue node, which handles streaming data; real-time and non-real-time modules both run on a compute node; database node serves for storing results. Parallelization for real-time and non-real-time modules has been limited to four threads per node for this experiment.

Queue and database nodes both work in micro batches mode, though this is just a matter of configuration from the point of view of the expected workload.

Virtual machines VM1 and VM2 are frontends that receive emulated workload with synthetically generated API requests. The workload is distributed by working hours to produce a realistic data set. Besides, some synthetic tasks run in a background to produce small load spikes.

We have full control over the API emulation process. Generated workload profile allows us to generate a response stream with a required mode of monitoring data, thus we can evaluate the performance of our detector under certain pre-designed conditions.

For four days prior to our experiment, we have generated a regular profile of disturbing signals to allow history module to adjust to “usual” regime when the experiment will begin. We have split our experiment into two stages to receive two streaming sets of monitoring data. In this way, monitoring data generated under specific workload patterns will represent different workload conditions.

The first scenario assumes both VMs workloads are in stable mode. Numbers of API requests per each period are set according to a regular profile, so all monitoring indicators should remain stable.

The second scenario emulated unstable workload mode is more complex. A number of requests to VM1 and VM2 at first drops to zero for half an hour and for the next half an hour increases in more than two times, if compared to a regular load for this time slot. Besides, part of the time after 17:45 VM2 has been switched off from balancing. This resulted in a slightly higher load to VM1, not anomalous, though. All changes were instant. A number of requests for both scenarios are shown in Fig. 3. Grey area at Stage 2 lies within 16:30-17:30 time interval, when we provoke our VMs to response with a shift in workload as well as expect our system to detect anomalous behavior.

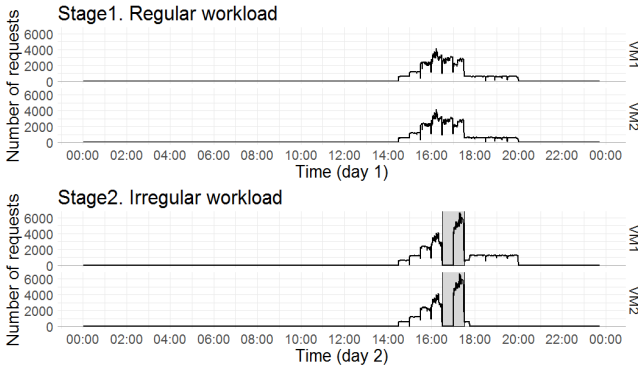


Fig. 3. Number of requests per minute

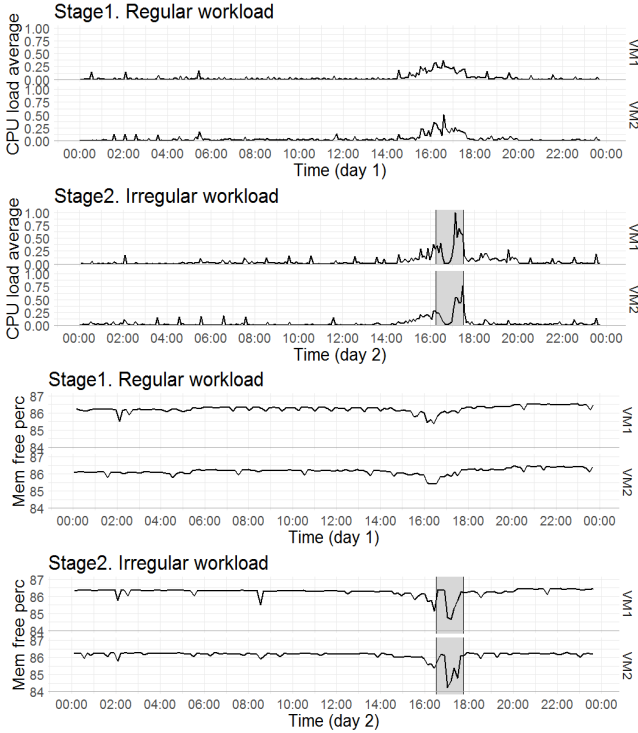


Fig. 4. Corresponding telemetry data

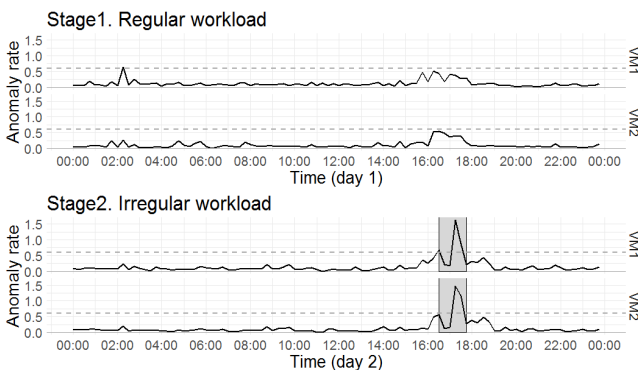


Fig. 5. Convolution of anomaly rate

B Generated data stream.

In our experiment, we limit a list of possible metrics to two telemetry time series, namely CPU load average (CPU_la) and free memory percentage (FMp). As mentioned above, our purpose is not to find the best metrics but rather to prove the concept.

The periodicity of observed values is 5 minutes for CPU_la and 9 minutes for FMp. For the experiment, we downsampled time series to 15 minutes intervals. The unsupervised historic module produces three streams according to the number of ensemble inputs (F_b, S_b, D_b). The fourth stream is the stream of the observed real-time telemetry (monitoring) data. Merged together in a real-time module, all streams form a fast parallel transformation pipeline that calculates anomaly score and forms an output stream of scores along with some interim calculations. This output stream sinks to a database for visualization, further analysis and interpretability.

In such a manner, we obtained data implemented as 4 data streams with numerous instances. Each time historic module processes over 1000 observations to produce forecast and statistics for the next 4 points. Snapshots of observed system behavior and calculated anomaly rates are shown in Fig. 4 and Fig. 5 correspondingly. Grey area on Figure 4 (Stage 2, both metrics) is a period of time when an expert sees some anomalies in CPU_la of FMp (16:15-17:30) if compared to regular behavior (Stage 1). Grey area on Figure 5 (16:30-17:45) is a time slot when anomaly is expected to be detected.

C Manually labeled classes.

Since we are pursuing to develop a tool for conducting an unsupervised anomaly detection, we do not need any previously labeled anomalies to implement our detector. However, to estimate the accuracy of the proposed detection algorithm, two classes of labels were applied: a class label False (no anomaly) is given by default, class label True (some anomalies) is assigned based on expert opinion as described above. It is necessary to emphasize once again that these labels are not used in the run-time of our anomaly detection algorithm.

V. RESULTS AND ESTIMATION OF THE APPROACH

For decision-making purposes whether the changes in mode denote a novelty, we have chosen the threshold to be equal 0.6, which stands for “approximately two-thirds of all diagnostics should point out that anomalous rates have been detected”.

It should be noted, our experiment missed the drop in CPU load caused by the drop in requests designed to be one kind of novelty, but it perfectly revealed unusual behavior expressed in the form of increased resource usage (see Fig. 4).

F1 score estimated for this experiment is 0.625 (precision is 0.833, recall is 0.5). Despite the fact that the estimates are not higher than 0.7, the conducted experiment has shown the viability of our approach and has highlighted directions for further study and improvement. More details on this one can find in the next section.

VI. CONCLUSIONS AND FUTURE WORK

The main contributions of the conducted study are:

- (1) A new modular approach to novelty detection has been developed based on the main idea of ensemble model training with splitting into real-time and non-real-time modules

(2) Experimental research has been implemented on the data center simulation framework under different scenarios of workload within the proposed method.

(3) Performance assessment, including precision and recall, of the proposed approach under experimental conditions shows an appropriate performance in both indicators.

Nevertheless, it should be noted that the current study has been carried out by a numerical experiment, this causes some limitations. In particular, it is advisable to conduct an advanced research of some theoretical issues:

- What kind of machine learning models should be included in the ensemble detector at the training stage on historical data to improve a quality of real-time detection?
- What kind of anomaly score function will give the best assessment of accuracy and sensitivity for the proposed approach?
- What length of historical data will be the best?
- Which frequency is optimal for retraining the models (once a day, every 3 hours or even dynamically chosen interval, instead of a predefined interval)?

Besides, the proposed approach, suitable for the streaming data anomaly detection, has been successfully validated but has not been compared with any of the existing methods. A comprehensive study of performance of the proposed approach versus other methods would be a good next step in this research.

ACKNOWLEDGMENT

We thank Oracle Corporation for supporting this research project and especially Maryna Chukhray and Andriy Rabochiy for their discussions, feedback and helpful suggestions.

Safe Harbor Statement. The following is intended for information purposes only, and may not be incorporated into any contract. It is not a commitment to deliver any material, code, or functionality, and should not be relied upon in making purchasing decisions. The development, release, and timing of any features or functionality described for Oracle's products remains at the sole discretion of Oracle.

REFERENCES

[1] M. Dias de Assunção, A. da Silva Veith, and R. Buyya, "Distributed data stream processing and edge computing: A survey on resource elasticity and future directions," *Journal of Network and Computer Applications*, vol. 103, pp. 1–17, 2018.

[2] M. Harvan, T. Locher, and A. C. Sima, "Cyclone: Unified Stream and Batch Processing," in 2016 45th International Conference on Parallel Processing Workshops (ICPPW), Philadelphia, PA, USA, pp. 220–229, 2016.

[3] W. Li, D. Niu, Y. Liu, S. Liu, and B. Li, "Wide-Area Spark Streaming: Automated Routing and Batch Sizing," *IEEE International Conference on Autonomic Computing (ICAC)*, Columbus, OH, USA, pp. 33–38, 2017.

[4] K. Vidyasankar, "On Atomic Batch Executions in Stream Processing," *Procedia Computer Science*, vol. 98, pp. 72–79, 2016.

[5] C. Klein, B. Donnellan, and M. Helfert, Eds., *Correlation-Model-Based Reduction of Monitoring Data in Data Centers*, Setúbal: SCITEPRESS - Science and Technology Publications Lda, 2016.

[6] P.-Y. Chen, S. Yang, and J. A. McCann, "Distributed Real-Time Anomaly Detection in Networked Industrial Sensing Systems," *IEEE Trans. Ind. Electron.*, vol. 62, no. 6, pp. 3832–3842, 2015.

[7] S. Y. Shin and J. C. Maldonado, Novelty detection algorithm for data streams multi-class problems. Coimbra, Portugal, ACM, March 18–22, 2013.

[8] D. Hong, D. Zhao, and Y. Zhang, "The Entropy and PCA Based Anomaly Prediction in Data Streams," *Procedia Computer Science*, vol. 96, pp. 139–146, 2016.

[9] C. C. Olson, K. P. Judd, and J. M. Nichols, "Manifold learning techniques for unsupervised anomaly detection," *Expert Systems with Applications*, vol. 91, pp. 374–385, 2018.

[10] Sajjad Kamali Siahroudi, Poorya Zare Moodi, and Hamid Beigy, "Detection of evolving concepts in non-stationary data streams: A multiple kernel learning approach," *Expert Systems With Applications*, pp. 187–197, 2018.

[11] S. Ahmad, A. Lavin, S. Purdy, and Z. Agha, "Unsupervised real-time anomaly detection for streaming data," *Neurocomputing*, vol. 262, pp. 134–147, 2017.

[12] M. Tennant, F. Stahl, O. Rana, and J. B. Gomes, "Scalable real-time classification of data streams with concept drift," *Future Generation Computer Systems*, vol. 75, pp. 187–199, 2017.

[13] E. Yu and P. Parekh, "A Bayesian Ensemble for Unsupervised Anomaly Detection," [Online] Available: <http://arxiv.org/pdf/1610.07677v1>.

[14] B. Krawczyk, L. L. Minku, J. Gama, J. Stefanowski, and M. Woźniak, "Ensemble learning for data stream analysis: A survey," *Information Fusion*, vol. 37, pp. 132–156, 2017.

[15] Z. Ding and M. Fei, "An Anomaly Detection Approach Based on Isolation Forest Algorithm for Streaming Data using Sliding Window," *IFAC Proceedings Volumes*, vol. 46, no. 20, pp. 12–17, 2013.

[16] A. Bifet, G. Holmes, B. Pfahringer, R. Kirkby, and R. Gavaldà, "New ensemble methods for evolving data streams," 15th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '09, Paris, France, p. 139, 2009.

[17] K. Noto, C. Brodley, and D. Slonim, "Anomaly Detection Using an Ensemble of Feature Models," (eng), *IEEE International Conference on Data Mining*, pp. 953–958, 2010.

[18] N. Golyandina and A. Korobeynikov, "Basic Singular Spectrum Analysis and forecasting with R," *Computational Statistics & Data Analysis*, vol. 71, pp. 934–954, 2014.

Data Stream Online Clustering Based on Fuzzy Expectation-Maximization Approach

Anastasiia O. Deineko
Artificial Intelligence Department
Kharkiv National University
of Radio Electronics
Kharkiv, Ukraine
anastasiya.deineko@gmail.com

Oleksandr O. Zayika
Artificial Intelligence Department
Kharkiv National University
of Radio Electronics
Kharkiv, Ukraine
zaiqa.alexander@gmail.com

Polina Ye. Zhernova
System Engineering Department
Kharkiv National University
of Radio Electronics
Kharkiv, Ukraine
polina.zhernova@gmail.com

Iryna Pliss
Control Systems Research Laboratory
Kharkiv National University
of Radio Electronics
Kharkiv, Ukraine
iryana.pliss@nure.ua

Boris Gordon
Computer Systems Department
Tallinn University of Technology
Tallinn, Estonia
boris.gordon@ttu.ee

Nelya Pabyrivska
Department of Mathematics
Lviv Polytechnic National University
Lviv, Ukraine
nelyapab@gmail.com

Abstract—In the paper the online fuzzy clustering recurrent procedure has been introduced that allows the forming of hyperellipsoidal clusters with an arbitrary orientation of the axes is proposed. Such clustering system is the generalization of a number of known algorithms, it is intended to solve tasks within the general problems of Data Stream Mining (DSM) and Dynamic Data Mining (DDM), when information is sequentially fed to processing in online mode.

Keywords— *big data; dynamic data mining; data stream mining; computational intelligence; EM-algorithm; fuzzy clustering; Kohonen's self-learning; soft clustering.*

I. INTRODUCTION

Clustering task is an integral part and an important direction in the global problem of Data Science and Data Mining [1, 2]. Many approaches and methods were proposed for solution of this problem [3-6]. They differ from each other in apriori assumptions, problem formulation and in used mathematical apparatus. Currently, the most intensively growing direction of Data Mining is *DSM* [7], in which data are fed to processing in the online mode, observation by observation. This direction is closely related with tasks of processing large amounts of data, so-called Big Data [8], when it is simply impossible to process increasing volume of data in the batch mode.

Computational Intelligence (*CI*) can be successfully used for many tasks of *DSM*. And first of all the methods based on soft computing and neural networks among which the Fuzzy Clustering (*FC*) methods [4, 9, 10] are the most wide spread. At the same time, the overwhelming number of known methods are oriented to the batch mode processing. And known intelligent systems of sequential data processing and first of all Kohonen's clustering neural networks, which also are known as Self-Organizing Maps [11], can solve crisp clustering tasks with the assumption of linearly separable classes.

Capabilities of crisp clustering algorithms are restricted by the fact that real data usually form overlapping classes, thus each vector-observation could belong to several classes at once, with different probabilities (or belonging) levels. In

this case soft calculations come to the fore. In the class of probabilistic methods most widely used is so-called Expectation-Maximization (*EM*) algorithm [10, 12-15]. And in the class of fuzzy methods the most popular is J.C. Bezdek's Fuzzy *C*-means (*FCM*) algorithm [9, 10]. It can be noted that in [15] the hybrid clustering algorithm has been discussed. It unites both of these approaches.

The mentioned clustering procedures which are based on soft computing are oriented to information processing only in batch mode. Naturally this fact makes usage of these methods in *DSM* possible. Note that in [16, 17] group of recursive *FC* algorithms were introduced. But clusters, what they are formed have a spherical shape. This fact limits their capabilities in situations where data form classes of an arbitrary form.

In this connection, it seems appropriate to develop recurrent procedures for probabilistic and fuzzy clusterization, which allow to process data in online mode and to form clusters of the hyperellipsoidal form with the axes of arbitrary orientation in the features space.

II. BATCH PROCEDURES FOR PROBABILISTIC AND FUZZY CLUSTERING (FC) IN THE CASE OF HYPERELLIPSOIDAL CLASSES

The batch clustering problem can be described in general case: it is assumed that the initial data array contains N multidimensional observations, which are described by vectors-features of order n $(x_1(k), \dots, x_n(k))$, $x_n(k) \in R^n$, $k = 1, 2, 3, \dots, N$ (k – number of observation in initial data array), which has to be partitioned for $m(1 < m < N)$ overlapping clusters.

In a standard *EM* approach, it is also assumed that the density of the distribution of observations in each cluster is Gaussian:

$$p_j(x) = \left((2\pi)^{\frac{n}{2}} \sqrt{\det \Sigma_j} \right)^{-1} \exp \left(-\frac{1}{2} (x - c_j)^T \Sigma_j^{-1} (x - c_j) \right),$$

$$j = 1, 2, \dots, m \quad (1)$$

where c_j – vector-centroid with order n of the j -th cluster,
 Σ_j – the correlation matrix of the j -th cluster of size $(n \times n)$:

$$\Sigma_j = \frac{1}{N} \sum_{k=1}^N (x(k) - c_j)(x(k) - c_j)^T \quad (2)$$

It is obvious that the joint density of the distribution of all data is described by the expression

$$\begin{aligned} p(x) &= \sum_{k=1}^m p_j p_j(x) = \\ &= \sum_{k=1}^m p_j \left((2\pi)^{\frac{n}{2}} \sqrt{\det \Sigma_j} \right)^{-1} \exp \left(-\frac{1}{2} (x - c_j)^T \Sigma_j^{-1} (x - c_j) \right) = (3) \\ &= \sum_{k=1}^m p_j \left((2\pi)^{\frac{n}{2}} \sqrt{\det \Sigma_j} \right)^{-1} \exp \left(-\frac{1}{2} d_M^2(x, c_j) \right) \end{aligned}$$

where p_j – a priori probabilities-weights that satisfy the obvious condition

$$\sum_{j=1}^m p_j = 1. \quad (4)$$

It can be noted that condition (4) completely coincides with the constraint on the levels of belonging of the k -th observation to the j -th cluster $u_j(k)$, which is the basis of the fuzzy c-means method

$$\sum_{j=1}^m u_j = 1. \quad (5)$$

In connection with that fact, *FCM* is sometimes called the method of fuzzy probabilistic clustering [10].

The main feature of the *EM* approach is that the exponent in (1), (3) contains the Mahalanobis distance between centroids c_j and observations $x(k)$

$$d_m^2(x(k), c_j) = (x(k) - c_j)^T \Sigma_j^{-1} (x(k) - c_j), \quad (6)$$

which allows, in contrast to *FCM* that restores spherical clusters to form classes in the form of hyperellipsoids with an arbitrary orientation of axes in the space of features.

The solution of the clustering problem in the context of the *EM* approach is related to the maximization of the log likelihood function

$$E(x(k), c_j, \Sigma_j, p_j) = \sum_{k=1}^m \log \left(\sum_{k=1}^m p_j p_j(x(k)) \right), \quad (7)$$

that leads to estimates [12]

$$\begin{cases} p_j(x(k)) = \exp \left(-\frac{1}{2} d_M^2(x(k), c_j) \right) / \sum_{i=1}^m \exp \left(-\frac{1}{2} d_M^2(x(k), c_i) \right), \\ c_j = \sum_{k=1}^N p_j(x(k)) x(k) / \sum_{k=1}^N p_j(x(k)). \end{cases} \quad (8)$$

A particular crisp version of the *EM* algorithm is the widely used k-means method coinciding with *EM* at $p_j = m^{-1}$ and identity matrix Σ_j .

K-means is based on minimizing the objective function

$$\begin{aligned} E(x(k), c_j) &= \sum_{k=1}^N \sum_{j=1}^m u_j(k) \|x(k) - c_j\|^2 = \\ &= \sum_{k=1}^N \sum_{j=1}^m u_j(k) d_E^2(x(k), c_j) \end{aligned} \quad (9)$$

where

$$u_j(k) = \begin{cases} 1, & \text{if } x(k) \in j\text{-th cluster,} \\ 0, & \text{otherwise.} \end{cases}$$

It should be noted that k-means is based on the Euclidean distance, although the method of Mahalanobis k-means is also known. It is based on the minimization of the goal function in the form

$$\begin{aligned} E(x(k), c_j) &= \sum_{k=1}^N \sum_{j=1}^m u_j(k) (x(k) - c_j)^T \Sigma_j^{-1} (x(k) - c_j) = \\ &= \sum_{k=1}^N \sum_{j=1}^m u_j(k) d_M^2(x(k), c_j). \end{aligned} \quad (11)$$

As a result of optimization (9), (11), it is not difficult to obtain estimates of the centroids coordinates in the form

$$c_j = \sum_{k=1}^N u_j(k) x(k) / \sum_{k=1}^N u_j(k) = \frac{1}{N_j} \sum_{x(k) \in u_j} x(k) \quad (12)$$

where N_j – the number of observations assigned to j -th cluster.

A generalization of crisp objective functions (9), (11) in the case of overlapping classes is fuzzy objective functions [18]

$$E(x(k), c_j, u_j) = \sum_{k=1}^N \sum_{j=1}^m u_j^\beta(k) d^2(x(k), c_j) \quad (13)$$

where β – a non-negative fuzzifier parameter. Such parameter determines the "blurring" of the boundaries

between classes (usually $\beta = 2$), $d^2(x(k), c_j)$ – estimate of Euclidean distance between $x(k)$ and c_j .

Minimization (13), taking into account the constraint (5), leads to the result [3]:

$$\begin{cases} u_j(k) = d^{1-\beta}(x(k), c_j) / \sum_{l=1}^m d^{1-\beta}(x(k), c_l), \\ c_j = \sum_{k=1}^N u_j^\beta(x(k), c_j) / \sum_{k=1}^N u_j^\beta(k), \end{cases} \quad (14)$$

that for $\beta = 2$ FCM takes the form:

$$\begin{cases} u_j(k) = \frac{d_E^{-2}(x(k), c_j)}{\sum_{l=1}^m d_E^{-2}(x(k), c_l)} = \frac{\|x(k) - c_j\|^{-2}}{\sum_{l=1}^m \|x(k) - c_l\|^{-2}}, \\ c_j = \sum_{k=1}^N u_j^2(k) x(k) / \sum_{k=1}^N u_j^2(k). \end{cases} \quad (15)$$

The first relation (14) can be easily transformed to the form

$$u_j(k) = \frac{1}{1 + \frac{d^{\beta-1}(x(k), c_j)}{\gamma_j}}, \quad (16)$$

$$\gamma_j = \left(\sum_{\substack{l=1, \\ l \neq j}}^m d^{\beta-1}(x(k), c_l) \right)^{-1},$$

corresponding to the generalized Gaussian function [19], with $\beta = 2$ we get

$$u_j(k) = \frac{1}{1 + \frac{d^2(x(k), c_j)}{\gamma_j}}, \quad (17)$$

$$\gamma_j = \left(\sum_{\substack{l=1, \\ l \neq j}}^m d^2(x(k), c_l) \right)^{-1},$$

corresponding to the Cauchy probabilities density function.

Thus, it can be noted that if the EM approach is based on the Gaussian distribution, then for fuzzy procedures, the Cauchy distribution is implicit. Among the fuzzy clustering procedures based on the objective function (13), the closest to the EM approach is the algorithm introduced in [20]. It uses as an estimate of the distance expression

$$\begin{aligned} d_{GG}^2(x(k), c_j) &= q_j (\sqrt{\det \Sigma_j})^{-1} \exp\left(-\frac{1}{2}(x-c_j)^T \Sigma_j^{-1}(x-c_j)\right) = \\ &= q_j (\sqrt{\det \Sigma_j})^{-1} \exp\left(-\frac{1}{2}d_M^2(x, c_j)\right) \end{aligned} \quad (18)$$

where

$$q_j = \sum_{k=1}^N u_j^\beta(k) / \sum_{k=1}^N \sum_{l=1}^m u_l^\beta(k) \quad (19)$$

Minimization of the (18), taking into account (5) and (19), leads to the result

$$\begin{cases} u_j(k) = d_{GG}^{1-\beta}(x(k), c_j) / \sum_{l=1}^m d_{GG}^{1-\beta}(x(k), c_l), \\ c_j = \sum_{k=1}^N u_j^\beta(k) x(k) / \sum_{k=1}^N u_j^\beta(k), \\ \Sigma_j = \sum_{k=1}^N u_j^\beta(k) (x(k) - c_j)(x(k) - c_j)^T / \sum_{k=1}^N u_j^\beta(k). \end{cases} \quad (20)$$

for $\beta = 2$ (20) becomes

$$\begin{cases} u_j(k) = d_{GG}^{=2}(x(k), c_j) / \sum_{l=1}^m d^{=2}(x(k), c_l), \\ c_j = \sum_{k=1}^N u_j^2(k) x(k) / \sum_{k=1}^N u_j^2(k), \\ \Sigma_j = \sum_{k=1}^N u_j^2(k) (x(k) - c_j)(x(k) - c_j)^T / \sum_{k=1}^N u_j^2(k) \end{cases} \quad (21)$$

close to (15) and is a generalization of FCM for the case of hyperellipsoidal clusters.

III. ADAPTIVE ONLINE PROCEDURES FOR PROBABILISTIC AND FUZZY CLUSTERING IN THE CASE OF HYPERELLIPSOIDAL CLUSTERS

Let's consider further a case when the data are fed to processing sequentially one after another in the form of a stream $x(1), x(2), \dots, x(k), x(k+1), \dots$, where k has the sense of the current discrete time. It is clear that the fuzzy clustering procedures discussed above in this case are ineffective. It is known that the optimization problem solution of the objective function (9) corresponding to the k-means method can be obtained with the help of the self-learning WTA-rule of the clustering neural network of T. Kohonen [21] in the form

$$c_j(k+1) = \begin{cases} c_j(k) + \eta(k+1)(x(k+1) - c_j(k)), \\ \text{if } c_j(k) - \text{"winner"}, \\ c_j(k) - \text{otherwise} \end{cases} \quad (22)$$

where $0 < \eta(k+1) < 1$ – the learning rate parameter chosen in the accordance with the stochastic approximation conditions.

Here it should be noted that it is possible to draw a clear analogy between self-learning according to T. Kohonen and the EM algorithm: the step of competition corresponds to the E-step (expectation), and the step of synaptic adaptation is the M-step (maximization). At the step of synaptic adaptation, a step of gradient minimization of the distance

$$d_E^2(x(k+1), c_j(k)) = \|x(k+1) - c_j(k)\|^2 \quad (23)$$

is realized, i.e., procedure (22) can be represented in the form

$$c_j(k+1) = \begin{cases} c_j(k) - \eta(k+1) \nabla_{c_j} d_E^2(x(k+1), c_j(k)), \\ \text{if } c_j(k) - \text{"winner"}, \\ c_j(k) - \text{otherwise.} \end{cases} \quad (24)$$

Similarly, the Mahalanobis metric (6), used in the EM algorithm, can be minimized [22]:

$$c_j(k+1) = \begin{cases} c_j(k) - \eta(k+1) \nabla_{c_j} d_M^2(x(k+1), c_j(k)), \\ \text{if } c_j(k) - \text{"winner"}, \\ c_j(k) - \text{otherwise,} \end{cases} \quad (25)$$

or

$$c_j(k+1) = \begin{cases} c_j(k) + \eta(k+1) \Sigma_j^{-1}(k) (x(k+1), c_j(k)), \\ \text{if } c_j(k) - \text{"winner"}, \\ \Sigma_j(k) = \frac{1}{k_j} \sum_{\tau=1}^k (x(\tau) - c_j(k)) (x(\tau) - c_j(k))^T = \\ = \Sigma_j(k-1) + \frac{1}{k_j} ((x(k) - c_j(k)) (x(k) - c_j(k))^T - \\ - \Sigma_j(k-1)), c_j(k) - \text{otherwise} \end{cases} \quad (26)$$

where the index k_j shows how many times the j -th neuron of Kohonen's SOM was the winner in the process of self-learning.

In the case of overlapping classes, procedure (26) can be supplemented by an estimate of the membership level of the first relation type (15):

$$u_j(k) = d_M^{-2}(x(k), c_j(k)) \Big/ \sum_{i=1}^m d_M^{-2}(x(k), c_i(k)) = \frac{\left((x(k) - c_j(k))^T \Sigma_j^{-1}(k) (x(k) - c_j(k)) \right)^{-1}}{\sum_{i=1}^m \left((x(k) - c_i(k))^T \Sigma_i^{-1}(k) (x(k) - c_i(k)) \right)^{-1}}. \quad (27)$$

The task of fuzzy objective function recurrent minimization of type (13) with constraint (5) reduces to the solution of the non-linear programming problem by the Arrow-Hurwitz-Uzawa procedure by optimizing the Lagrange function. In this case [16]:

$$\begin{cases} c_j(k+1) = c_j(k) + \eta(k+1) u_j^\beta(k+1) (x(k+1) - c_j(k)), \\ u_j(k+1) = d_E^{1-\beta}(x(k+1), c_j(k)) \Big/ \sum_{i=1}^m d_E^{1-\beta}(x(k+1), c_i(k)), \end{cases} \quad (28)$$

and for $\beta = 2$ FCM:

$$\begin{cases} c_j(k+1) = c_j(k) + \eta(k+1) u_j^2(k+1) (x(k+1) - c_j(k)), \\ u_j(k+1) = \|x(k+1) - c_j(k)\|^2 \Big/ \sum_{i=1}^m \|x(k+1) - c_i(k)\|^2, \end{cases} \quad (29)$$

where the factors $u_j^\beta(k+1)$ and $u_j^2(k+1)$ play the role of the neighborhood function in the WTM-rule of self-learning, instead of the traditionally used Gaussians, generalized Gaussian is used in (28), and in (29) - Cauchian, while the width parameter of these functions is given automatically.

As for the Gath-Geva algorithm [20], described by the expression (18-21), recurrent modifications were introduced in [23]. However, they are not related to optimization procedures. Thus, in [23] a simplified procedure of the form

$$\begin{cases} c_j(k+1) = c_j(k) + \eta(k+1) (x(k+1) - c_j(k)), \\ \text{if } c_j(k) - \text{"winner"}, \\ \Sigma_j(k+1) = (1 - \eta(k+1)) \Sigma_j(k) + \eta(k) * \\ * (x(k+1) - c_j(k)) (x(k+1) - c_j(k))^T \end{cases} \quad (30)$$

that is essentially a WTA-rule of self-learning, supplemented with the procedure for the correlation matrix correcting. In this case, this matrix does not influence the centroids correction process.

More flexible is the algorithm proposed in [23], where an additional variable of accumulated memberships were introduced into consideration:

$$U_j(k+1) = \sum_{\tau=1}^{k+1} u_j^\beta(\tau) = U_j(k) + u_j^\beta(k+1). \quad (31)$$

In this case, the FC procedure becomes the form

$$\begin{cases} c_j(k+1) = c_j(k) + \frac{u_j^\beta(k+1)}{U_j(k+1)} (x(k+1) - c_j(k)), \\ \Sigma_j(k+1) = U_j(k) / U_j(k+1) * \\ * \left(\Sigma_j(k) + \frac{u_j^\beta(k+1)}{U_j(k+1)} (x(k+1) - c_j(k)) (x(k+1) - c_j(k))^T \right), \\ u_j(k+1) = d_{GG}^{1-\beta}(x(k+1), c_j(k)) \Big/ \sum_{i=1}^m d_{GG}^{1-\beta}(x(k+1), c_i(k)). \end{cases} \quad (32)$$

The algorithm (32) coincides with (28) for $\eta(k+1) = U_j^{-1}(k+1)$ and differs only in the used distance estimate $d_{GG}^2(x(k+1), c_j(k))$ instead of $d_E^2(x(k+1), c_j(k))$. The correlation matrix $\Sigma_j(k)$ does not affect the process of centroids correction.

Returning to the algorithm (25) based on the gradient of the Mahalanobis distance, it is easy to introduce a recurrent version of the Gath-Geva algorithm, which is a generalization of the procedures (25-26, 28):

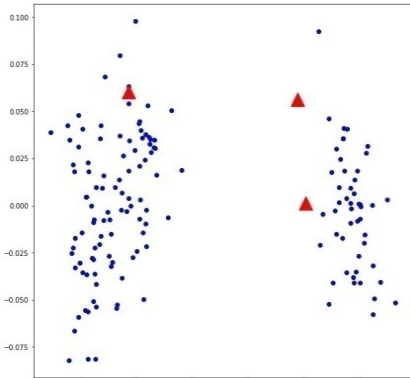
$$\begin{cases} c_j(k+1) = c_j(k) + \frac{u_j^\beta(k+1)}{U_j(k+1)} \Sigma_j^{-1}(x(k+1) - c_j(k)), \\ \Sigma_j(k+1) = \frac{U_j(k)}{U_j(k+1)} * \\ * \left(\Sigma_j(k) + \frac{u_j^\beta(k+1)}{U_j(k+1)} (x(k+1) - c_j(k))(x(k+1) - c_j(k))^T \right), \\ u_j(k+1) = d_{GG}^{1-\beta}(x(k+1), c_j(k)) / \sum_{l=1}^m d_{GG}^{1-\beta}(x(k+1), c_j(k)). \end{cases} \quad (33)$$

For $\beta = 2$ we obtain

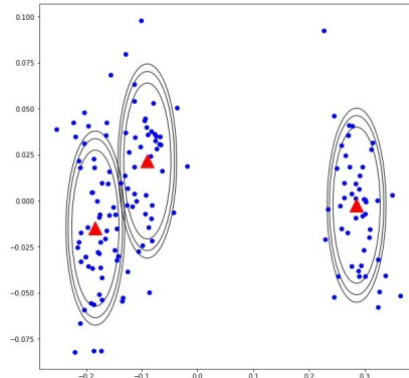
$$\begin{cases} U_j(k+1) = U_j(k) + u_j^2(k+1), \\ c_j(k+1) = c_j(k) + \frac{u_j^2(k+1)}{U_j(k+1)} \Sigma_j^{-1}(x(k+1) - c_j(k)), \\ \Sigma_j(k+1) = \frac{U_j(k)}{U_j(k+1)} * \\ * \left(\Sigma_j(k) + \frac{u_j^2(k+1)}{U_j(k+1)} (x(k+1) - c_j(k))(x(k+1) - c_j(k))^T \right), \\ u_j(k+1) = d_{GG}^{-2}(x(k+1), c_j(k)) / \sum_{l=1}^m d_{GG}^{-2}(x(k+1), c_j(k)). \end{cases} \quad (34)$$

IV. RESULTS OF EXPERIMENTS

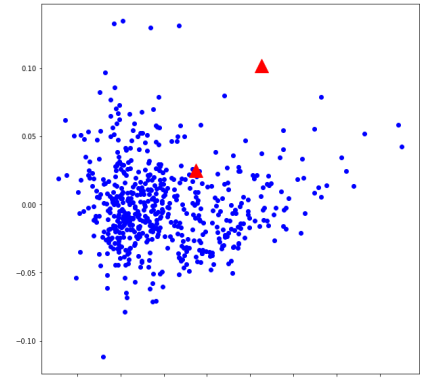
Tree data set from *UCI Machine Learning Repository* [24] are used in the experimental analysis. The information about used data sets is shown in Table I.



a) Initial centroids coordinates "Iris"



c) Final centroids coordinates "Iris"



d) Initial centroids coordinates "WDBC"

TABLE I. THE DESCRIPTION OF THE DATA SETS

Data sets	Properties		
	Attributes	Classes	Number of samples
Iris	4	3	150
WDBC	30	2	569

The performance of described in these paper systems were compared in series of experiments. For comparison of proposed soft clustering system, the standard *FCM* algorithm and standard *EM* algorithm were taken. The clustering accuracy of proposed soft clustering system was measured and compared with *FCM* and *EM* algorithms. The clusterization results were shown in Table II. These clustering results of the proposed soft clustering system, *FCM* algorithm and standard *EM* algorithm were estimated using the well-known *Xie-Benlie criterion* for fuzzy clustering. From Table II easy to see that proposed soft clustering system demonstrated a better performance of clustering quality. The changes centroids coordinates from initial initialization to the final iteration are shown at the Fig. 1.

TABLE II. THE MEAN CLUSTERING ACCURACIES OF THE COMPARED ALGORITHMS

Algorithms for comparison	Clustering accuracies	
	Iris	WDBC
FCM	0,82	0,86
EM	0,84	0,85
Soft clustering system	0,89	0,90

V. CONCLUSION

The online fuzzy clustering problem was considered. The recurrent procedure has been introduced that allows the forming of hyperellipsoidal clusters with an arbitrary orientation of the axes. The proposed procedure is the generalization of a number of known algorithms, it is quite simple in computational implementation and is intended to solve tasks within the general problem of Data Stream Mining, when information is sequentially fed to processing.

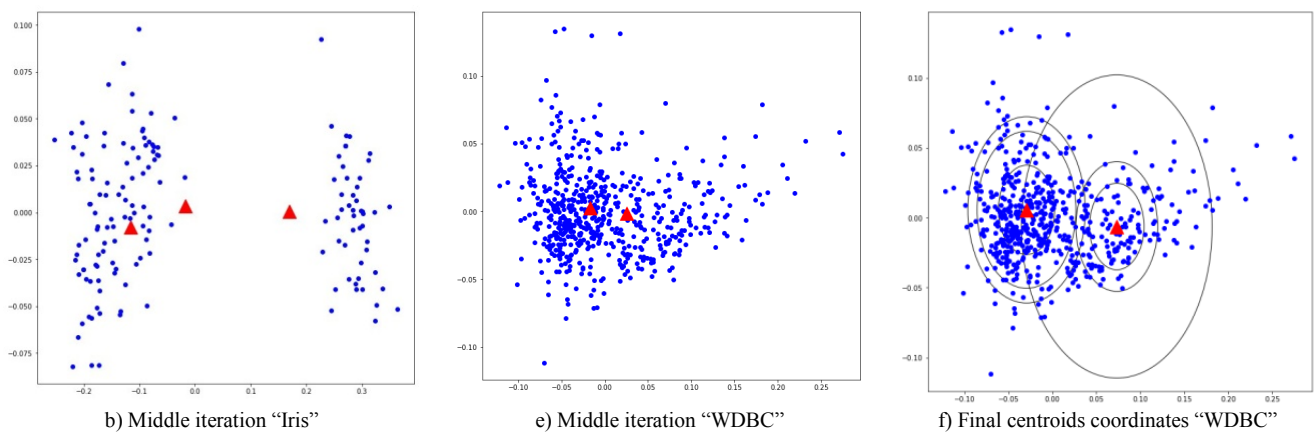


Fig. 1. The changes centroids coordinates

REFERENCES

- [1] C. C. Aggarwal, Data Mining. Cham: Springer, Int. Publ., Switzerland, 2015.
- [2] M. Bramer, Principles of Data Mining. Springer-Verlag London, 2016.
- [3] A. Bifet, R. Gavaldà, G. Holmes, and B. Pfahringer, Machine Learning for Data Streams with Practical Examples in MOA. The MIT Press, 2018.
- [4] F. Höppner, F. Klawonn, R. Kruse, and T. Runkler, Fuzzy Cluster Analysis: Methods for Classification, Data Analysis and Image Recognition. John Wiley & Sons. Chichester, 1999.
- [5] C. C. Aggarwal and C. K. Reddy, Data Clustering. Algorithms and Application. Boca Raton: CRC Press, 2014.
- [6] R. Xu and D. C. Wunsch, Clustering. IEEE Press Series on Computational Intelligence. Hoboken, NJ: John Wiley & Sons, Inc., 2009.
- [7] A. Bifet, Adaptive Stream Mining: Pattern Learning and Mining from Evolving Data Streams, IOS Press, 2010.
- [8] J. Kacprzyk, and W. Pedrycz, Springer Handbook of Computational Intelligence, Berlin Heidelberg: Springer, Verlag, 2015.
- [9] K.-L. Du and M. N. S. Swamy, Neural Networks and Statistical Learning. London: Springer-Verlag, 2014.
- [10] J.-C. Bezdek, Pattern Recognition with Fuzzy Objective Function Algorithms, N.Y.: Plenum Press, 1981.
- [11] Ye. V. Bodyanskiy, A. O. Deineko, and Y. V. Kutsenko, "On-line kernel clustering based on the general regression neural network and T. Kohonen's self-organizing map," Automatic Control and Computer Sciences, 51(1), pp. 55-62, 2017.
- [12] J. Keller, J. C. Bezdek, R. Krishnapuram and N. Pal, Fuzzy Models and Algorithms for Pattern Recognition and Image Processing. The Handbook of Fuzzy Sets. Kluwer, Dordrecht, Netherlands: Springer, vol. 4, 1999.
- [13] B. Quost, and T. Denceux "Clustering and classification of fuzzy data using the fuzzy EM algorithm," Fuzzy Sets and Systems. vol. 286, pp. 134-156, March 2016.
- [14] J. Yu, Ch. Chaomu, and M. S. Yang, "On convergence and parameter selection of the EM and DA-EM algorithms for Gaussian mixtures," Pattern Recognition, vol. 77, pp. 188-203, May 2018.
- [15] X. L. Meng and D. B. Rubin, "Maximum likelihood estimation via the ECM algorithm: a general framework," Biometrika, vol. 80, pp. 267-278, 1993.
- [16] Ye. Bodyanskiy, "Computational intelligence techniques for data analysis," Lecture Notes in Informatics, Bonn: GI, pp. 15 - 36, 2005.
- [17] Ye. Gorshkov, V. Kolodyazhnyi and Ye., Bodyanskiy, "New recursive learning algorithms for fuzzy Kohonen clustering network," 17th Int. Workshop on Nonlinear Dynamics of Electronic Systems, Rapperswil, Switzerland, pp. 58-61, 2009.
- [18] L. Jain and C. Mumford, Computational Intelligence, Collaboration, Fuzzy and Emergence, Berlin: Springer, Verlag, 2009.
- [19] S. Osowski, Sieci neuronowe do przetwarzania informacji, Warszawa: Oficjalna Wydawnicza Politechniki Warszawskiej, 2006.
- [20] A. B. Geva and I. Gath "Unsupervised optimal fuzzy clustering," Pattern Analysis and Machine Intelligence, vol. 2, n.7, pp. 773-787, 1989.
- [21] T. Kohonen, Self-Organizing Maps. Berlin: Springer-Verlag, 1995.
- [22] Ye. Bodyanskiy, A. Deineko, Y. Kutsenko and O. Zayika, "Data streams fast EM-fuzzy clustering based on Kohonen's self-learning," 1th IEEE International Conference on Data Stream Mining & Processing (DSMP 2016), Lviv, Ukraine, pp. 309-313, 2016.
- [23] A. B. Geva, "Clustering as a basis for evolving neuro-fuzzy modeling," Evolving Systems, pp. 59-71, 2010.
- [24] UCI Repository of machine learning databases. CA: University of California, Department of Information and Computer Science. [Online]. Available: <http://www.ics.uci.edu/~mllearn/MLRepository.html>

Information Technology of Process Modeling in the Multiparameter Systems

Solomija Ljaskovska
Department of designing and operation of machines
Lviv Polytechnic National University
Lviv, Ukraine
solomiam@gmail.com

Igor Malets
Department of Project Management, Information
Technologies and Telecommunications
Lviv State University of Life Safety
Lviv, Ukraine
igor.malets@gmail.com

Yevgen Martyn
Department of Project Management, Information
Technologies and Telecommunications
Lviv State University of Life Safety
Lviv, Ukraine
evmartyn@gmail.com

Oleksandr Prydatko
Department of Project Management, Information
Technologies and Telecommunications
Lviv State University of Life Safety
Lviv, Ukraine
o_prydatko@ukr.net

Abstract— Information graphics technologies of designing models of the processes of multiparameter technical systems are argued and developed in order to increase the effectiveness of determining the influence of many operating parameters on their dynamics. The requirements for model designs are formulated on the basis of the formed numbers of different measurements of multidimensional spaces. Its implementation is proposed by a complex combination of the mathematical description of the parameters interconnection and the use of rational geometry. The features of the implementation of models are shown with the number of possible assumptions reduced.

Keywords— *information technology, modeling, multi-parameter system, processes, applied geometry*

I. INTRODUCTION

Open systems that surround a person and part of which it is, are multiparameter. Their research, the discovery of useful features and the creation of more perfect directed intellectual activity of a scientist. The isolation of the characteristic features of the system leads to the identification of its significant parameters by accepting one or another number of assumptions. As the historical development of information technology neglected many factors of influence on the behavior of the system was reduced to the adoption of a minimum number of significant parameters that distinguish the studied system among others. In the process of creating and designing target models, it is important to take into consideration the ways in which the parameters of individual units of the studied systems are presented, for example, using clearly defined or fuzzy sets [1]. Construction of models of systems with the ability to study not only static but also dynamic characteristics requires the involvement of both classical and new methods [4].

This approach made it possible to create almost identical models for different physical entities and apply similar research methods to them, in particular, mechanical [12, 6] or mechatronic systems [11]. For example, the system of two differential equations of the first order with defining and constant parameters a, A, b, B, R, c, D is a basic mathematical model of many multi-parameter systems of

diverse purposes, which, using a finite number of assumptions, are reduced to two-parameter systems. To them, as an example of IT technology in improving the quality of educational processes in the training of rescue workers [3], widely known biological systems, the operating elements of which are related to the model Lotka-Volterra, fire and technical systems, which, by a number of assumptions, are reduced to dual systems [4], direct or alternating current motors when powered from an electric network of infinite power [5] and others. Such models corresponded or sometimes correspond to the operational requirements within the limits of acceptable for engineering calculations of accuracy.

$$\begin{aligned} \frac{dy}{dt} &= \frac{1}{a}(A - Ry - cx); \\ \frac{dx}{dt} &= \frac{1}{b}(By - D) \end{aligned} \quad (1)$$

The growth of requirements to systems, primarily to the technical, the maintenance of the technological requirements for the accuracy of the parameters prompts the development of modern methods and models, research, and therefore, on the contrary, reducing the number of assumptions, and as a consequence, requires the development and use of new information technology as an environment for the implementation of modern models, methods, algorithms for calculating the values of parameters of multiparameter objects, systems or processes. For example, research of the dynamics of a mechanism with a DC motor is possible based on a mathematical model using (1), which reproduces processes with sufficient accuracy. However, the higher harmonics of the power supply system of the thyristor converter significantly affect the quality of communication. Research of such influences requires the development and use of other specialized models.

In the scientific research [14], theoretically based propositions concerning the modeling and development of certain types of equipment with the use of artificial neural networks are proposed.

The proposed approach to the scientific and practical process of model development is based on the fact that a wide class of systems reflects its behavior in universal models, which often differ in system variables and parameters [7].

The development of models in the classical version occurs sequentially from physical representations to the detection of the method of representing the interconnections of significant parameters with a rational number of assumptions. Due to the change of the philosophy of information technology, the transition to computerized means of scientific research, the development of the research process takes place due to the complexity of models taking into account sufficient for engineering calculations of the number of assumptions regulated by the capabilities of the computer, provided that all the relationships of parameters that are submitted mathematical or geometric means, equal regardless of their importance [14,15].

II. SCIENTIFIC AND APPLIED PRINCIPLES OF CONSTRUCTING MODELS

The process of accumulation of human knowledge is continuous and inexhaustible. One of the means of cognition is simulation. The process of modeling in science is carried out for a thorough study of the properties of the object, the identification of the laws of the mutual influence of the connections of its parameters and the impact on them to optimize the functioning and establishment of useful properties.

In modeling, numerical values of parameters, are presented by graphic dependencies [16] which are used in the analysis of the system [17, 18].

The simulation process consists of two active elements, a simulation object and, in fact, its model, which in essence represents a dual system. Formation of an object model can be given, in particular, as follows (Fig. 1).

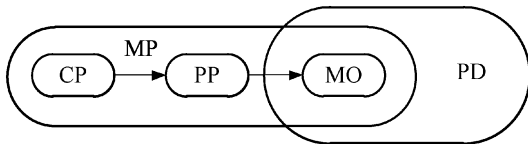


Fig. 1. The diagram of process for the development model of object MO

The process of constructing an object model begins because of the need to obtain additional knowledge about the object, based on the infinite fluidity of knowledge, which can be given by an infinite-dimensional linear vector space, mutually perpendicular orcs of which are the constituent parts of general knowledge. Such a space as the central in the process of modeling the CP space, interacting with the scope of the parameters of the object PP, forms a modeling space MP. Models of objects MO, configured to conduct research on the interconnection of their parameters, are represented by real numbers. Consequently, by immersing the modeling space MP in the field of real numbers PD, we obtain a model of the object of the study of MO as a common object of the modeling space MP and the field of real numbers PD. The model of the object of the MO is realized mainly in the multidimensional Euclidean space, the dimensionality of which as a derivative of the dimensionality of the central space CP is determined by the number of independent essential parameters of the investigated object.

In the presence of a positively directed axis, for the simulation of the flow of processes in time, space half-space is used as a space of multiparameter state with the parameters of the technical system (Fig. 2 a, c). Projection of the state space in a direction parallel to its axis, we obtain a phase space (Fig. 2 d) or a phase space with two and interconnected differential equations of essential parameters (Fig. 2 b).

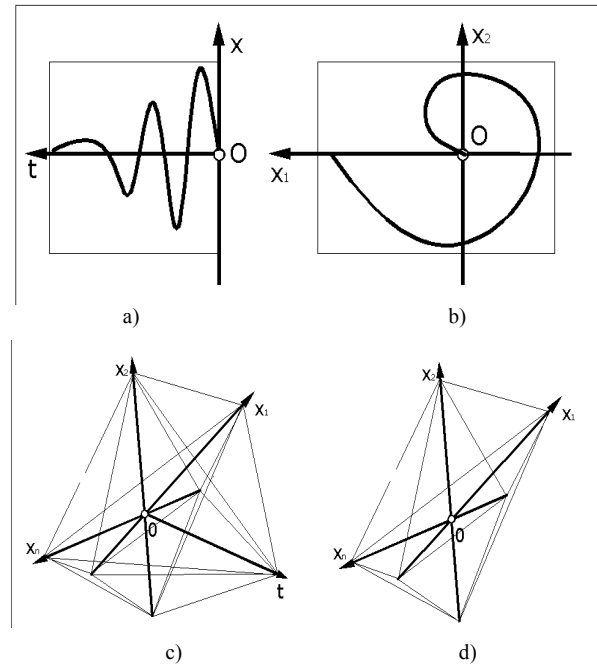


Fig. 2. Spaces of state and phase spaces of multiparameter technical systems

The integral curves (Fig. 2a) and the phase trajectories (Fig. 2b) are presented in the two-dimensional coordinate planes of the multidimensional space of the state of the technical system with measurements of the variables, for example, the time, current and frequency dependencies of the electric motor from time (Fig. 3).

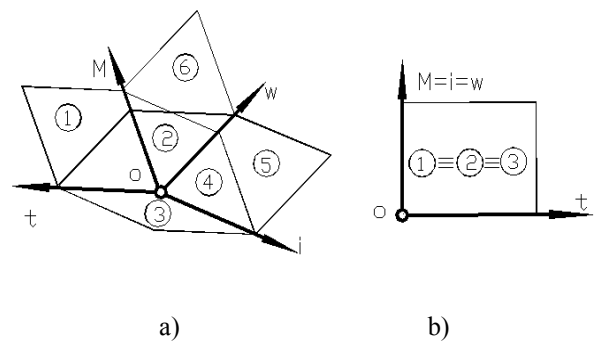


Fig. 3. Geometric model and complex drawing for dependencies of engine parameters

The rotation of the coordinate planes relative to the coordinate axes gives the Cartesian coordinate system (Fig. 3b) with the combined planes of projections as a complex drawing of the spatial geometric model of the dependencies of the engine parameters.

The overlap of geometric images in the combined planes of projections can be avoided using the proposed integrated drawings. For a four-dimensional Euclidean space, we map all the fields in two-dimensional coordinate planes (Fig. 4).

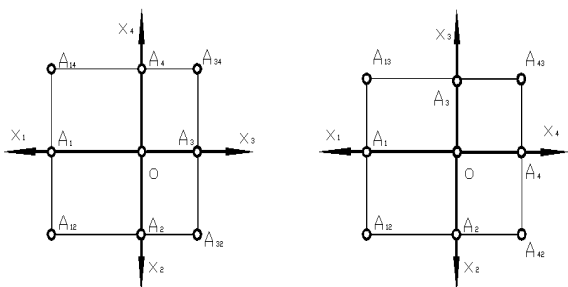


Fig. 4. Comprehensive drawing of a four-dimensional Euclidean space

This drawing, provided the smallest number of coordinate two-dimensional planes, has the form (Figs. 5a, b), with a generalization of it on three-dimensional coordinate planes (Fig. 5c, d).

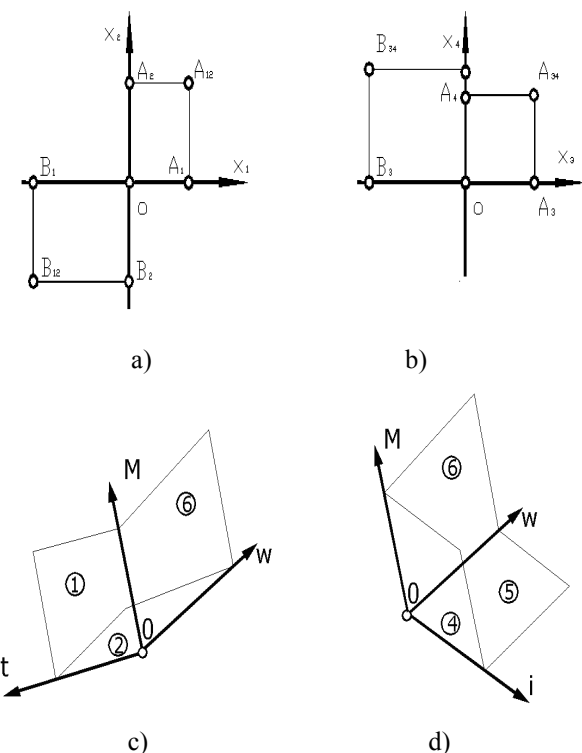


Fig. 5. Variants of complex drawings for computer visualization

The integral curves of the process of changing the time of two parameters are given by the three-space line of the Euclidean three-dimensional space [8,9]. To form a multi-spatial space line, it is quite sufficient that the change of each of the parameters is reflected by the integral curve in the corresponding coordinate plane of the multidimensional state space. Each integral curve represents the direction of a multidimensional cylinder, the intersection of which forms the geometric multidimensional space. For its unambiguous definition it is necessary to have a minimal, but sufficient number of integral curves in coordinate planes, which determines the completeness of representation of the geometric image. Two integral curves in two coordinate

planes of a three-dimensional space determine the position of a three-spatial integral curve: the curves form guides of two-dimensional cylinders, whose intersection is determined by a one-dimensional multidimensional, linear curve, three-dimensional space. Two of the six coordinate planes of the four-dimensional space pair the links of four parameters, for example, and and. Such connections in the form of plane curves determine the position of guides of three-dimensional cylinders whose intersections in the four-dimensional Euclidean space determine the position of the two-dimensional multidimensional. In addition, there is a change in time only for the parameter z. The minimum number of d multi-species, which determines the completeness of the mapping curves of the dimension of the transition multiparameter process, determine, drawing dependence [10]:

$$\eta = \sum_{i=1}^d m_i - n(d-1), \quad (2)$$

For the same measurements of multidimensional cylinders with guides with integral curves, their number is determined from (3):

$$d = \frac{n-1}{n-m} \quad (3)$$

In three-dimensional space the number of two-dimensional cylinders of measurements, is. For a four-dimensional space of state, when, we have.

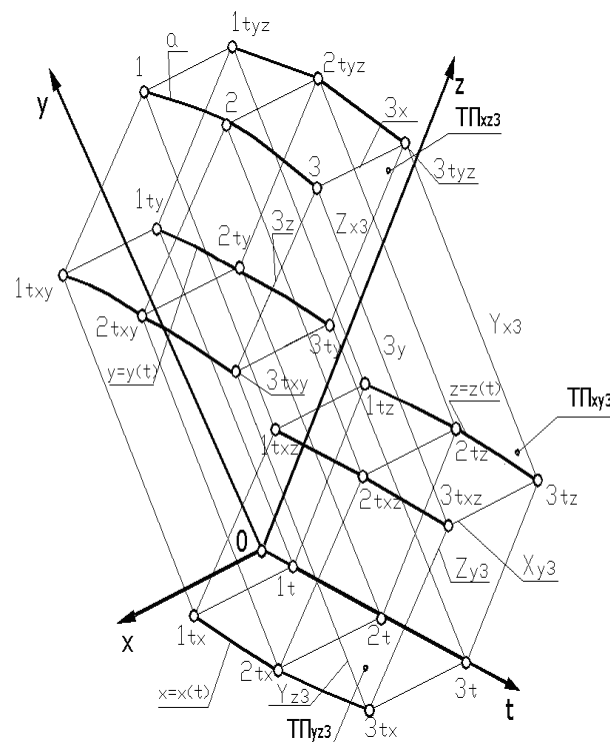


Fig. 6. Formation of the integral curve a in the space of the Oxyzt state

We set the completeness of the task of the integral curve of the four-dimensional space Oxyzt on the basis of (3). We will accept three integral curves as guides of three-dimensional cylinders with generating two-dimensional

planes (Fig. 6). The interconnection of three-dimensional cylinders forms a curve as a one-dimensional four-spatial manifold of this space. Each point of the curve, for example, 3 is formed by the cross section of the corresponding geometric images.

3D dimensional plane $3_1 3_{1x} 3_{1z} 3_{1y}$ $m_{III} = 3$ with the foll 3_1 on the axis Ot crosses each of three three-dimensional cylinders of dimensionality $m_c = 3$ in a plane $T\Pi_i$ dimensionality $r_n = m_c + m_{III} - n = 2$. The dimension of the geometric images r_i of the intersection of the hyperplane and each of the guides $x = x(t), y = y(t), z = z(t)$ $r_i = m_n + m_{III} - n = 0$

where $m_n = 1$ is the dimension of the guide of the three-dimensional cylinder. These geometric images are given points $3_{1x}, 3_{1z}, 3_{1y}$. The planes III_{yz3} i III_{xy3} , III_{xz3} i III_{xy3} , III_{xz3} i III_{xz3} intersect in points $3_{1xz}, 3_{1yz}$, and 3_{1xy} the set of which defines the position of the intersection of the two-dimensional and three-dimensional hyperplanes. The point 3_{1xz} determines the position of the 3_y intersection of the two-dimensional plane III_{yz3} and the three-dimensional plane $3_1 3_{1z} III_{xy3}$. The point 3_{1xy} determines the position of the line 3_z of intersection of the two-dimensional plane III_{yz3} and $3_1 3_{1y} III_{xz3}$ the point 3_{1yz} determines the position of the intersection 3_x of the two-dimensional plane III_{xy3} and the three-dimensional hyperplane $3_1 3_{1y} III_{yz3}$. Direct $3_x, 3_y$ and 3_z intersect at point 3, the set of which forms the curve α of the four-dimensional space of the Oxyzt state of the technical system, and substantiates the assertion about the completeness of representation of the integral curves of multidimensional phase and space of the state of the technical systems. Trajectories of phase spaces are obtained by projection of integral curves of state space in the direction parallel to the axis Ot of this space, which describe the transition process in a multiparameter system by differential equations of the first order

$$\frac{dx_i}{dt} = f_i(x_1, x_2, \dots, x_j, \dots, x_n, t). \quad (4)$$

The integral curve, for example $x_1 = x_1(t)$, is a directing line of the cylinder with generator $n-1$ - dimensional subspaces, which are parallel to $n-1$ - dimensional coordinate subspace $Ox_2x_3\dots x_i\dots x_n$. The dimension of each cylinder is $k_i = 1 + (n-1) = n$.

Dimensionality of r_{k12} section of two arbitrary cylinders with dimension $k_1 = k_2 = n$ $n+1$ - dimensional state space $Ox_1x_2x_3\dots x_i\dots x_n t$ is $r_{k12} = k_1 + k_2 - (n+1) = n + n - n - 1 = n - 1$, and its section using the third cylinder with dimensionality $k_3 = n$ forms multi-species with dimensionality $r_{k3} = r_{k12} + k_3 - (n+1) = n - 2$. The dimensionality of the

cross-section $n-1$ - dimensional surface by the next multidimensional cylinder decreases by one.

III. IT-IMPLEMENTATION OF GEOMETRIC MODELING OF PROCESSES

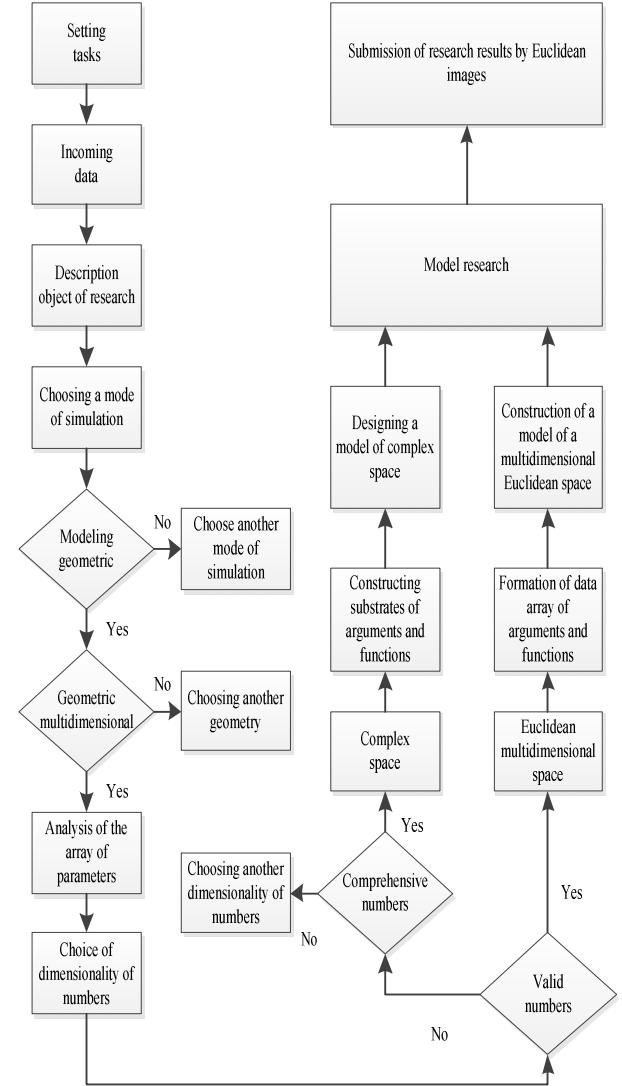


Fig. 7. Diagram of IT-modelling

Finally, dimensionality of r_{kn} multi-species as section $n-1$ - dimensional surface using n - dimensional cylinder is $r_{kn} = n - (n-1) = 1$, which is the dimensional n - spatial one-dimensional line $n+1$ - dimensional space of the system state. Projection of the curve as a guide of the projection cylinder of dimensionality $l = 2$ in the subspace of variables $Ox_1x_2\dots x_i\dots x_n$ as n - dimensional phase space we get a geometric image of dimensionality

$$q = l + n - (n+1) = 1,$$

which represents the dimensionality of the phase trajectory as a projection of the integral state space curve $Ox_1x_2\dots x_i\dots x_n t$ in the subspace of variables $Ox_1x_2\dots x_i\dots x_n$ as the phase space of the system.

Geometric modeling of processes (Fig. 7) involves the selection of all stages of the basic stages, which include, in particular, the choice of input parameters and the use for simulation of the numbers of the corresponding measurements (real, complex, etc.), the method of geometric modeling, design and research of the model with presentation of research results in the Euclidean space.

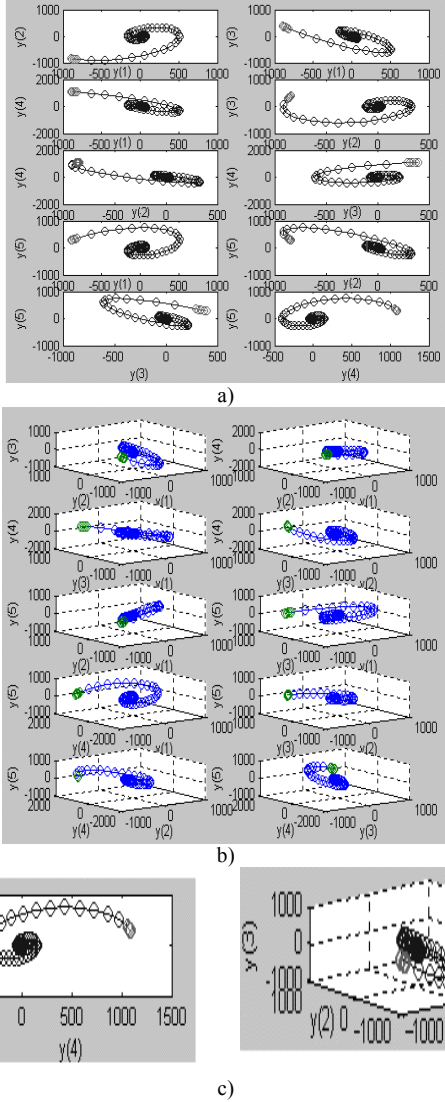


Fig. 8. Projections of the phase trajectory

The choice of the calculator determines to a greater extent the effectiveness of the research. Taking into account the possibility of projection of phase trajectories into two- and three-dimensional phase planes, it is effective to use the tools of computer mathematics Matlab.

An example of the realization of the projection of the phase trajectory of the solution of the differential equation, in particular, of the fifth order

$$\frac{d^5 y}{dt^5} + \frac{d^4 y}{dt^4} + \frac{d^3 y}{dt^3} + \frac{d^2 y}{dt^2} + \frac{dy}{dt} = 0 \quad (5)$$

indicates the need to reduce its order and to form a system of differential equations of the first order. For given values of the initial conditions y_0 and the integration

time $tspan$ of the use of options 'odephas2' and / or 'odephas3' commands $options$, it is possible to obtain solutions as projections of the phase trajectory (5) in two-dimensional (Fig. 8a), three-dimensional (Fig. 8b) projection planes and their combination (Fig. 8c).

IV. PRACTICAL IMPLEMENTATION OF GEOMETRIC SIMULATION OF PROCESS

A. Drive with asynchronous electric fire pump

The drive of low-power fire pumps is carried out using asynchronous short-circuited motors. The joint work is investigated by the analysis of so-called dynamic mechanical characteristics $M = M(\omega)$ in a two-dimensional space. Such curves represent one of the projections of an integral curve in a two-dimensional plane (Fig. 9) as one of the solutions of the system (6) of differential equations [3]

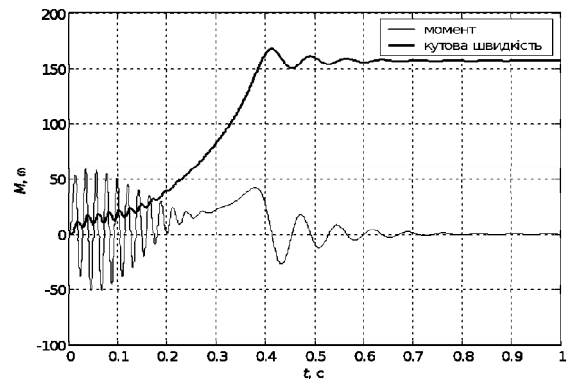


Fig. 9. Projection of the integral induction motor curve

$$\begin{aligned} \frac{d\psi_{x1}}{dt} &= u_m \cos \gamma + \omega_0 \psi_{y1} - \omega_0 \alpha_s \psi_{x1} + \omega_0 \alpha_r k_r \psi_{x2}; \\ \frac{d\psi_{y1}}{dt} &= u_m \sin \gamma - \omega_0 \psi_{x1} - \omega_0 \alpha_s \psi_{y1} + \omega_0 \alpha_r k_r \psi_{y2}; \\ \frac{d\psi_{x2}}{dt} &= \psi_{y2} (\omega_0 - \omega) - \omega_0 \alpha_r \psi_{x2} + \omega_0 \alpha_s k_s \psi_{x1}; \\ \frac{d\psi_{y2}}{dt} &= -\psi_{x2} (\omega_0 - \omega) - \omega_0 \alpha_r \psi_{y2} + \omega_0 \alpha_s k_s \psi_{y1}; \\ M &= \frac{3}{2} \frac{p \omega_0 k_2}{x_s \sigma} (\psi_{x2} \psi_{y1} - \psi_{x1} \psi_{y2}); \\ \frac{d\omega}{dt} &= \frac{p}{I} (M - M_c). \end{aligned} \quad (6)$$

where $u_{u1}, u_{v1}, u_{u2}, u_{v2}, \psi_{u1}, \psi_{v1}, \psi_{u2}, \psi_{v2}$,

$$\omega_0, \omega, \omega_k, \alpha_s = \frac{r_1}{\sigma x_s}; \quad \sigma_r = \frac{r_2}{\sigma x_r}, \quad \sigma = 1 - \frac{x_0^2}{x_s x_r} = 1 - k_r k_s,$$

r_1, r_2, x_0, x_s, x_r – asynchronous motor parameters.

M, M_s – electromagnetic and moment of loading of asynchronous engine; p – number of pairs of poles of the asynchronous motor; I – moment of inertia of the system, brought to the shaft of the induction motor.

Projections of the phase trajectory of the transition process in the asynchronous motor (Fig.10) make it possible to conduct research of all its determining parameters.

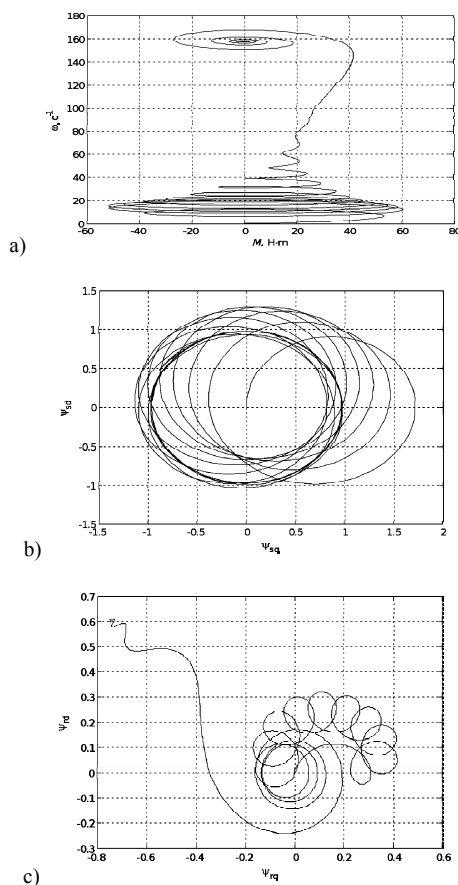


Fig. 10. Projection of the phase path of the start of the asynchronous motor

B. Drive with DC motor and periodic load torque

The co-use of two- and three-dimensional planes of phase spaces illustrates the example of the implementation of the system of differential equations (7) of the DC motor and the periodic moment of the load on the shaft

$$\begin{aligned} u &= L \frac{di}{dt} + i \times R + C_e \omega; \\ I \frac{d\omega}{dt} &= C_M i - M_c, \end{aligned} \quad (7)$$

where L, R, c_e, c_m - engine parameters u, i - voltage and current.

V. CONCLUSION

For the first time, the completeness of presentation of graphical information means of mapping of integral curves and phase trajectories of multidimensional phase spaces of the state of technical systems is substantiated. The development of graphical information technology tools expands the capabilities of model development and the study of the determinants of the parameters of processes of arbitrary material multiparameter systems, regardless of the physical content of the parameters, with the involvement of

graphic capabilities of IT technologies. Further research relates to IT technologies for geometric modeling of processes by reducing the number of equations describing the state of a system with the use of numbers of higher measurements, in particular complex numbers.

REFERENCES

- [1] S. Briot, and W. Khalil, *Dynamics of Parallel Robots: From Rigid Bodies to Flexible Elements*. Springer International Publishing Switzerland, 2015.
- [2] Yj. Zhao, "Dynamic optimum design of a three translational degrees of freedom parallel robot white considering anisotropic property," *Robotics and Computer-Integrated Manufacturing*, vol. 29(4), pp. 100-102, 2013.
- [3] O. Prydatko, and I. Pasnak, "Investigation of the processes of the information technologies integration into the training of specialists at mine rescue departments," *Scientific bulletin of National Mining University*, is. 1 (157), pp. 108-113, 2017.
- [4] P. Chmiel, Y..Martyn, J. Olenjuk and Ya.. Pidgorodecky "Metody reprezentacji modeli w zarzadzaniu zorientowanym projektowo transgranicznych jednostek operacyjno - ratowniczych," *Technika, Informatyka, Inzynieria bezpieczenstwa.- Czestochowa*, pp. 31-48, 2014.
- [5] V. Lobov, K. Lobova, and Ye. Fortuna, "Comparison of mechanical properties of asynchronous electric motors at various schemes of paravetric control," *Scientific bulletin of National Mining University*, is. 1 (157), pp. 88-92, 2017.
- [6] D. Pilchicovs, and E. Dzelzitis, "Evaluation of Efficiency Improvement Potential Applying Proportional Pressure Control of Variable Speed Pumps in Water Supply," *International Journal of Engineering Science Invention*, vol. 2, is. 9, pp. 29-38, 2013. ISSN (Online): 2319-6734, ISN (Print): 2319-6726.
- [7] O. Gumen, N. Spodyniuk, M. Ulewicz, and Y. V.Martyn, "Research of thermal processes in industrial premises with Energy - saving technologies of heating," *Diagnostics: Collection of scientific works, Poland: Polish Society for Technical Diagnostics*, no.18(2), pp. 43-49, 2017.
- [8] B. Arnold, *Teoria katastrof*. M.: Nauka, 1990.
- [9] Anton I. Guda, and A. I. Mikhalyov, "Criteria synthesis problem for the chaotic systems identification," *IEEE 1st International Conference on Data Stream Mining and Processing (DSMP 2016)*, Lviv, Ukraine, pp. 125-128, 2016.
- [10] S.M. Koval'ov, M.S. Humen, Pustul'ha S.I. at el. *Prykladna heometriya ta inzhenerna hrafika*. Luts'k: LDTU, 2006.
- [11] Micro-Cap 11 Electronic Circuit Analysis Program. Users Guide. © Spectrum Software. 1982-2014. Jon-line Available at: <http://www.spectrum-soft.com/download/ug111.pdf>
- [12] P. Kolpachyan and Alexander Zarifyan Jr., "Study of the asynchronous traction drivers operating modes by computer simulation. Part 2. Simulation results and analysis," *Transport problems*, vol. 10, is. 3, pp. 5-15, 2015.
- [13] L. House-Peters, and Heejum Chang, "Urban water demand modeling: Review of concept, methods and organizing principles," *Water resource research*, vol. 47, pp. 1-15, 2011.
- [14] K. Frank J. *General Ellipse Packing in Optimized Regular Polygons*, (Submitted for Publication February 2016) / Frank J. Kampas, Ignacio Castillo, Janos D. Pinter // *Global Optimization Submissions*. - 2016. (http://www.optimization-online.org/DB_FILE/2016/03/5348.pdf).
- [15] J. Kalrath, and S. Rebennack "Cutting ellipses from area-minimizing rectangles," *Journal of Global Optimization*, vol. 59 (2-3), pp.405-437, 2014.
- [16] W. X. Xu, H. S. Chen, and Z. Lv, "An overlapping detection algorithm for random sequential packing of elliptical particles," *Physica*, vol. 390, pp. 2425-2467, 2011. doi:10.1016/j.physa.2011.02.048.
- [17] S. Kramer, R. Gritzki, A. Perschke, M. Roesler, and C. Felsmann, "Numerical simulation of radiative heat transfer in indoor environments on programmable graphics hardware," *International Journal of Thermal Sciences*, vol. 96, pp. 345-354, 2015.
- [18] E.-H. Lee, D.-Y. Yang., "Experimental and numerical analysis of a parabolic reflector with a radiant heat source." *International Journal of Heat and Mass Transfer*. Vol. 85, pp. 860-864, 2015.

Scenario of Interaction of the Mobile Technical Objects in the Process of Transmission of Data Streams in Conditions of Impacting the Powerful Electromagnetic Field

Gennadiy Churyumov
Photonics and Laser Engineering
Kharkiv National University of Radio
Electronics
Kharkiv, Ukraine
g.churyumov@ukr.net

Vladimir Tokarev
Electronic Computers Department
Kharkiv National University of Radio
Electronics
Kharkiv, Ukraine
tokarev.v@ukr.net

Vitalii Tkachov
Electronic Computers Department
Kharkiv National University of Radio
Electronics
Kharkiv, Ukraine
vitalii@tkachov.com

Stanislav Partyka
Electronic Computers Department
Kharkiv National University of Radio
Electronics
Kharkiv, Ukraine
stanislav.partyka@nure.ua

Abstract — This paper analyzes the current state of the problem of obtaining and processing big data streams by a group of mobile technical objects (drones) under the influence of powerful electromagnetic field. Recommendations regarding to the design of mobile systems characterized by increased survivability are given.

Keywords — data stream, mobile system, drone, electromagnetic field, microwave pulse, survivability.

I. ACTUALITY

The paradigm of the development of modern information systems is their progressive and ever-increasing integration into various spheres of activity (public administration, industry, energy, communications, etc.) and the simultaneous complication of the functions they perform. Successful tests of algorithms for organizing and functioning of complex mobile systems for registering and transmitting data streams (swarm algorithms) have given rise to a new direction in the design of information systems [1, 4]. Such systems are characterized by sophisticated architecture, heterogeneity, structural diversity, multifunctionality, and so on. In terms of composition, such systems can be divided into a number of subsystems that are distributed in space; they are mobile and functioning asynchronously among themselves for the performance of a single goal. The considered technical objects are designed to solve the tasks of information registration, temporary buffering and real-time processing of data streams. Under the task of information registration, one should understand photo and video recording, monitoring of the state of electromagnetic and radiation environment, etc. Temporary buffering allows you to preserve the integrity of data when being processed and transmitted. The set of problems being solved by such systems allows the performing of both current and new classes of tasks [4].

The unifying factor in the design of such systems is the assured transmission of data streams using wireless

technologies. In this case, the factor of external influence on the functioning of a complex mobile system is of particular interest. The external influence should be understood here as meteorological (natural) effects (for example, lightning) and artificial (for example, a powerful electromagnetic field). With external electromagnetic effects on a complex mobile system, an important role is played by the problem of increasing survivability [8]. For example, the work [1] presents a complex mobile system consisting of a number of drones, united in a single network. The communication channels are organized on the base of IEEE 802.11 standard family to communicate with each other. The successful achievement of the stated objectives depends on the characteristics of the network organization and the external effect of natural and / or artificial origin. The presented system is characterized by low noise immunity under the conditions of a powerful electromagnetic field. That reduces the data transfer rate and provides the integrity of the data streams transferred.

There are the solutions on the use of drones as unmanned retransmission nodes to support a wireless telecommunication link between two nodes [2]. When the propagation range of a radio signal from stationary nodes does not justify itself, and the power increase mode of transmitters does not guarantee their stealthiness (makes them radio beacons), the mobile relay nodes are used then to amplify wireless communication channels, acting as a "the last mile". The solutions have been analyzed where the mobile objects can act as communication repeaters: ground transport vehicles, helicopters or geostationary satellites. However, these facilities have certain limitations in performance under unfavorable conditions: inaccessible areas (mountains, caves, ice roads) and high operating costs. The authors propose and justify the strategy of using drones as mobile relay communication nodes to overcome these limitations [2]. The essential shortcoming is the lack of

organizational arrangements for the deployment of such systems and the security of the data transmitted by them.

The work [3] presents the method of data collection from stationary objects provided by drones. Its essence lies in the preliminary estimation of the amount of data received from each node and, accordingly, in the choice of appropriate data transmission technology in order to minimize the total data acquisition time from all stationary nodes. The disadvantage of this solution is the technical aspect: the use of narrow-band antennas while using high-speed data transfer standards. The task of adjusting with stationary nodes may take longer than the use of lower-speed standards and omnidirectional antennas. The work [6] deals with the problem of transferring large data streams in mobile engineering systems in which the drone and the sensor network are regarded as the objects. The modern systems of collecting data by sensor networks from geographically distributed points have been analyzed. The efficiency has been shown of using drones for data collection from sensor networks. Several approaches have been proposed to solve the problem of data transmission using the Big Data technology. The work [6] describes the ideal conditions under which the system functions, and the data rate indicators correspond to the maximum possible within the chosen data transmission standard, and which requires the addition of a mathematical apparatus for describing the scenarios of the drone behavior under impact conditions. In many cases, the abovementioned systems that operate in the conditions of the destructive external environment are subjected to additional requirements for reliability and survivability. This is done to ensure the safety and integrity of the accumulated data as a result of the functional task of the drone for their further transfer to the processing center. For example, the swarm of drones can be considered as an information and communication system for providing access to the Internet in remote locations under natural conditions, and in the case of using ultra-broadband communication technology, the system can be characterized as possessing the property of survivability in the event of an external destructive effect on blocking out the signals of control and data transmission [5]. The downside is the unsolved problems of finding the compromise in the power-producing aspect that ensures the operation of the drone and the use of energy-intensive technology for the assured data transmission. Taking into account the absence of a universal approach to solving the problem of ensuring the electromagnetic stability of a group of mobile technical objects (drones) in the conditions of the assigned task to provide the guaranteed high-speed data transmission in the studies [1-3, 6], this line of the research is topical today. The objective of this work is to develop the scenarios for the interaction of a group of mobile technical objects while providing the assured data transmission under the conditions of external electromagnetic influence.

II. DESCRIPTION OF THE SYSTEM

A group of mobile technical objects (drones) will be considered as a data system. Each of the drones can simultaneously perform the following functions: information recording, primary processing, transmission, reception, storage, and data destruction. Let us also notice that the additional conditions for the operation of the drone are reception, acknowledgment, transmission of control commands and broadcast messaging. According to the

functions, the drone can be described by a set of the group parameters:

- the parameters relating to the subsystem of reception and transmission of service information (for example, control commands);
- the parameters relating to the subsystem of information registration, its primary processing, temporary data storage, data reception and transmission.

This study does not consider the features of controlling a group of drones (parameters relating to the subsystem of reception and transmission of service information). It is assumed that the control with the account of the external impact is based on already known algorithms with weak a priori and weak a posteriori info ware for solving group control problems under the conditions of organized counteraction in dynamic, nondeterministic environments [4]:

- if the task is to control a single object, then the “drone-environment” system is continuous, and its functionality is described by a system of differential equations as following:

$$\dot{S} = \tilde{f}(A(t), S(t), g(t), t) \quad (1)$$

where $\dot{S} = \frac{dS(t)}{dt}$ – is the derivative of the steady-state vector-function of the “Drone-Environment” system; $A(t) = [a_1(t), a_2(t), \dots, a_m(t)]^T$ – is the vector-function describing the set of actions of the drone; $S(t) = f_s(A(t))$ – is the function of actions of the drone; $g(t)$ – is the external influence;

- if the task is to control a group of objects in a destructive external environment, then the functional specifying the objective of the functioning of a group of objects can be represented in the form of:

$$Y_c = \int_{t_0}^{t_f} [F(\mathfrak{R}(t), E(t), A_c(t)) - G(A_c(t), g(t))] dt = \int_{t_0}^{t_f} F(\mathfrak{R}(t), E(t), A_c(t)) - \int_{t_0}^{t_f} G(A_c(t), g(t)) dt \rightarrow \max \quad (2)$$

where $\mathfrak{R}(t)$ – are the states of the drones in the group; $E(t)$ – is the state of an external environment; $A_c(t)$ – the actions of the drones; $G(A_c(t), g(t))dt$ – the function of the external negative influence.

At present, it is of interest to develop the scenario for the interaction of drones within the solution of the assured data transmission resulting from the registration of the target information. For this purpose, let us consider the following subsystems of the drones: information registration subsystem, primary processing subsystem, temporary storage subsystem, data receiving and transmission subsystem and the subsystem of recording the intensity of electromagnetic interference (sensor) (Fig. 1).

The hardware-software subsystem of data reception and transmission is characterized by the following: the standard of transmitted data depending on the selected frequency, the speed of data reception and transmission, the sustainability and integrity of the data received. For example, such a subsystem may be presented by a set of radio modules

operating within the standards of the IEEE 802.11 family, and the sustainability and integrity of the received data at the level of data transfer protocols.

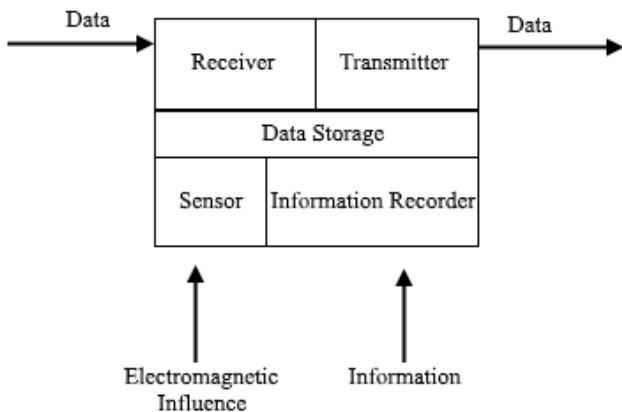


Fig. 1. The Subsystems of the Drone

The subsystem of temporary data storage is characterized by the capacity, the load monitoring function, and the function of verifying the integrity of stored data.

The sensor carries out the function of recording the intensity of external electromagnetic effect in a wide range of frequencies, for the drone to choose the optimal standard for the data transferring among other drones and for the generating the service messages about the location of the electromagnetic impact zone. It is important that the information received from the sensor is the influential factor when choosing the data exchange standards for the subsystem of data transmission and reception. In fact, the information recorder performs the information registration with its subsequent conversion into the data storage format in the temporary storage subsystem. A photo and video camera or a microphone can take on the role of the recorder.

III. SCENARIOS OF INTERACTION

Let us consider the scenarios of a group interaction of drones when solving the task of assured data transmission under the conditions of external electromagnetic influence.

The pulses of electromagnetic radiation, like any radio signal, are characterized by the propagation distance and the attenuation in space. However, the impact of electromagnetic radiation pulses has a different effect on the information system as they are approaching their source. While regarding a group of drones as the information system, it is possible to distinguish conditionally three zones of electromagnetic radiation influence (Fig. 2).

In the zero area there is no evidence of electromagnetic effect registered by the sensor, or the effect occurs to be below the noise level. The drone operates in the routine mode. Depending on the task assigned, the drone performs data reception and transmission using the most high-speed standards, for example, IEEE 802.11ac (5 GHz frequency, 433 Mbps data rate, 150 m transmission range in the open space), in the real time mode.

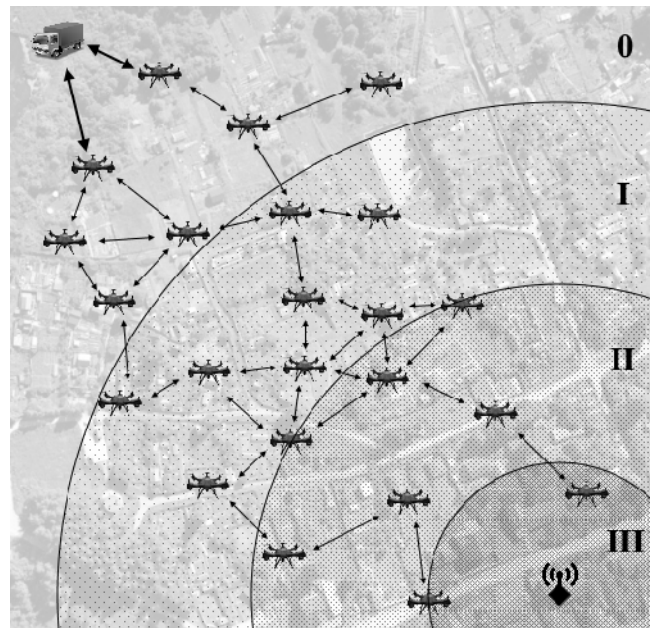


Fig. 2. The zones of electromagnetic radiation impact on the group of drones

The data warehouse is not enabled when rebroadcasting and is involved in the process of converting data from the information recorder into the required format for the further data transmission to the data center. In the first zone, the sensor registers the electromagnetic effect, but its impact does not provide any tangible losses in the process of exchanging the data among the drones. The drone selects the optimal data transmission standard in accordance with its speed characteristics. The most high-speed frequency standard at which the effect is minimal is selected depending on the intensity and frequency of the electromagnetic effect. For example, if the electromagnetic interference covers the range from 2 up to 6 GHz, then it is advisable to go to the IEEE 802.11ad standard, which operates in the 60 GHz range. The disadvantage of this standard is the condition of forming a beam that provides reliable communication within less than 15 meters, which poses the problem of increasing the density of drones in space for high-speed data exchange.

In the second zone, the sensor registers a rapidly growing level of electromagnetic effect, the narrow-band interference appears, aimed to suppress the specific ranges in which data is exchanged; the data flow rate is dropping significantly due to the increase in the noise threshold when approaching the source region of electromagnetic pulses. The optimal solution to maintain the assured data transmission is to use the ultra-wideband communication technology at close ranges when low energy consumptions, using ultra-wideband signals with extremely low power spectral density as a carrier [11]. The use of an ultra-wide frequency band (at least 500 MHz) at distances up to 20 m makes it possible to achieve a data transfer rate of up to 50 Mbps [11]. In the third zone, a destructive effect of the electromagnetic field on the semiconductor component base of the drone occurs. This leads to the appearance and development of degradation processes in its microstructural elements and to the consequent malfunctions of normal operation of objects. The stored information is either distorted or destroyed completely [1]. The drone operates in an emergency mode. At the same

time, it is supposed to manage to solve the task of broadcast announce message about the critically powerful effect of an electromagnetic field in a given location. If possible, the data in the storage are destroyed.

IV. CONCLUSIONS

The task of the paper is to develop a scenario for the interaction of a group of mobile objects (drones) under the conditions of powerful electromagnetic influence. The definitions of the concepts of mobile objects and the consequences of the influence of powerful electromagnetic effects on them have been given. The analysis has been carried out of the problem of providing guaranteed delivery of large amounts of data under the said conditions. The scenario has been proposed for the interaction of a group of drones within a framework of the common task they perform. The variants have been considered of increasing the survivability of mobile objects under the conditions of powerful electromagnetic effect.

The work has been carried out within the framework of the research study "Creation of Scientific and Methodological Foundations for Ensuring the Survivability of Network Information Exchange Systems under the External Influence of High-Power Microwave Radiation" on the basis of "Reconfigurable and Mobile Systems laboratory of Kharkiv National University of Radio Electronics.

REFERENCES

- [1] I. V. Ruban, G. I. Churyumov, V. V. Tokarev, and V. M. Tkachov, "Provision of Survivability of Reconfigurable Mobile System on Exposure to High-Power Electromagnetic Radiation," Selected Papers of the XVII International Scientific and Practical Conference on Information Technologies and Security (ITS 2017). - CEUR Workshop Processing. Kyiv, Ukraine, pp. 105-111, November 30, 2017.
- [2] S. J. Kim, G. J. Lim, and J. Cho, "Drone Relay Stations for Supporting Wireless Communication in Military Operations," /International Conference on Applied Human Factors and Ergonomics, Springer, Cham, pp. 123-130, 2017.
- [3] V.M. Tkachov, and V.V. Tokarev, "Sposib peredachi tsyfrovyykh danykh multykopternoiu systemoiu mizh sehmentamy rozpodilenoii sensornoii merezhi ta bazovoiu stantsieiu" : pat. 118921 Ukraina: MPK 2017.01, H04W 64/00, H04W 84/18 (2009.01), G06F 17/40 (2006.01), Kharkivskiy natsionalnyi universytet radioelektroniky. Vol. 16, 2017.
- [4] I. A. Kaljaev, A. R. Gajduk, and S. G. Kapustin, Modeli i algoritmy kollektivnogo upravlenija v gruppah robotov. M.: FIZMATDIT, 2009.
- [5] D. J. Seo et al., "Object following method for a differential type mobile robot based on Ultra Wide Band distance sensor system," Control, Automation and Systems (ICCAS), 17th International Conference on. IEEE, pp. 736-738, 2017.
- [6] V. M. Tkachov, V. V. Tokarev, V. O. Radchenko, and V. O. Lebediev, "Problema peredachi danykh typu Big Data u mobilnii systemi "Multykopter - sensorna merezha," Systemy upravlinnia, navihatsii ta zviazku, Poltava, Ukraine, no. 2(42). pp.154-157, 2017.
- [7] Z. Hu, et al., "Analytical Assessment of Security Level of Distributed and Scalable Computer Systems," International Journal of Intelligent Systems and Applications, vol. 8, no. 12, pp. 57, 2016.
- [8] O. H. Dodonov, M. H. Kuznetsova, and O. S. Horbacyk, "Metodolohichni aspekta stvorennia korporatyvnykh informatsiino-analitychnykh system pidvyshchenoi zhyvuchosti," Reiestratsiia, zberihannia i obrobka danykh, vol.14, no. 3, pp. 58-69, 2012
- [9] Y. Bodyanskiy, "Computational Intelligence Techniques for Data Analysis," Leipziger Informatik-Tage, pp. 15-36, 2005.
- [10] V. A. Gadyshchev, A. S. Krutolapov, and D. A. Sychev, "Matematicheskaja model' informacionnogo obmena v setjah peredachi danykh," Vestnik Voronezhskogo instituta GPS MChS Rossii. no. 1 (2). pp. 14-17, 2012.
- [11] M. G. Di Benedetto, and G. Giancola, Understanding Ultra Wide Band Radio Fundamentals. Pearson Education, 2004. — Pearson Education, 2004. — 528 p.

Informational System of Project Management in the Areas of Regional Security Systems' Development

Oleksandr Prydatko
Department of Project Management,
Information
Technologies and Telecommunications
Lviv State University of Life Safety
Lviv, Ukraine
o_prydatko@ukr.net

Olga Smotr
Department of Project Management, IT
and Telecommunications
Lviv State University of Life Safety
Lviv, Ukraine
olgasmotr@gmail.com

Yurii Borzov
Department of Project Management,
Information
Technologies and Telecommunications
Lviv State University of Life Safety
Lviv, Ukraine
borzovuo@ukr.net

Ivan Solotvinskyi
Department of Information
Technologies and Technical Means of
Training
Lviv State University of Life Safety
Lviv, Ukraine
jonh7282@ukr.net

Oleksii Didyk
Kherson National Technical University
Kherson, Ukraine
olexii.didyk@gmail.com

Abstract — the analysis of the subject area shows the complexity of the project management in the areas of regional life security systems development due to large data streams and far-flung network of communications between them. Structural-logical scheme has been constructed in the form of a graph of possible system conditions. The scheme opens the full essence of data streams management in the projects of regional security systems development. The structure of interconnection inside the data streams set has been investigated. The structure of the information system of project management in the areas of regional security systems' development has been developed.

Keywords — data stream, information system, project management, life safety.

I. INTRODUCTION

The concept of regional life security systems' development, in accordance with the state policy, should be considered as a complex organizational and technical system, which has many data communications. Management of such systems should be considered in terms of a project-oriented approach, since the development and management of complex systems requires the implementation of individual projects, portfolios, or even programs that cannot be realized without analyzing and processing large data streams. The set of development projects of such systems includes many projects that vary in multisectoral objectives, resource constraints, team models, time frames. However, before moving on to the process of substantiating the volume of data streams for the successful implementation of these projects, an analysis of existing achievements in this area should be conducted. Questions related to the declared topics are studied both in the light of information technologies and project management. In particular, the development of information technologies for operational management in emergency situations, where the life security systems are used actively, is considered in [1, 2]. The basic scientific aspects of project management for the complex systems development are described in [3, 4, 5]. A number of scientific works are devoted to project management issues for regional life security systems development, in particular [6, 7, 8]. An overview of scientific works can be continued

for a long time; however, even the above list shows that there is clearly an unexplored field of the overall problem, namely the development of informational systems for collecting, structuring, storing and processing of data in order to support the process of making effective decisions on the regional life security systems development. There is a number of works devoted to the development of algorithms for processing many simulated data, in particular in the papers [9, 10, 11], a multi-model algorithm for target tracking maneuvering based on the second-order Markov chain is proposed. The results of the work demonstrate the effectiveness of multimodal interaction algorithms in comparison with algorithms of interaction of several models and algorithm of a reliable model that can be used for efficient processing of data flow. As to the basic concepts of portfolio management and development programs, they can be found in the works [12, 13, 14]. Scientific concepts of managing changes in complex organizational and technical systems can be found in a number of works on project management, in particular in papers [15, 16] where a significant part is also devoted to issues of regional development. Consequently, according to the announced problem and the existing achievements in the research area, the aim of the study is as follows: to explore the plurality of data streams and to develop the structure of informational decision-making system for the developing of regional life security systems.

II. ANALYSIS OF THE DATA STREAM CONTENT

For a good understanding of the complexity of the regional life security systems development, consider its main stages using the graph of possible system conditions. Thus, the sequence of the process of restructuring the existing regional systems can be described as a graph $G(X, U)$, using the expression (1):

$$\begin{aligned} X &= \{r_1, d, f_1, t, v, g, n, r_2, f_2, s, k, p, l, z\}; \\ U &= \{u_1, u_2, u_3, u_4, u_5, u_6, u_7, u_8, u_9, u_{10}, u_{11}, u_{12}, u_{13}, u_{14}, u_{15}, \\ &u_{16}, u_{17}, u_{18}, u_{19}, u_{20}, u_{21}, u_{22}, u_{23}, u_{24}, u_{25}\}, \end{aligned} \quad (1)$$

where r_1 - information gathering about the road network of the region; d - information gathering about the location of existing rescue units; f_1 - information gathering about the rescue equipment on the balance sheet of rescue units; t - information gathering about the most remote settlements of the region; v - distribution of areas of responsibility between existing and projected rescue units; g - determination of staffing number and optimal places of rescue teams' disposition; n - information gathering about the fire and manmade load of the analyzed region; r_2 - information gathering on the features of the analyzed region (water sources, hi-rise buildings, etc.); f_2 - determination of the types and required quantity of rescue equipment; s - determination of the optimal staff structure of the unit; k - determination of the qualification requirements for the personnel of the team; l - training of staff at specialist educational establishments, p - organization of property complexes' transfer; z - organization of property complexes' procurement; u_1 - u_{25} - edges describing transitions between system conditions.

For a good understanding of the connections between the possible system conditions the graph $G(X, U)$ is matrix-based. The aim of the matrix-based representation is to indicate communications between the elements (subsets) of the development process. That is why the adjacency matrix is presented for a non-oriented graph:

$$G = \begin{matrix} & r_1 & d & f_1 & t & v & g & n & r_2 & f_2 & s & k & p & l & z \\ \begin{matrix} r_1 \\ d \\ f_1 \\ t \\ v \\ g \\ n \\ r_2 \\ f_2 \\ s \\ k \\ p \\ l \\ z \end{matrix} & \begin{pmatrix} 0 & 1 & 0 & 1 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 1 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 \end{pmatrix} \end{matrix} \quad (2)$$

The adjacency matrix (2) reflects the connections between the stages of the regional life security systems development. However, for a better representation of the sequence of this process, it is necessary to show the transition routes between possible system conditions in the form of an incident matrix that takes the form:

$$G = \begin{matrix} & r_1 & d & f_1 & t & v & g & n & r_2 & f_2 & s & k & p & l & z \\ \begin{matrix} u_1 \\ u_2 \\ u_3 \\ u_4 \\ u_5 \\ u_6 \\ u_7 \\ u_8 \\ u_9 \\ u_{10} \\ u_{11} \\ u_{12} \\ u_{13} \\ u_{14} \\ u_{15} \\ u_{16} \\ u_{17} \\ u_{18} \\ u_{19} \\ u_{20} \\ u_{21} \\ u_{22} \\ u_{23} \\ u_{24} \\ u_{25} \end{matrix} & \begin{pmatrix} -1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ -1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ -1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ -1 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & -1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & -1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & -1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & -1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & -1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & -1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & -1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & -1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & -1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & -1 \end{pmatrix} \end{matrix} \quad (3)$$

The matrix representation of the graph $G(X, U)$ opens the full essence of the interconnections between the various stages of the regional life security systems development. The resulting adjacency matrix and the incident matrix, based on it, allow us to construct a graph of possible system conditions in a geometric form. The geometric representation of the graph with the existing transition routes between possible system conditions is presented in Fig. 1. The subsets described by the expression (1) are shown as the graph nodes. The geometric representation illustrates graphically the possible information connections and the role of a particular set in the structure of the development process.

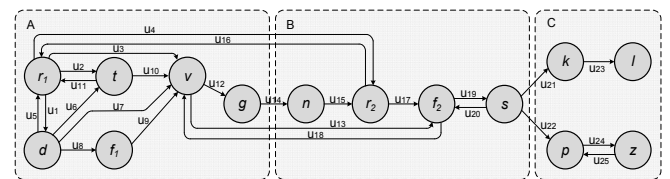


Fig. 1. Graph of possible system conditions in a geometric form

Figure 1 illustrates the connections and the relationships between the different conditions of the development process, which is divided into three main stages. The figure shows all the complexity of the process, which requires the establishment of appropriate information support at each stage. The following is a description of each of the stages in the context of its place in the system.

The first stage (A) determines the normative quantity and optimal places for the rescue teams. At this stage such criteria as the road network of the region $\{r_1\}$, the location of existing rescue units $\{d\}$, the rescue equipment on the balance sheet of rescue units $\{f_1\}$, the time of arrival to the most remote settlements of the region $\{t\}$ (it should not exceed 20 minutes) must be taken into account. According to these criteria, the areas of responsibility of existing and

projected (reformed) rescue units $\{v\}$ should be formed. This must provide the basis for determination of staffing number and optimal places of rescue teams' disposition $\{g\}$.

The second stage (B) involves determining the structure of existing and projected (reformed) rescue units. The main criteria determining the structure of existing and projected (reformed) rescue units are: the fire and manmade load of the analyzed region $\{n\}$ (for example: the presence of critical infrastructure, the average daily temperature in summer, the area of woodlands and dry woodland grass, the avalanche hazards, etc.), features of the region $\{r_2\}$ in terms of water sources' availability, quantity of high-rise buildings, etc. These criteria give grounds for determination of the types and required quantity of rescue equipment $\{f_2\}$. This will provide the basis for building the optimal staff structure of the unit $\{s\}$ (the required number of drivers, rescuers, etc.).

Assembling (equipping) of projected rescue teams (C) is the final stage in regional life security systems' development. The final stage involves determination of the qualification requirements for the personnel of the team $\{k\}$, training of the personnel at educational establishments of SES of Ukraine $\{l\}$, organization of property complexes' transfer from the state to communal property $\{p\}$ (real estate, fire and rescue vehicles, equipment, etc.), or, in the case of impossibility, their procurement $\{z\}$.

At the moment, three main issues of informational support are clearly expressed. Two of them are: determination of staffing number and optimal places of rescue teams' disposition (at the first stage), which depends on the information gathering and processing by the criteria $\{r_1\}$, $\{d\}$, $\{f_1\}$, $\{t\}$; determination of the types and required quantity of rescue equipment, as well as determination of the optimal staff structure (at the second stage), taking into account the criteria $\{n\}$, $\{r_2\}$. At the third stage (C), in addition to regulating the legal framework for the property complexes' transfer and training of staff at specialist educational establishments, the development of training programs in accordance with the proper qualification requirements must be performed. It depends on the criterion $\{k\}$. It is safe to declare that solving these tasks is impossible without proper informational support. Informational support for this process consists in collecting the necessary information, forming the knowledge base on it and using it as a source data for calculating $\{v\}$, $\{g\}$, $\{f_2\}$, $\{s\}$, $\{l\}$, $\{p\}$, $\{z\}$. Due to the complexity of the identified informational support process, a detailed analysis of the volume of data streams using the conceptual apparatus of the set theory should be conducted. The declared task is presented in the next section.

III. ANALYSIS OF THE VOLUME OF THE DATA STREAM

The volume of the data stream in the investigated case is proposed to analyze using the conceptual apparatus of the theory of sets, where the volume of relevant information, required for the work of the system, is presented in the form of sets (subsets). We will start the analysis from the geometric representation of the sets (A), (B) and (C) with the corresponding subsets. The model for collecting and processing information is considered in the form of universum U . The universum includes three sets with corresponding subsets, namely a set of information for determining the normative quantity and optimal places for the rescue units (A); a set of information for determining the structure of existing and projected (reformed) rescue units

(B); a set of information support for assembling (equipping) of projected rescue teams (C). Combining operations exist in the system of presented sets. They are commutative; therefore, taking into account the ratio of sets to the universum, their interconnection can be represented as follows:

$$U \supseteq \bigcup_{i=1}^3 (A, B, C) \neq \emptyset. \quad (4)$$

For a good understanding of the investigated environment and the structure of the model, we describe each of the sets in more detail.

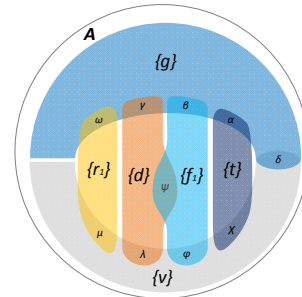


Fig. 2. A set of information for determining the normative quantity and optimal places for the rescue units

The set (A) in the model is represented as a comprehending set and contains the following subsets:

$$A = \{r_1, d, f_1, t, v, g\}, r_1 = \overline{1}, c, d = \overline{1}, i, f_1 = \overline{1}, j, t = \overline{1}, q, v = \overline{1}, m, g = \overline{1}, a, \quad (5)$$

where c - the volume of data on the road infrastructure; i - volume of data on the network of existing rescue units; j - the volume of data on varieties and volumes of available rescue equipment in existing rescue units; q - the volume of data about time and routes of arrival within the area of responsibility; m - the volume of data on the methodology for determining the boundaries of areas of responsibility; a - the volume of data on the methodology for determining normative quantity and optimal places for the rescue teams.

The next set (B) is also endowed with a number of combining and intersection operations.

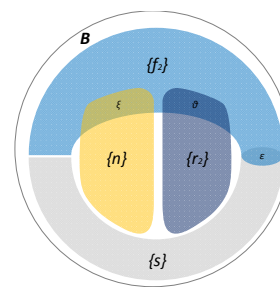


Fig. 3. A set of information for determining the structure of existing and projected (reformed) rescue units

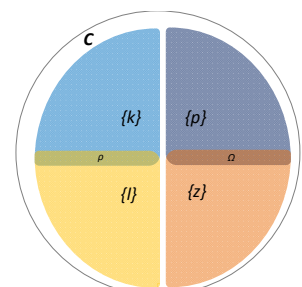


Fig. 4. A set of information for projected rescue teams' equipping

Similar to the previous case, the set (B) is represented as a comprehending set and contains the following subsets:

$$B = \{n, r_2, f_2, s\}, n = \overline{1}, x, r_2 = \overline{1}, e, f_2 = \overline{1}, y, s = \overline{1}, u, \quad (6)$$

where x - the volume of data on fire and manmade load of the analyzed region; e - the volume of data on the features of the analyzed region; y - the volume of data on methods for

determining the types and required quantity of rescue equipment; u – the volume of data on methods for determining the optimal staff structure of the rescue unit.

And the last set (C) expresses the assembling (equipping) of projected rescue teams. Let's analyze this set. The set (C) is also represented as a comprehending set and contains the following subsets:

$$C = \{k, l, p, z\}, k = \overline{1}, l = \overline{1}, h, p = \overline{1}, b, z = \overline{1}, w, \quad (7)$$

where o - the volume of data on determining the qualification requirements for the personnel of the team; h - the volume of data on personnel training programs due to qualification requirements; b – the volume of data on mechanisms for the property complexes' transfer; w – the volume of data on the mechanisms of the property complexes' procurement.

IV. SIMULATION OF THE STRUCTURE OF INTERCONNECTIONS IN INFORMATION FLOWS

The analysis of the volume of data streams gives us only a generalized presentation about the structure of the informational support process. For a more detailed study of these processes, it is necessary to study the interconnections of the data streams set and their structure. Considering the significant amount of obtained results, we will present only the final results of the research. In the set (A) both combining and intersection operations between subsets exist. Taking into account the peculiarities of sets and their interconnections, we make a generalized description of the existing connections between the data streams in the set (A):

$$[r_1 \cap (g \cap v)] \cup [d \cap (g \cap v) \cap f_1] \cup [t \cap (g \cap v)] \Rightarrow A \supseteq [r_1 \cup (d \cap f_1) \cup t] \cap (g \cap v) \neq \emptyset. \quad (8)$$

Expression (8) shows that determining normative quantity and optimal places for the rescue teams is closely connected (intersecting) with determining the boundaries of areas of responsibility. In turn, the sets $\{g\}$ and $\{v\}$ depend on the combining of $\{r_1\}$, $\{d\}$, $\{f_1\}$ and $\{t\}$. In the following a generalized description of the existing connections between the data streams in the set (B) is studied. It is submitted in a form of a model:

$$B \supseteq (n \cup r_2 \cup s) \cap f_2 \neq \emptyset. \quad (9)$$

Expression (9) clearly shows the complementarity of information (combining) between subsets $\{n\}$, $\{r_2\}$ and $\{s\}$ in the set (B), as well as their resulting effect (intersection) on the subset $\{f_2\}$. Although, in a time perspective, the volume of data on determining the optimal staff structure of the rescue unit is based on previously obtained data on the list of required rescue equipment. A generalized description of the existing connections between the data streams in the set (C) is as follows:

$$C \supseteq (k \cap l) \cup (p \cap z) \neq \emptyset. \quad (10)$$

Expression (10) shows a complementary connection between two intersections. As a result, the reformed rescue team can be produced by the qualified personnel $\{k\}$, $\{l\}$ and the necessary property and equipment $\{p\}$, $\{z\}$. Thus, the study of the interconnections of the data streams set and their structure in the investigated system shows the entire complexity of the regional security systems' development.

Data streams sets include combining and intersection of different sources of information that have the resulting influence on related information centres. It is necessary to note that most elements of the data stream are fundamental to the implementation of the next stage in the development system. Based on the result of this study we have formulated the necessity of developing the information system allowing to perform the data flow management in the projects of regional security systems' development

V. STRUCTURE OF THE INFORMATIONAL SYSTEM

The structure of the informational system should reflect the interconnections between input data stream that is processed using the proper techniques and the methods for obtaining (modelling) the final result. The final result is usually presented in the form of proposals for the number, location, structure and logistics of the rescue units. The prototype of an informational decision-making system should fulfil the following functions:

- gathering, arrangement (clustering) and preservation of data arrays about the features of the analyzed region;
- updating and editing data depending on external conditions;
- processing the input data depending on the task and the region;
- implementation of methods for determining key indicators regarding the number, location and optimal structure of rescue units;
- interaction with existing methods and approaches to the definition of individual components of the development process;
- validity check of the obtained results;
- formation of reports with proposals for the adoption of management decisions.

The above functions of the system allow the user formulating managerial decisions on the planning of activities for the regional life security systems' development. In accordance with the described sets of the data streams and the main functions of the informational system, the structure of such system should include:

- a unit of the user, which is responsible for input the initial data and obtaining the results in the form of the final report;
- a database for organizing, structuring and clustering the collected information;
- a knowledge base for appropriate processing of initial data;
- a module for estimation factors and initial modelling (determination of staffing number and optimal places of rescue teams' disposition);
- key modelling module (determination of the optimal staff structure of the unit);
- module of final modelling (proposals of equipping of projected rescue teams);
- module for analyzing the final result.

In accordance with the main elements of the structure of the informational system of project management in the areas of regional security systems' development, there is a need to

construct a graphical model. The structure is depicted in Figure 5.

VI. CONCLUSION

Through mathematical and geometric description of the basic processes of regional life security systems' development, the structural-logical scheme has been constructed in the form of a graph of possible system conditions. This scheme discovers the sequence of informational support of creating and reforming the rescue units. Geometrical description for the data stream model in the process of regional life security systems' development has been presented. For this purpose the conceptual

apparatus of the theory of sets has been used. The structure of interconnections inside the data streams set has been grounded. Using the simulation methods, the structure of the informational system of project management in the areas of regional security systems' development has been prepared. This system allows gathering, preservation and processing of initial data in order to make effective managerial decisions. Substantiation of appropriate methods and procedures for processing the initial data set, as well as development of software for the practical implementation of the declared informational system, has been further developed.

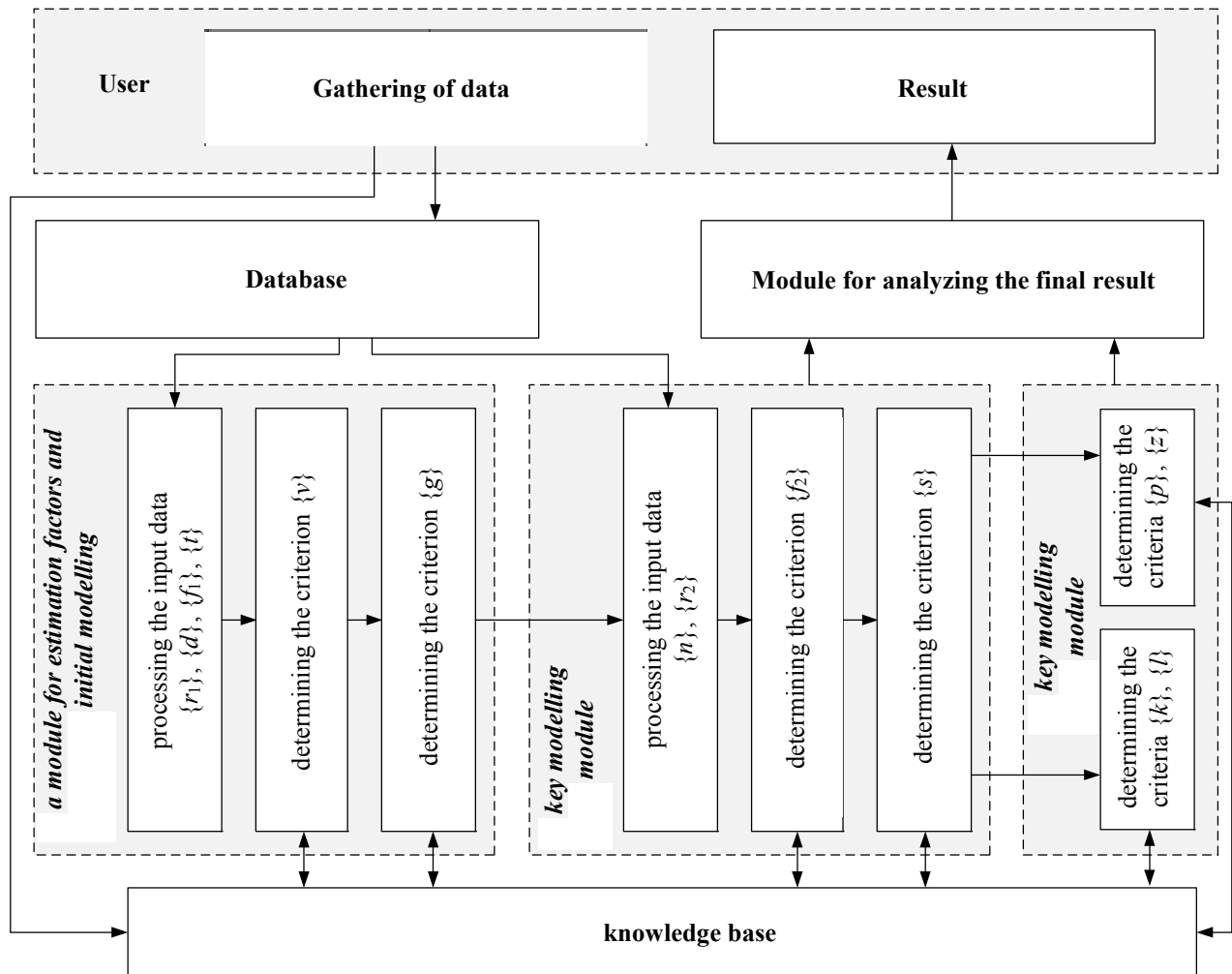


Fig. 5. The structure of the informational system of project management in the areas of regional security systems' development

REFERENCES

- [1] T. Ye. Rak, "Information technologies for assessing the professional level of operators for work in conditions of risk," Proceedings of IPEM of NAS of Ukraine, no. 62. pp. 202-207, 2012.
- [2] T. Ye. Rak, "Information technologies for monitoring the state of active production systems in extreme situations," Proceedings of IPEM of NAS of Ukraine, no.63. pp. 189-195, 2012.
- [3] O. B. Danchenko, and V. V. Lepskiy, "Modern models and methods of project, portfolio and program management", Management of the complex systems' development, no. 29, pp. 46-54, 2017.
- [4] O. G. Timinsky, and I. O. Marushchak, "Analysis of the value approach-based project of the regional network proactive creation," Management of the complex systems development, no 28, pp. 53 – 67, 2017.
- [5] A. A. Beloshchitskiy, Yu. I. Minaeva, and G. A. Filimonov, "Comparative models under uncertainty conditions for solving management problems," Management of the complex systems development, no. 25, pp. 91 – 95, 2016.
- [6] O. V. Bugrov, and O. O. Bugrova, "Functional analysis of regional development programs," Management of the complex systems development, no. 26. pp. 30 – 36, 2016.
- [7] Yu. I. Gritsyuk, I. O. Malets, and T. Ye. Rak, "Mathematical models of portfolio management for improvement of life safety systems," Computer Science and Information Technologies: Bulletin of the National University "Lviv Polytechnic", no. 672, pp. 110-119, 2010.
- [8] Yu. P. Rak, and O. B. Zachko, "Methods of analysis and estimation of the level of safety of vital activities of regions of Ukraine in

- conditions of realization of regional development projects," Project management and production development, no. 2 (26), pp. 29-39, 2008.
- [9] J. Lan, X. R. Li, Vesselin P. Jilkov, Chundi Mu. "Second-Order Markov Chain Based Multiple-Model Algorithm for Maneuvering Target Tracking," IEEE Transactions on Aerospace and Electronic Systems, vol. 49, no.1, pp.3-19, 2013.
- [10] J. Lan, X. R. Li, and Chundi Mu, "Best Model Augmentation for Variable-Structure Multiple-Model Estimation," IEEE Aerospace and Electronic Systems, vol. 47, no.3, pp.2008-2025, 2011
- [11] J. Lan, and X. R. Li, "Equivalent-Model Augmentation for Variable-Structure Multiple-Model Estimation," IEEE Aerospace and Electronic Systems, vol. 49, no.4, pp.2615-2630, 2013.
- [12] A. A. Matveev, D. A. Novikov, and A.V. Tsvetkov, "Project Portfolio Management Models and Methods," PMSOFT, Moscow, Russia. 2005.
- [13] A Guide to the Knowledge of Project Management (PMBOK®) – Fourth Edition [electronic resource]. – Available from <http://www.uapa.ru/media/uploads/attachment/source/2012/12/PMbok4.pdf>
- [14] Standard for project and program management P2M [electronic resource]. – Available from http://www.pmaj.or.jp/ENG/p2m/p2m_guide/p2m_guide.html
- [15] B. J. Weiner, "A theory of organizational readiness for change," Implementation Science, vol. 4, pp. 1-9. 2009. available at: www.ncbi.nlm.nih.gov/pmc/articles/PMC2770024/ (accessed December 14, 2012),
- [16] K. Malik, Human Development Report 2013. The Rise of the South: Human Progress in a Diverse World. Published for the United Nations Development Programme (UNDP), 2013. available: http://hdr.undp.org/sites/default/files/reports/14/hdr2013_en_compl etc.pdf.

Online Ranking Learning on Clusters

Leonid Lyubchik

*Computer Mathematics and Data Analysis Dept.
National Technical University "Kharkiv Polytechnic
Institute"*

Kharkiv, Ukraine
lyubchik@kpi.kharkiv.edu

Galyna Grinberg

*Economic Cybernetics and Management Dept.
National Technical University "Kharkiv Polytechnic
Institute"*

Kharkiv, Ukraine
glngrinberg@gmail.com

Abstract—Online data stream ranking learning problem is considered using training data in the form of a sequence of identical items series, described by a number of features and relative rank within the series. It is assumed that feature values and relative ranks of the same items may vary slightly for different series of observations, and there are stable groups of items with similar properties. In this regard, the problem of learning to rank on clusters is stated, while training dataset consist of estimates of centers of clusters and average rank of the items inside each cluster. A unified approach to ranking learning on clusters using kernel models of utility function is proposed. Recurrent algorithms for estimating the parameters of a utility function model as well as recurrent ranking learning algorithm in the space of conjugate variables are developed.

Keywords—data stream, kernel function, online learning, ranking learning, recurrent estimation, regularization, utility function.

I. INTRODUCTION

The problem of ranking learning has been intensively studied recently due to broad variety of important practical applications, ranging from classical problems of multi-criteria choice of alternatives and decision-making [1] to modern ones, such as information retrieval, machine translation, computational linguistics and biology [2]. The goal of learning to rank is automatic building of a ranking model using training data, which consists of lists of subject items with some partial order specified between items in each list, usually set by indicating some numerical or ordinal score for all items. It is assumed, that ranking mechanism, determined by preferences of users, experts or perhaps some artificial ranking system, is usually unknown. Therefore, ranking learning, in fact, is data-based modeling of this mechanism, so that the results of ranking of elements of a new lists will be similar to rankings in the training data in some sense.

Especially intensive the problem of ranking learning is studied in connection with the tasks of Internet data processing with the purpose of information retrieval. Ranking learning is widely used in such problems as document retrieval, recommendation system development, search engine modeling and over [3]. For example, in a document retrieval problem, for any given query, a ranking model assigns a relevance score to each document in obtained collection, and then ranks the documents in decreasing order of relevance scores. Therefore, the training data consists of queries and ranked sequence of documents. In this formulation, in particular, the important problem of adaptive modeling of search engine with unknown ranking

mechanism is considered, using information concern search results, corresponding to a certain sequence of queries [4].

Learning to rank belongs to the class of supervised machine learning problem, using given training sample consist of some items with measured features and labels, representing its ranks. The purpose of learning is to obtain some ranking function estimate, which provides similar ranking results on the test sample. In turn, ranking function is usually found by empirical risk minimization, determined by averaging of certain loss function on the training data sample [5].

As a ranking function model, in preference learning framework, it is often used latent utility function, describes expert or user preferences. Utility function is usually specified in the form of a scalar positive function defined in feature space, while larger values of the utility function correspond to larger values of the ranks.

In practice, linear models of utility function in the form of a weighted sum of features were widely used, while weights determine the relative features importance [6], these weights are found by applying expert or statistical methods.

In fact, the linear utility function model does not always adequately reflect the real structure of user preferences, and the structure of the utility function, reflecting the actual ranking mechanisms, can be significantly more complex.

At present, a number of heuristic approaches to the choice of utility function non-linear model structure are proposed as a fairly simple functional dependencies [7, 8], but revealing its true form remains a difficult task.

More general non-linear models of utility functions can be chosen in the form of a linear combination of some pre-determined coordinate functions. In order to build a qualitative approximate model, it is necessary to use a large number of coordinate functions, the consequence is the need for high dimension model parameters vector estimating, which leads to significant computational problems.

Since the latent utility function can have a very complex structure and previous information about its structure is usually absent, it is advisable to use kernel-based machine learning technique [9]. In the framework of this method, utility functions estimate can be represented as a linear combination of kernel functions at training points. However, due to "kernel trick", there is no need for preliminary specification of a set of coordinate functions, which makes it possible to build models of limited complexity that successfully approximate rather complex utility functions [10].

In many practical applications training data are generated as a data streams in the form of consecutive series of observations, arrives continuously in variable time [11]. Under the frequently changing items features and ranking results, it is advisable to use online ranking learning methods that provide the opportunity for effective training of ranking model in real time [12, 13].

The peculiarity of considered problem is in the fact that for the same items in different series of data streams the observed values of features and relative ranks can vary, because its properties and user preferences can change over time. Consequently, it is impossible to assign to each particular item the exact rank and it is reasonable to use some averaged ranks as supervised information. It is assumed that feature values and relative ranks of the same items may vary slightly for different series of observations, and there are stable groups of items with similar properties. This predetermines the need for prior aggregation of the ranked items into certain groups of similar properties by clustering them in feature space. This aggregation allows moving from the problem of ranking objects to the problem of ranking clusters. In this regard, it is reasonable to use *ranking learning on clusters* approach [14], based on preliminary clustering of ranking items followed by utility function model building using average ranks of items inside each cluster.

In the tasks of online ranking learning, the data stream is formed as sequences of series of the same items consist from number of observations of features and relative ranks within each series. The specificity of online ranking learning using data stream requires learning algorithms in recurrent form, wherein the number of the iteration step coincides with the number of series of observations.

In this paper, a unified approach to the cluster ranking recurrent learning algorithms development using kernel models of utility function is proposed. First, recursive algorithms for estimating the parameters of a linear utility function model are considered. In this case, expert estimates of features weights are used as *a priori* information for regularizing the estimation problem, realizing, in fact, optimal concordation of expert and statistical estimates [9, 15]. Further, based on kernel approach, a learning non-linear utility function model is obtained, while the estimates of the parameters of the linear model are used for regularization of optimized functional of empirical risk. On top of that, another one algorithm of recurrent ranking learning in the space of conjugate variables is also proposed.

II. PROBLEM STATEMENT

Consider the ranking learning problem for the set of same items $x \in \Omega$, characterized by its feature vector $\mathbf{x}^T = (x^1, x^2, \dots, x^N)$.

Supposed that training data are generated in real time and is representable as a sequence of observations series. Each series includes observations on the entire set of ranked items and has fixed length L .

In such a case, in each data stream series n training dataset are presented by the data matrix $\mathbf{X}_n = \{x_i^j(n)\}_{i,j=1}^{L,N}$, consists from feature observations

$$\mathbf{x}_i^T(n) = (x_i^1(n), x_i^2(n), \dots, x_i^N(n)), \quad i=1, \dots, L.$$

Each item within the any stream observation series is assigned its relative rank $r_n(\mathbf{x}_i(n))$, $1 \leq i \leq n$, defined by some ranking function. It is assumed that the specified ranking function is unknown, and only relative ranks for any objects in each series are available to observation.

The ranking function is usually described by some scalar positive continuous utility function $f(\mathbf{x})$, such that $f(\mathbf{x}_i) > f(\mathbf{x}_j)$ if $r_n(\mathbf{x}_i(n)) < r_n(\mathbf{x}_j(n))$.

The problem of data stream learning to rank is to restore unknown utility function by finding its estimate $\hat{f}(\mathbf{x})$ using available sequence of observations series $\{\mathbf{X}_n, \mathbf{r}_n\}$, $\mathbf{r}_n^T = (r_n(\mathbf{x}_1(n)), \dots, r_n(\mathbf{x}_L(n)))$.

Suppose that using the appropriate clustering method the set of ranking items is divided into a set of M clusters $\{\Omega_m\}_{m=1}^M$ in feature space, described by a set of parameters $(\bar{\mathbf{x}}^m, \bar{r}^m)$, $m=1, M$

$$\bar{\mathbf{x}}^m = \frac{1}{|J^m|} \sum_{i \in \Omega_m} \mathbf{x}_i, \quad \bar{r}^m = \frac{1}{|J^m|} \sum_{i \in \Omega_m} r(\mathbf{x}_i), \quad (1)$$

where $\bar{\mathbf{x}}^m$ is center vector of m -th cluster and \bar{r}^m is average rank of items belonging to the same cluster, $J^m = \{i | x_i \in \Omega_m\}$, $m=1, M$.

Then, to restore the utility function, aggregated training data is used in the form of a sequence of estimated centers of clusters and average ranks of items inside the corresponding cluster for each data stream series of observations $\bar{\mathbf{X}}_n^T = (\bar{\mathbf{x}}^1(n), \dots, \bar{\mathbf{x}}^M(n))$, $\bar{\mathbf{r}}_n^T = (\bar{r}^1(n), \dots, \bar{r}^M(n))$.

We take utility function model in the quasilinear form $f(\mathbf{x}) = \varphi^T(\mathbf{x})\mathbf{c}$, where $\mathbf{c}^T = (c^1, \dots, c^D)$ – vector of utility model parameters, $\varphi^T(\mathbf{x}) = (\varphi^1(\mathbf{x}), \dots, \varphi^D(\mathbf{x}))$ – vector of model coordinate functions, D is a model dimension.

Then the problem of ranking learning based on streaming data by restoring the utility function on clusters is reduced to finding model parameters estimates $\hat{\mathbf{c}}_n$ using a sequence of aggregated streaming training data $\{\bar{\mathbf{X}}_n, \bar{\mathbf{r}}_n\}$, $n=1, 2, \dots$

In the kernel-based learning framework coordinate functions are taken hereby that its scalar products will be positive definite functions $\varphi^T(\mathbf{x})\varphi(\mathbf{x}') = \kappa(\mathbf{x}, \mathbf{x}')$, at that utility function model are linear combination of kernel function, located in centers of clusters

$$\hat{f}(\mathbf{x}) = \sum_{m=1}^M d_m \cdot \kappa(\mathbf{x}, \bar{\mathbf{x}}^m), \quad (2)$$

where d_m , $m=1, M$ – kernel-based utility function model parameters are determined by center and average rank of clusters estimates.

Then the problem of online rankings kernel-based learning on clusters is reduced to the construction of a recurrent algorithm for estimating parameters d_m , $m = \overline{1, M}$ of kernel model (2) based on training data stream $\{\bar{\mathbf{X}}_n, \mathbf{r}_n\}$, $n = 1, \dots$.

III. LINER UTILITY FUNCTION MODEL IDENTIFICATION

We first consider the problem of estimating the parameters of the linear model of utility function

$$f(\mathbf{x}) = \mathbf{x}^T \mathbf{w}, \quad (3)$$

where $\mathbf{w}^T = (w^1, w^2, \dots, w^N)$ – vector of linear utility function model parameters.

To construct model parameter estimates, streaming training data $\{\bar{\mathbf{X}}_n, \mathbf{r}_n\}$, $n = 1, \dots$ is used. Because the elements of this sequence are random vectors and matrices, at the data preprocessing stage it is advisable to smooth the sequence of the training sample elements using a suitable current averaging algorithm, for example, the method of exponential smoothing:

$$\begin{aligned} \hat{\mathbf{X}}_{n+1} &= \tau_x \cdot \hat{\mathbf{X}}_n + (1 - \tau_x) \cdot \bar{\mathbf{X}}_{n+1}, \\ \hat{\mathbf{r}}_{n+1} &= \tau_r \cdot \hat{\mathbf{r}}_n + (1 - \tau_r) \cdot \bar{\mathbf{r}}_{n+1}, \quad 0 < \tau_x, \tau_r < 1. \end{aligned} \quad (4)$$

Then the problem of estimating the parameters of the linear model of utility function is reduced to the problem of multi-dimensional dynamic linear regression

$$\hat{\mathbf{r}}_{n+1} = \hat{\mathbf{X}}_{n+1} \mathbf{w} + \mathbf{e}_{n+1}^w, \quad (5)$$

where $\mathbf{e}_{n+1}^w = (e_{n+1}^1, e_{n+1}^2, \dots, e_{n+1}^M)^T$ – vector of average rank on clusters estimation errors.

Let us find the current estimate of the vector of linear model parameters from the condition of minimization of a one-step regularized functional with constraints

$$\begin{aligned} R_{n+1}^w(\mathbf{w}) &= \|\mathbf{e}_{n+1}\|^2 + \alpha \|\mathbf{w} - \hat{\mathbf{w}}_n\|^2 \rightarrow \min_w, \\ \mathbf{e}_{n+1}^w &= \hat{\mathbf{r}}_{n+1} - \hat{\mathbf{X}}_{n+1} \mathbf{w}, \end{aligned} \quad (6)$$

where $\alpha > 0$ – regularization parameter, and the previous estimation of the parameter vector $\hat{\mathbf{w}}_n$ is used as *a priori* information for regularization at data stream series $n+1$, which provides the possibility of obtaining a recurrent estimate.

To solve the optimization problem with constraints, we use Lagrange function

$$\begin{aligned} L(\mathbf{w}, \mathbf{e}_{n+1}^w, \boldsymbol{\mu}) &= \\ &= 0.5 \cdot R_{n+1}^w(\mathbf{w}) + \boldsymbol{\mu}^T (\hat{\mathbf{r}}_{n+1} - \hat{\mathbf{X}}_{n+1} \mathbf{w} - \mathbf{e}_{n+1}^w), \end{aligned} \quad (7)$$

where $\boldsymbol{\mu}$ are Lagrange multipliers.

Using optimality conditions for (7) in the form of the Kuhn-Tucker:

$$\begin{aligned} \alpha \cdot (\mathbf{w} - \hat{\mathbf{w}}_n) - \hat{\mathbf{X}}_{n+1}^T \boldsymbol{\mu} &= 0, \quad \boldsymbol{\mu} = \mathbf{e}_{n+1}^w, \\ \hat{\mathbf{r}}_{n+1} - \hat{\mathbf{X}}_{n+1} \mathbf{w} - \mathbf{e}_{n+1}^w &= 0, \end{aligned} \quad (8)$$

we obtain an explicit expression for $\hat{\mathbf{w}}_{n+1}$ estimate in the form of a recurrent estimation algorithm

$$\begin{aligned} \hat{\mathbf{w}}_{n+1} &= \Psi_{n+1}^{-1}(\alpha) \cdot (\alpha \cdot \hat{\mathbf{w}}_n + \hat{\mathbf{X}}_{n+1}^T \hat{\mathbf{r}}_{n+1}), \\ \Psi_{n+1}^{-1}(\alpha) &= \alpha \cdot \mathbf{I}_N + \hat{\mathbf{X}}_{n+1}^T \hat{\mathbf{X}}_{n+1}. \end{aligned} \quad (9)$$

The obtained algorithm, which relates to the class of one-step regularized projection identification algorithms, allows tracing slow changes in cluster parameters, and the choice of the regularization parameter provides a balance between its tracking and filtering properties (9).

IV. RECURSIVE NONLINEAR PREFERENCE LEARNING

Measurement equation for quasilinear utility function model identification at stream series $n+1$ may be represented as following:

$$\hat{\mathbf{r}}_{n+1}^m = \hat{f}(\hat{\mathbf{x}}_{n+1}^m) = \boldsymbol{\varphi}^T(\hat{\mathbf{x}}_{n+1}^m) \mathbf{c} + e_{n+1}^m, \quad m = \overline{0, M}. \quad (10)$$

In matrix form this equation is $\hat{\mathbf{r}}_{n+1} = \boldsymbol{\Phi}_{n+1}^T \mathbf{c} + \mathbf{e}_{n+1}$,

where $\hat{\mathbf{r}}_{n+1} = (\hat{r}_{n+1}^1, \hat{r}_{n+1}^2, \dots, \hat{r}_{n+1}^M)^T$ – observation vector, composed from average rank on clusters estimates, $\boldsymbol{\Phi}_{n+1} = (\boldsymbol{\varphi}(\hat{\mathbf{x}}_{n+1}^1), \boldsymbol{\varphi}(\hat{\mathbf{x}}_{n+1}^2), \dots, \boldsymbol{\varphi}(\hat{\mathbf{x}}_{n+1}^M))$ – feature matrix estimate, $\mathbf{e}_{n+1} = (e_{n+1}^1, e_{n+1}^2, \dots, e_{n+1}^M)^T$ – average rank estimation errors.

Introduce kernel matrix $\mathbf{K}_{n+1} = \boldsymbol{\Phi}_{n+1}^T \boldsymbol{\Phi}_{n+1}$,

$$\mathbf{K}_n = \| \| k_{q,s} \| \|, \quad k_{q,s} = \kappa(\hat{\mathbf{x}}_n^q, \hat{\mathbf{x}}_n^s), \quad q, s = \overline{1, M},$$

where $\kappa(\mathbf{x}, \mathbf{x}')$ is an appropriate kernel function.

The utility function model parameters estimates $\hat{\mathbf{c}}_{n+1}$ at any data stream series $n+1$ may be obtained as a solution of regularized constrained optimization problem

$$\begin{aligned} R_{n+1}^c(\mathbf{c}) &= \|\mathbf{e}_{n+1}\|^2 + \beta \|\mathbf{c} - \mathbf{c}_{n+1}^0\|^2 \rightarrow \min_c, \\ \mathbf{e}_{n+1} &= \hat{\mathbf{r}}_{n+1} - \boldsymbol{\Phi}_{n+1}^T \mathbf{c}, \end{aligned} \quad (11)$$

where \mathbf{c}_{n+1}^0 – vector of *a priori* value of utility function model parameters for data stream series $n+1$, $\beta > 0$ – regularization parameter.

To solve the optimization problem (11) we use the Lagrange function

$$L(\mathbf{c}, \mathbf{e}_{n+1}, \boldsymbol{\lambda}) = 0.5 \cdot R_n(\mathbf{c}) + \boldsymbol{\lambda}^T (\mathbf{r}_{n+1} - \Phi_{n+1}^T \mathbf{c} - \mathbf{e}_{n+1}), \quad (12)$$

where $\boldsymbol{\lambda}^T = (\lambda_1, \dots, \lambda_M)$ – vector of Lagrange multipliers.

Using the conditions for optimality for (12)

$$\begin{aligned} \mathbf{c} &= \mathbf{c}_{n+1}^0 + \beta^{-1} \Phi_{n+1} \boldsymbol{\lambda}, \quad \boldsymbol{\lambda} = \mathbf{e}_{n+1}, \\ \hat{\mathbf{r}}_{n+1} - \Phi_{n+1}^T \mathbf{c} &= \mathbf{e}_{n+1}, \end{aligned} \quad (13)$$

model parameters and conjugate variables optimal estimates can be presented in the form

$$\begin{aligned} \hat{\mathbf{c}}_{n+1} &= \Phi_{n+1} \mathbf{A}_{n+1}^{-1}(\beta) \hat{\mathbf{r}}_{n+1} + (\mathbf{I}_D - \Phi_{n+1} \mathbf{A}_{n+1}^{-1}(\beta) \Phi_{n+1}^T) \mathbf{c}_{n+1}^0, \\ \hat{\boldsymbol{\lambda}}_n &= \mathbf{A}_{n+1}^{-1}(\beta) (\hat{\mathbf{r}}_{n+1} - \Phi_{n+1}^T \mathbf{c}_{n+1}^0), \\ \mathbf{A}_{n+1}(\beta) &= \beta^{-1} \mathbf{I}_M + \mathbf{K}_{n+1}. \end{aligned} \quad (14)$$

The use of kernel approach requires the elimination of direct evaluation of model parameters. To do this, we express *a priori* value of utility function model parameters \mathbf{c}_{n+1}^0 through available estimates, as which we choose utility function linear approximation parameters $\hat{\mathbf{w}}_{n+1}$ estimates defined by algorithm (9).

To do this, we take the linear model $f^0(\mathbf{x}) = \mathbf{x}^T \mathbf{w}$ as the first approximation of utility function. In accordance with this assumption, we find an optimal *a priori* value of utility function model parameters \mathbf{c}_{n+1}^0 at stream series $n+1$ from the condition of best approximation of linear utility function model values $\hat{\mathbf{r}}_{n+1}^0 = \hat{\mathbf{X}}_{n+1} \hat{\mathbf{w}}_{n+1}$, estimated on measured data $\hat{\mathbf{X}}_{n+1}$, by *a priori* average rank vector $\mathbf{r}_{n+1}^0 = \Phi_{n+1}^T \mathbf{c}_{n+1}^0$.

Consequently, to find *a priori* value \mathbf{c}_{n+1}^0 , consider auxiliary optimization problem for regularized functional:

$$\begin{aligned} Q_0(\mathbf{c}^0) &= \|\zeta\|^2 + \gamma \|\mathbf{c}^0\|^2 \rightarrow \min_{\mathbf{c}^0}, \\ \zeta &= \hat{\mathbf{r}}_{n+1}^0 - \mathbf{r}_{n+1}^0 = \hat{\mathbf{X}}_{n+1} \hat{\mathbf{w}}_{n+1} - \Phi_{n+1}^T \mathbf{c}^0. \end{aligned} \quad (15)$$

where $\gamma > 0$ – regularization parameter.

Using Lagrange function for constrained optimization problem

$$L(\mathbf{c}_0, \zeta, \mathbf{v}) = 0.5 \cdot Q_0(\mathbf{c}_0) + \mathbf{v}^T (\hat{\mathbf{X}}_{n+1} \hat{\mathbf{w}}_{n+1} - \Phi_{n+1}^T \mathbf{c}_0 - \zeta), \quad (16)$$

where $\mathbf{v}^T = (v_1, \dots, v_n)$ – appropriate vector of Lagrange multipliers, we can obtain its solution of problem (15) as

$$\begin{aligned} \mathbf{c}_{n+1}^0 &= \Phi_{n+1} \mathbf{B}_{n+1}^{-1}(\gamma) \hat{\mathbf{X}}_{n+1} \hat{\mathbf{w}}_{n+1}, \\ \mathbf{B}_{n+1}(\gamma) &= \gamma \mathbf{I}_M + \mathbf{K}_{n+1}. \end{aligned} \quad (17)$$

Taking into account the obvious kernel relation

$$\begin{aligned} \chi_{n+1}^T(\mathbf{x}) &= \boldsymbol{\varphi}^T(\mathbf{x}) \cdot \Phi_{n+1} = \\ &= (\kappa(\mathbf{x}, \hat{\mathbf{x}}_{n+1}^1), \kappa(\mathbf{x}, \hat{\mathbf{x}}_{n+1}^2), \dots, \kappa(\mathbf{x}, \hat{\mathbf{x}}_{n+1}^M))^T, \end{aligned} \quad (18)$$

the optimal utility function nonlinear model estimate $\hat{f}_{n+1}(\mathbf{x}) = \boldsymbol{\varphi}^T(\mathbf{x}) \hat{\mathbf{c}}_{n+1}$ takes the following form:

$$\begin{aligned} \hat{f}_{n+1}(\mathbf{x}) &= \chi_{n+1}^T(\mathbf{x}) \cdot (\mathbf{A}_{n+1}^{-1}(\beta) \hat{\mathbf{r}}_{n+1} + \\ &+ [\mathbf{I}_M - \mathbf{A}_{n+1}^{-1}(\beta) \mathbf{K}_{n+1}]) \mathbf{B}_{n+1}^{-1}(\gamma) \hat{\mathbf{X}}_{n+1} \hat{\mathbf{w}}_{n+1} \end{aligned} \quad (19)$$

Thus, utility function model estimate depends only of kernel function, located in the estimated clusters centers

$$\hat{f}_{n+1}(\mathbf{x}) = \sum_{m=1}^M d_m(\hat{\mathbf{x}}_{n+1}, \hat{\mathbf{r}}_{n+1}, \hat{\mathbf{w}}_{n+1}) \cdot \kappa(\mathbf{x}, \hat{\mathbf{x}}_{n+1}^m), \quad (20)$$

where kernel model coefficients d_m , $m = \overline{1, M}$ are calculated recurrently using available data stream.

V. CONJUGATE VARIABLES LEARNING ALGORITHM

A different way of estimating the parameters of the utility function can be proposed using the algorithm for recurrent estimation of conjugate variables [16].

Using the relation (13) with the parameters corresponding to stream series $n+1$ $\mathbf{c}_{n+1} = \beta^{-1} \Phi_{n+1} \boldsymbol{\lambda}_{n+1} + \mathbf{c}_{n+1}^0$ and multiplying (13) by matrix Φ_{n+1}^T , we obtain corresponding measurement equation for conjugate variables

$$\begin{aligned} \hat{\mathbf{r}}_{n+1} &= \beta^{-1} \mathbf{K}_{n+1} \boldsymbol{\lambda}_{n+1} + \Phi_{n+1}^T \mathbf{c}_{n+1}^0 + \mathbf{e}_{n+1}, \\ \mathbf{c}_{n+1}^0 &= \Phi_{n+1} \mathbf{B}_{n+1}^{-1}(\gamma) \hat{\mathbf{X}}_{n+1} \hat{\mathbf{w}}_{n+1}. \end{aligned} \quad (21)$$

Introduce a local identification criterion [16] as the moving estimation cost includes regularization term, determined by conjugate variables estimate $\hat{\boldsymbol{\lambda}}_n$ at previously data stream series $n+1$:

$$\begin{aligned} R_{n+1}(\boldsymbol{\lambda}_{n+1}) &= \delta \cdot (\boldsymbol{\lambda}_{n+1} - \hat{\boldsymbol{\lambda}}_n)^T \mathbf{K}_{n+1} (\boldsymbol{\lambda}_{n+1} - \hat{\boldsymbol{\lambda}}_n) + \\ &+ \left\| \hat{\mathbf{r}}_{n+1} - \beta^{-1} \mathbf{K}_{n+1} \boldsymbol{\lambda}_{n+1} - \mathbf{K}_{n+1} \mathbf{B}_{n+1}^{-1}(\gamma) \hat{\mathbf{X}}_{n+1} \hat{\mathbf{w}}_{n+1} \right\|^2. \end{aligned} \quad (22)$$

where $\delta > 0$ – regularization parameter.

Such a choice of the regulariser restricts the rate of change of estimates $\hat{\boldsymbol{\lambda}}_n$, which ensures effective smoothing of conjugate variables estimates.

Condition of optimality for the problem of functional (22) minimization, leads to normal matrix equations

$$\begin{aligned} (\beta^{-2} \mathbf{K}_{n+1}^T \mathbf{K}_{n+1} + \delta \cdot \mathbf{K}_{n+1}) \boldsymbol{\lambda}_{n+1} &= \delta \mathbf{K}_{n+1} \hat{\boldsymbol{\lambda}}_n + \\ &+ \beta^{-1} \mathbf{K}_{n+1}^T [\hat{\mathbf{r}}_{n+1} - \mathbf{K}_{n+1} \mathbf{B}_{n+1}^{-1}(\gamma) \hat{\mathbf{X}}_{n+1} \hat{\mathbf{w}}_{n+1}]. \end{aligned} \quad (23)$$

From (13), (21), (23) we obtain recurrent estimation algorithm for conjugate variables

$$\begin{aligned} \hat{\lambda}_{n+1} = & \delta \cdot \mathbf{A}_{n+1}^{-1}(\rho) \cdot \mathbf{K}_{n+1} \cdot \hat{\lambda}_n + \\ & + \beta \cdot \mathbf{A}_{n+1}^{-1}(\rho) \cdot (\hat{\mathbf{r}}_{n+1} - \beta \cdot \mathbf{K}_{n+1} \mathbf{B}_{n+1}^{-1}(\gamma) \mathbf{X}_{n+1} \hat{\mathbf{w}}_{n+1}), \end{aligned} \quad (24)$$

where $\rho = \beta^2 / \delta$.

Finally, the utility function model estimate based on the recursively estimated conjugate variables can be represented as

$$\begin{aligned} \hat{f}_{n+1}(x) = & \chi_{n+1}^T(x) \cdot [\rho^{-1} \mathbf{A}_{n+1}^{-1}(\rho) \mathbf{K}_{n+1} \hat{\lambda}_n + \\ & + \mathbf{A}_{n+1}^{-1}(\rho) \hat{\mathbf{r}}_{n+1} - \\ & - (\mathbf{I}_M - \mathbf{A}_{n+1}^{-1}(\rho) \mathbf{K}_{n+1}) \mathbf{B}_{n+1}^{-1}(\gamma) \hat{\mathbf{X}}_{n+1} \hat{\mathbf{w}}_{n+1}]. \end{aligned} \quad (25)$$

The convergence of recurrent algorithm for estimating conjugate variables (26) can be provided by an appropriate choice of the regularization parameter δ .

Thus, recurrent algorithms (19) and (26) define computational procedures for identifying kernel models of utility function on clusters in feature space that can be used to find estimates of the ranks of new items with similar features.

VI. CONCLUSION

The proposed approach to the problem of online learning to rank using training data stream is based on combining of dynamic items clustering in feature space and recurrent utility function estimating on clusters. The obtained model of learning ranking function is a linear combination of kernel functions with recurrently tuning parameters, at that the complexity of utility function model is determined by the number of clusters.

Implementation of the proposed method of online ranking learning on clusters first of all involves the improvement of algorithms of both cluster parameter and average clusters rank estimating. For this purpose it seems expedient to use semi-supervised clustering methods with partial use of data describes relative ranks. Further development of the proposed approach can be carried out in the direction of

optimizing the number of clusters in feature space using the complexity ratings of learning ranking model.

REFERENCES

- [1] . Figueira, S. Greco and M. Ehrgott, Multiple criteria decision analysis: state of the art surveys. Springer, Switzerland, 2005.
- [2] H. Li, Learning to rank for information retrieval and natural language processing. Morgan & Claypool, Toronto, 2011.
- [3] T. Y. Liu, "Learning to rank for information retrieval," Foundations and trends in information retrieval, vol. 3, no. 3, pp.225–331, 2009.
- [4] L. Hang, "A Short Introduction to Learning to Rank," IEICE Trans. Information & Systems, vol. E94-D, no.10, October 2011.
- [5] J. Furnkranz and E. Hullermeier, "Pairwise preference learning and ranking," Lecture Notes in Computer Science, vol. 2837, pp. 145-156, 2003.
- [6] V. Noghin, "Relative importance of criteria: a quantitative approach," Journal Multi-Criteria Decision Analysis, v. 6, pp. 355-363, 1997.
- [7] J. Barzilai, "Measurement and preference function modeling", Intern. Trans. in Operational Research, vol. 12, pp. 173-183, 2005.
- [8] V. Podvezko, and A. Podvezko, "Dependence of multicriteria evaluation result on choice of preference functions and their parameters," Technological and Economic Development of Economy. Baltic Journal of Sustainability, 16(1), pp.143–158, 2010.
- [9] L. Lyubchik and G. Grinberg, "Preference function reconstruction for multiple criteria decision making based on machine learning Approach," Recent developments and new directions in Soft Computing, L.A. Zadeh et al. (Eds), Springer, pp. 53-63, 2014.
- [10] M. Espinoza, J. Suykens, and B. Moor, "Kernel based partially linear models and nonlinear identification," IEEE Transactions on Automatic Control, 2005, vol. 50 (10), pp. 1602–1606.
- [11] J. A. Papini, S. de Amo, A. Kardec and S. Soares, "FPS Mining: A fast algorithm for mining user preferences in data streams," Journal of Information and Data Management, vol. 5, no. 1, pp. 4–15, February 2014.
- [12] K. Hofmann, S. Whiteson, and M. de Rijke, "Balancing exploration and exploitation in learning to rank online", In ECIR 2011: 33-rd European Conference on Information Retrieval. Springer, pp, 251-263, April 2011.
- [13] K. Hofmann, A. Schuth, S. Whiteson, and M. de Rijke, "Reusing historical interaction data for faster online learning to rank for information retrieval", In WSDM 2013: International Conference on Web Search and Data Mining, ACM, 2013.
- [14] L. Lyubchik and G. Grinberg, "Real time recursive preference learning to rank from data stream," 1-th IEEE International Conference on Data Stream Mining & Processing, Lviv, Ukraine, pp. 280=285, 2016.
- [15] V. Strijov, and V. Shakin, "Index construction: the expert-statistical method," Environmental research, engineering and management, no. 4 (26), pp.51-55, 2003.
- [16] L. Lyubchik, V. Kolbasin and R. Shafeev, "Nonlinear signal reconstruction based on recursive moving window kernel method," 8-th International Conference on Intelligent Data Acquisition and Advanced Computing Systems (IDAACS'2015), Warsaw, Poland, pp. 298-302, 2015.

Time Series Classification Based on Fractal Properties

Vitalii Bulakh
dept. Applying mathematics
Kharkiv National University of
Radioelectronics
Kharkiv, Ukraine
bulakhvitalii@gmail.com

Lyudmyla Kirichenko
dept. Applying mathematics
Kharkiv National University of
Radioelectronics
Kharkiv, Ukraine
lyudmyla.kirichenko@nure.ua

Tamara Radivilova
dept. of Infocommunication
Engineering
Kharkiv National University of
Radioelectronics
Kharkiv, Ukraine
tamara.radivilova@nure.ua

Abstract – The article considers classification task of fractal time series by the meta algorithms based on decision trees. Binomial multiplicative stochastic cascades are used as input time series. Comparative analysis of the classification approaches based on different features is carried out. The results indicate the advantage of the machine learning methods over the traditional estimating the degree of self-similarity.

Keywords – multifractal time series, binomial stochastic cascade, classification of time series, Hurst exponent, Random Forest

I. INTRODUCTION

Many complex processes have a fractal structure and their dynamics are represented by time series possessing fractal properties. Such processes include various information processes in communication networks, including hacker attacks.

Distributed Denial of Service (DDoS) attack is hacker attack on the computer system in order to force it to failure that is creation of such conditions at which conscientious system users can not access the provided system resources, or this access is difficult. Failure of the system can be a step towards mastering it. Currently, DoS and DDoS-attacks are the most popular, since they allow to force almost any system to failure without leaving legally significant evidence. One solution to the problem of early intrusion detection is the development of a classifier that would determine the probability that incoming traffic contains an attack.

Recent researches show that network traffic has fractal properties and one of the characteristic features of intrusion is changing of fractal characteristics in the traffic containing attack. In [1-3] a number of methods of detection based on changing traffic fractal structure containing attacks. An important step in the intrusion detection methods is classification of the test traffic on the fractal properties, in particular, range of values of the Hurst parameter.

In many cases, the problems of recognizing and classifying fractal series take place. Most often, such tasks are solved by estimating and analyzing fractal characteristics [4-7]. However, in recent years, there has been a growing interest in machine learning methods to analyze and classify fractal series [8-10]. The aim of this paper is a comparative analysis of the classification of fractal stochastic time series performed by meta-algorithms using decision tree methods.

II. CHARACTERISTICS OF SELF-SIMILAR AND MULTIFRACTAL PROCESSES

The self-similarity of random processes is to preserve statistical characteristics when changing the time scale. A stochastic process $X(t)$ is self-similar with a parameter H if the process $a^{-H}X(at)$ is described by the same finite-dimensional distribution laws as $X(t)$. The parameter H , $0 < H < 1$, called the Hurst exponent, represents the self-similarity degree. Along with this property, the Hurst exponent characterizes the measure of the long-term dependence of the stochastic process. The moments of the self-similar random process satisfy the following scaling relation

$$E\left[|X(t)|^q\right] \propto t^{qH}. \quad (1)$$

Multifractal random processes are inhomogeneous fractal ones and have more flexible scaling laws for moment characteristics [11]:

$$E\left[|X(t)|^q\right] \propto t^{qh(q)}, \quad (2)$$

where $h(q)$ is generalized Hurst exponent, in the general case, a nonlinear function for which the value $h(q)$ at $q = 2$ coincides with the value of Hurst exponent H . For time series that correspond to a monofractal process, the generalized Hurst exponent $h(q)$ is constant: $h(q) = H$.

III. ESTIMATING OF FRACTAL CHARACTERISTICS BY TIME SERIES

There are many methods for estimating the fractal characteristics by time series. One of the most popular is the method of multifractal detrended fluctuation analysis (MFDFA) [12]. MFDFA is powerful tool for statistical processing of time-dependent processes.

According to the MFDFA method, the fluctuation function $F^2(\tau) = \frac{1}{\tau} \sum_{t=1}^{\tau} (y(t) - Y_m(t))^2$ is calculated for each

segment of length τ , where $y(t)$ is input cumulative time series; function $Y_m(t)$ is local m -polynomial trend within the given segment. If the investigated series $y(t)$ has a long-term dependence the averaged function $F(\tau)$ has scaling $F(\tau) \propto \tau^H$.

In the study of multifractal properties the dependence of the fluctuation function $F_q(\tau)$ of a parameter q is

considered: $F_q(\tau) = \left\{ \frac{1}{N} \sum_{i=1}^N [F^2(\tau)]^{\frac{q}{2}} \right\}^{\frac{1}{q}}$. If the investigated

series is multifractal and has a long-term dependence, the fluctuation function $F_q(\tau)$ has scaling $F_q(\tau) \propto \tau^{h(q)}$. Estimate of the Hurst exponent H can be represented by confidence interval within which the true value H is found [13-15]:

$$\hat{H} + \Delta - t_\alpha S < H < \hat{H} + \Delta + t_\alpha S, \quad (3)$$

where $\hat{H} = \hat{H}(N)$ is obtained evaluate of H ; N is the time series length; $\Delta = \Delta(N)$ is the calculated mean bias of the estimate, $S = S(N)$ is the calculated standard deviation; α is required significance level; t_α is the quantile of the simple normal distribution. Values Δ and S can be obtained numerically for different lengths of time series.

IV. CLASSIFICATION OF TIME SERIES BY DECISION TREE METHODS

The decision tree method is applicable to solving classification problems arising in various fields and it is considered one of the most effective. It consists in the process of dividing the original data set into groups, until homogeneous subsets are obtained. The set of rules that give such a partition allows to make a conclusion, produced based on an evaluation of some input attributes for new data.

The algorithm of learning a tree acts on the principle of recursive partitioning. Partitioning the data set is based on the use of the most suitable for this feature. In the tree a corresponding decision node is created, and the process continues recursively until the stop criterion is fulfilled.

The models that adapt their state in the learning process in accordance with the training set, such as decision trees, are unstable: even a small change in the training set may cause to significant changes in the structure of the tree. In other words, by making even minor changes in the training data, we will always receive a different model. But the original and modified models will operate approximately the same way and with comparable accuracy: minor changes in the training data will not lead to a change in the basic regularities. In this case, it is expedient to use ensembles of models. The ensemble of models as a whole can be considered as a composite model consisting of separate basic models. The components of the ensemble can be both the same type and different.

One of the first and most famous types of ensembles is the bagging method based on the statistical method of bootstrap aggregating [16]. Bootstrap aggregating is a

computer method for studying the distributions of probability distribution statistics, obtained by multiple generation of samples based on a single sample.

Bagging is a classification technique where all elementary classifiers are trained and operate independently of each other. The idea is that the classifiers do not correct each other's mistakes, but compensate them for voting. The basis of the bagging method is the classification technology, called "perturbation and combination". Perturbation is understood as the introducing of some random changes in training data and the construction of several alternative models on the modified data, followed by a result combination. From a single training set several samples containing the same number of objects are extracted by sampling. Each of the received samples is used to train one of the ensemble models. If the ensemble is built on the basis of models of various type, then each type has its own learning algorithm.

To obtain the result of the work of models ensemble the following combination methods are usually used: voting (a class that was chosen by a simple majority of ensemble models) or averaging which can be defined as the simple average of the outputs of all models (if weighted averaging is performed, then the model outputs are multiplied by the corresponding weights). The effectiveness of bagging is achieved due to the fact that the basic algorithms, trained in different subsamples, are obtained quite different and their mistakes are mutually compensated for in the voting process and also because the emission objects may not be included in some training subsamples.

Random forest is also a method of bagging, but unlike its main version, the random forest has several features: 1) it uses within itself an ensemble of only regression or classifying decision trees; 2) in the sampling algorithm, in addition to random selecting training objects random selection of features is also carried (usually the new number of features is equal to the square root of the total number); 3) for each subsample, the decision tree is built up to the full completion of the training objects and it is not subjected to post pruning. [17]

V. INPUT DATA

Numerous studies of processes in information networks have shown that network traffic has the multifractal properties. One of the frequently used models of the multifractal process is the conservative stochastic binomial multiplicative cascade [18]. In its construction, the initial time interval is divided into two equal intervals, to which the weighting factors w_1 and $1-w_1$ are assigned. The weighting coefficients are the independent values of some given random variable. If we choose a random variable defined over the interval $[0, 1]$, then the sum of the coefficients at each iteration is equal to unity. On the second iteration, two new independent random values w_2 and w_3 are added. We obtain four intervals with weight coefficients $w_1 w_2$, $w_1(1-w_2)$, $(1-w_1)w_3$ and $(1-w_1)(1-w_3)$. With the number of iterations $n \rightarrow \infty$, we arrive at the limiting measure, which is an inhomogeneous fractal set. For each iteration n , $n \gg 1$, we have a time series (cascade) of length 2^n with multifractal properties.

The fractal characteristics of stochastic multiplicative cascade obtained using beta distribution random variable $Beta(\alpha, \beta)$ are completely determined by the parameters $\alpha, \beta > 0$ [19]. Fig. 1 shows typical cascades with different Hurst exponent: the cascade with $H = 0.7$, is on the left, on the right is the cascade with values $H = 0.9$.

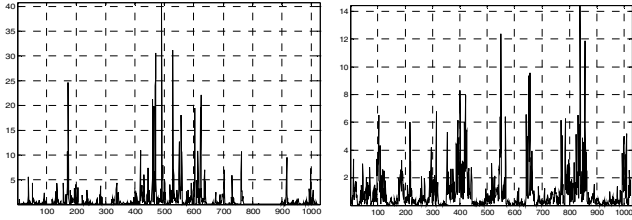


Fig. 1. Multifractal cascades: on the left $H = 0.7$, on the right $H = 0.9$

VI. EXPERIMENT AND RESULTS

Fractal time series for classification were obtained by generating stochastic binomial cascades on the basis of a symmetric beta distribution. In this case, building the cascade process allows to obtain multifractal time series with the Hurst exponent in the range $H \in (0.5, 1)$ and they can be split into several classes according to the Hurst exponent values.

Each class is a set of generated time series with a Hurst exponent, which lies in a specified range of values. To generate cascades within each class values H are chosen using a uniform distribution. To form series classes the ranges of Hurst exponent are changed in the interval $(0.5, 1)$ in increments of 0.05. The minimum and maximum values of the H were chosen at 0.51 and 0.99, respectively. Thus, the training of models was carried out on 11 classes, where $H \in \{[0.51, 0.525), [0.525, 0.575), [0.575, 0.625), \dots, [0.975, 0.99]\}$.

A preliminary comparative analysis of the classification of fractal time series by meta-algorithms using decision tree methods was carried out. The results showed that the best results are given by the method of Bagging and Random Forest which use regression trees. Therefore, to carry out the classification using the methods of Machine Learning, the Random Forest method with regression trees was selected.

In the work, three different approaches to determine the time series belonging to one of the classes were considered. In the first approach, the classification was carried out directly from the time series values by the Random Forest method, where the object was the cascade time series, and the features were the values of this series. When using decision tree regression the result is the probability of cascade series matching to a given class. The probabilities are calculated by the formula: $P_i = 1 - |m_i - C|$, where m_i is the regression result for the i -th example, C is theoretically known class number.

The second approach to the classification of multifractal cascades is also based on machine learning, but in this case, the statistical and multifractal characteristics obtained from time series were used for the classification: standard deviation, maximum value and median of series, mean and standard deviation of the generalized Hurst exponent $h(q)$,

the values $h(1)$, $h(2) = H$, and the range $\Delta h = h(0.1) - h(5)$. In this case, the objects were cascade series, and the features were the estimates of characteristics calculated for each cascade.

The third way of classification is based on direct estimation of the Hurst exponent by time series and determination of the confidence interval in accordance with (3). The choice of the class was defined as the probability of finding the true value of the Hurst exponent within the confidence interval in relation to the class width.

To build decision tree models, Python with libraries that implement machine learning methods was used. The training of models for each class was conducted on 500 examples of time series and tested on 50 test cases. In the case of using confidence intervals, the classification was carried out for the test sample. The classification was performed for time series with different lengths, but for compare the results, the main attention was paid to the series of length 512 and 4096 values.

As a classification result, the histograms of the probabilities of class determination for each range of the Hurst exponent were obtained. Fig. 2 shows the histograms obtained for the class $0.725 \leq H < 0.775$ corresponding to the classification results for cascades of length 4096 values. Such distributions are typical for all cascade classes, including for series with greater number of values.

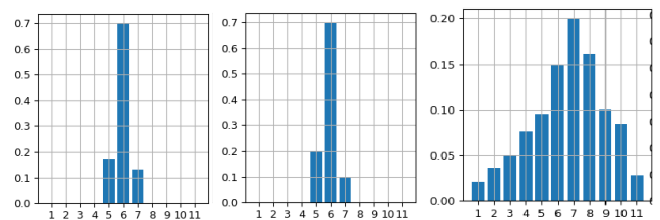


Fig. 2. Distribution of probabilities of determining the class number for different classification approaches: by time series values (left), by time series characteristics (middle), by estimation of Hurst exponent (right)

In many classification tasks, including detecting a discord or changing one of the system indicators, it is required to determine the probability of the object belonging to one of the two classes (for example, to detect that the server is being attacked). Therefore, in the work the classification of cascade series into 2 classes was carried out, for which the Hurst exponent was chosen in the ranges $0.51 \leq H < 0.7$ and $0.7 \leq H \leq 0.99$. The table presents the average probability of class determination and time of training model depending on the classification method for both cases: for 11 and 2 time series classes. Thus, the obtained results indicate a great advantage of machine learning methods over traditional methods of estimating fractal characteristics when classifying time series by fractal properties. However, for training, a sufficient number of time series with known properties is necessary, which is often problematic. Obviously, an interesting and promising task is to develop a classification methods based on the simulated time series with given fractal and probabilistic properties for using them along with real data as a training sample.

Classification performed on the basis of training by the time series values showed a slightly higher accuracy than on the basis fractal and statistical characteristics. However, it should be noted that in this case the number of features is

determined by the series length. It requires learning anew when changing the length of the time series, moreover, the time of leaning increases nonlinearly with increasing length of the time series.

TABLE I. AVERAGE PROBABILITY OF CLASS DETERMINATION

	Length	Time series values		Time series characteristics		Estimate H
		<i>P</i>	<i>time</i>	<i>P</i>	<i>time</i>	<i>P</i>
1 classes	512	0.75	15 min	0.74	1 min	0.19
	4096	0.81	120 min	0.75	1.5 min	0.23
2 classes	512	0.97	5 sec	0.96	15 sec	0.75
	4096	0.97	10 sec	0.96	25 sec	0.78

The time of learning the model by characteristics weakly depend on the series length and it is substantially smaller for large series. It is worth emphasizing that in the second case, the possibility of using the learned model does not depend on the series length: with a small length of the series, only the accuracy of the classification can suffer because of less estimation accuracy of characteristics.

VII. CONCLUSION

In this paper, a comparative analysis of the classification of multifractal stochastic time series using meta-algorithms based on decision trees has been performed. Binomial multiplicative stochastic cascades were used as input time series. Time series were split into several classes depending on their degree of self-similarity.

Three different approaches were used for the classification. In the first, the classification was carried out by the Random Forest method directly by the time series values. In the second, Random Forest was also used, but the features for classification were statistical and multifractal characteristics of the time series. In the third case, the probability of belonging to the class was determined by estimating the Hurst exponent by the time series.

The results of the classification showed the advantage of machine learning methods over the traditional method of estimating the Hurst exponent, especially with a short length of the time series.

The obtained results can be used for practical applications related to the classification or clustering of real time series with fractal properties, for example, detecting DDoS attacks.

In our future research we intend to concentrate on the study and classification of traffic realizations that includes DDoS attacks and without that ones and for various transmission protocols in order to develop a method of early intrusion detection based on the use machine learning methods.

REFERENCES

[1] G. Kaur, V. Saxena and J. Gupta, "Detection of TCP targeted high bandwidth attacks using self-similarity", Journal of King Saud University, Computer and Information Sciences, pp. 1-15, 2017. doi: 10.1016/j.jksuci.2017.05.004.

[2] R. Deka and D. Bhattacharyya, "Self-similarity based DDoS attack detection using Hurst parameter," Security and Communication Networks, vol. 9, no. 17, pp. 4468-4481, 2016. doi: https://doi.org/10.1002/sec.1639

[3] S. M. Popa and G. M. Manea, "Using Traffic Self-Similarity for Network Anomalies Detection," 20th International Conference on Control Systems and Computer Science, Bucharest, pp. 639-644, 2015. doi: 10.1109/CSCS.2015.89

[4] A. Banerjee, S. Sanyal, T. Guhathakurata, R. Sengupta and D. Ghosh, "Categorization of stringed instruments with multifractal detrended fluctuation analysis". [Online]. Available: https://arxiv.org/ftp/arxiv/papers/1601/1601.07709.pdf. [Accessed: 20-Mar-2018].

[5] Ł. Korus and M. Piórek, "Compound method of time series classification," Nonlinear Analysis: Modelling and Control, vol. 20, no. 4, pp. 545-560, 2015.

[6] A. Alghawli, and L. Kirichenko, "Multifractal Properties of Bioelectric Signals under Various Physiological States," Information Content & Processing International Journal, vol. 2, no.2, pp.138-163, 2015.

[7] A. Nepochenko, "Rhinomanometric signal processing for selection of formalized diagnostic criterion in rhinology," Telecommunications and Radio Engineering, vol. 74, no. 14, pp. 1285-1294, 2015.

[8] A. Coelho and C. Lima, "Assessing fractal dimension methods as feature extractors for EMG signal classification," Engineering Applications of Artificial Intelligence, vol. 36, pp. 81-98, 2014.

[9] S. P. Arjunan, D. K. Kumar and G. R. Naik, "A machine learning based method for classification of fractal features of forearm sEMG using Twin Support vector machines," Annual International Conference of the IEEE Engineering in Medicine and Biology, Buenos Aires, pp. 4821-4824, 2010. doi: 10.1109/IEMBS.2010.5627902

[10] H. Zhang, P. Chang-Shing, and C. Qingsheng, "An improved algorithm for feature selection using fractal dimension," 2nd International Workshop on Databases, Documents, and Information Fusion, Karlsruhe, Germany, July 4-5, 2002 pp.1-8.

[11] I. Ivanisenko, L. Kirichenko and T. Radivilova, "Investigation of multifractal properties of additive data stream," 2016 IEEE First International Conference on Data Stream Mining & Processing (DSMP), Lviv, pp. 305-308, 2016. doi: 10.1109/DSMP.2016.7583564

[12] J. Kantelhardt, S. Zschiegner, E. Koscielny-Bunde, S. Havlin, A. Bunde and H. Stanley, "Multifractal detrended fluctuation analysis of nonstationary time series," Physica A: Statistical Mechanics and its Applications, vol. 316, no. 1-4, pp. 87-114, 2002.

[13] L. Kirichenko, T. Radivilova, and I. Zinkevich, Comparative Analysis of Conversion Series Forecasting in E-commerce Tasks. In: Shakhovska N., Stepashko V. (eds) Advances in Intelligent Systems and Computing II. CSIT 2017. Advances in Intelligent Systems and Computing, vol 689. Springer, Cham, 2018.

[14] L. Kirichenko, T. Radivilova and V. Bulakh, "Generalized approach to Hurst exponent estimating by time series," Informatics Control Measurement in Economy and Environment Protection, vol. 8, no. 1, pp. 28-31, 2018.

[15] L. Kirichenko, T. Radivilova and I. Zinkevich, "Forecasting weakly correlated time series in tasks of electronic commerce," 12th International Scientific and Technical Conference on Computer Sciences and Information Technologies (CSIT), Lviv, pp. 309-312, 2017.

[16] L. Breiman, "Bagging predictors", Machine Learning, vol. 24, no. 2, pp. 123-140, 1996. Doi: 10.1023/A:1018054314350

[17] L. Breiman, "Random Forests", Machine Learning, vol. 45, no. 1, pp. 5-32, 2001. Doi: 10.1023/A:1010933404324

[18] R. H. Riedi. "Multifractal processes," in Doukhan P., Oppenheim G., Taqu M. S. (Eds.), Long Range Dependence: Theory and Applications: Birkhuser, 2002, pp.625-715.

[19] L. Kirichenko, T. Radivilova and E. Kayali, "Modeling telecommunications traffic using the stochastic multifractal cascade process," Problems of Computer Intellectualization, pp.55-63, 2012.

Identifining European E-Learner Profile by Means of Data Mining

Olga Zavgorodnia
*Computer Systems and Technologies
department*
*Simon Kuznets Kharkiv National
University of Economics*
Kharkiv, Ukraine
olga.zavgorodnia@hneu.net

Ivan Mikheev
*Department of Computer Science and
Information Technologies*
*Kharkiv National University of Civil
Engineering and Architecture*
Kharkiv, Ukraine
i.a.mikheev@gmail.com

Oleksandr Zyma
Tourism Department
*Simon Kuznets Kharkiv National
University of Economics*
Kharkiv, Ukraine
zima@hneu.edu.ua

Abstract—The urgency of using e-learning in higher education is not in doubt at this time. The trend of using e-learning in the current practice of educational institutions and business structures of various levels has become global, due not only to the peculiarities of the development of educational systems in the world, but also to economic factors, which in many cases are the main driver of changes and innovations in the educational sphere. In spite of all the difficulties of deployment, publishing of those systems and troubles of their efficiency maintenance, their usage is spreading in tertiary education as well as in vocational education. In order to gain profits from e-learning materials Ukrainian universities search the possibilities to enter European e-learning market. For this reason it is vital to understand European e-learner, his needs and features. Therefore the European e-learner profile identification was performed on the basis of open European statistical datasets by the means of developed data conversion mechanism.

Keywords—*e-learner profile, data mining, statistical databases, data conversion mechanism*

I. INTRODUCTION

The development of information-communication technologies (ICT) and their broad incorporation in all spheres of social and business life influences tremendously on learners of all kinds. The deepening of integration processes in the context of informatization of economic activity creates information-rich environments that makes a necessarily for the personnel to be trained to perform the specified ICT tasks (those that require ICT skills). The need for developed countries' economies in ICT skills is due to the general introduction and profound restructuring of the economy through the use of ICTs: the Internet in particular and ICT as a whole change the nature of work performed by employees, reduce the amount of routine manual and mental labor [1], form new requirements for personnel [2]. Furthermore, e-learning is highly effective in case of resemblance of approaches of work and learning, e.g. in work based on ICT usage [3, p. 59].

The intrusion of ICT technologies into educational process of all levels is status quo nowadays. The educational systems of the world embrace ICT in educational process not only because innovation should be an essential part of up-to-date education, but because of impossibility of not using ICT – it became a demand of most stakeholders of education. And there are many economic drivers that stimulate the ICT inclusion in educational system.

The results of search queries on "e-learning" in Google gives a possibility to state its high popularity – about 29.4 million results. As some research state [4], those results prove to be important.

According to the e-learning market research held in 2016 [5], in 2015 the size of the market was estimated to be over \$165 billion. The expected growth of the global e-learning market is expected to be by 5% between 2016 and 2023, so the expected capacity of the global market will be \$240 billion in 2023. Such growth rates and market volumes may seem dubious at first glance, but the availability of other comparative studies significantly reduces the level of skepticism. Thus, according to [6], the capacity of the global e-learning market in 2015 was \$165,21 billion, and is expected to reach \$275.10 billion by 2022 growing by 7,5% during the forecast period. Forecasts of the growth of the global e-learning market are not uniform and unambiguous. According to another forecast [7], the growth of the global e-learning market in 2015–2020 should occur at a less rapid pace. Such a forecast is more cautious about the prospects of developing regions of the world in electronic learning.

However, the reports combine an understanding of the trends in the significant growth of e-learning educational markets around the world. Moreover, in the forecasts, significant pace of development is expected from Africa, Eastern Europe and the Russian Federation in electronic learning implementation.

Brazil, Mexico and Argentina are currently leaders in the implementation of eLearning in Latin America, expecting significant growth in the Dominican Republic (36%), Honduras (34%) and El Salvador (32%) [7]. Expectations about the pace of growth of the e-learning market in Asia are uneven; India, China, and Australia are expected to be the drivers of the region's development (according to the generally accepted division for parts of the world, Australia is not part of Asia, but the authors of studies [5] and [7] have included it in this region).

The actual growth of the e-learning market by 2015 in the regions of the world indicates the presence of significant opportunities and potential of the market, including Eastern Europe. Indeed, Eastern European markets capacity in monetary terms are low in comparison to the e-learning markets of North America, Asia, and Western Europe, but the growth prospects remain extremely attractive to all educational institutions in the world.

The growth rate of the global e-learning market (and its attractive capacity in cash equivalents) retains increased interest in higher education institutions prior to the introduction of e-learning systems and deepens the desire to participate in the distribution of such market revenues.

According to studies conducted by the European Association of Universities [8, 9], the vast majority of European universities have already implemented e-learning in the current educational process or in the process of such implementation. Exact figures for the coverage of the educational process of higher education institutions by e-learning systems are not available in official statistical sources, so we will rely on the results of sample surveys. Thus, a study under the auspices of the European Association of Universities [8] in 2014 covered 249 institutions of higher education in Western and Eastern Europe (36 countries), it is one third of all higher education institutions in Europe that are part of the European University Association, and study [9] held in 2015, covering 451 higher education institutions in Europe and representing 10 million students (46 countries), which is roughly a quarter of all students of European higher education institutions. According to [8], e-learning has been implemented by almost all institutions of higher education or in the process of such implementation. Thus, the vast majority of institutions surveyed (91%) have introduced blended-learning systems, which involve integrating e-learning into a traditional system of teaching disciplines. However, 82% also noted the availability of teaching disciplines only through e-learning tools, i.e. the full use of e-learning distance learning.

Consequently, it can be considered proven that almost all higher education institutions in Europe are involved in e-learning, but the degree of coverage of curricula by e-learning systems is not uniform. Only half of the institutions indicated that e-learning was implemented in the whole institution [8, 9], while only one third of institutions involved all or most students in e-learning. Non-continuous, partial use of e-learning systems may be due to the attempts of careful implementation in order to find financially efficient and robust models [8, p. 8] to introduce e-learning to all disciplines in the short term.

Also, we use the data from the elective study of e-learning under the auspices of EDUCAUSE [10] held in 2013, which covered 311 higher education institutions (EDUCAUSE members). According to a study [10], more than 80% of institutions of higher education have at least several disciplines using e-learning systems, while more than half offer a significant number of electronic distance courses (disciplines).

A separate direction in electronic learning is MOOCs (Massive Open Online Courses) that as a phenomenon appeared on the educational background only in 2008, however, the first discipline in this format was introduced in 2012. Thus, according to [11, p. 14], the leader in the implementation of the MOOCs is North America, which is followed by Europe with a significant margin. According to a study by the European Association of Universities [8], 12% of the polled institutions of higher education offer MOOCs, 42% plan to create such courses, and 39% of respondents do not plan to implement the MOOCs in the near future. In addition, this study did not reveal correlations between the implementation of other forms of e-learning and the implementation of the MOOCs.

Consequently, the introduction of e-learning in a variety of forms into the world of educational space is beyond doubt. Under this light the task of entering of Ukrainian educational institutions on this market is highly promising. But in order to enter this market a multifacet research should be held to understand the common e-learner profile. That might give an opportunity to customize e-learning courses to the e-learners' needs.

II. DATA SOURCES ABOUT E-LEARNERS

The gaining popularity of e-learning and blended learning technologically enriched education lead the boom of scientific research on various effects of such educational forms using and resulting problems. The numerous research in specialized scientific journals are printed, i.e. Creative education, Electronic journal of e-learning, European journal of open and distance and e-learning, Interactive multimedia electronic journal of computer-enhanced learning, International journal of emerging technologies in learning, International journal of advanced corporate learning, International review of research in open and distributed learning, Journal of online learning and teaching, The American journal of distance education, etc. (mentioned in alphabetical order). As an example there might be mentioned the papers on e-learners' profile building [12–14].

The research held under the auspices of corresponding associations and organizations are also the source of needed data. They are: Accreditation Organisation of the Netherlands and Flanders (NVAO), Commonwealth of Learning (COL), EDUCAUSE, European University Association, European University quality in eLearning (UNIQUE), United Nations Educational, Scientific and Cultural Organization (UNESCO), etc. (mentioned in alphabetical order). Their reports contain a great variety of needed data with the only (but most essential) pitfall – usually they show some snaps of current state but they lack data for dynamic analysis due to gaps or incomparable data [8–11].

All those sources of data about e-learners are surely important though they are not enough for multifacet research of the market, because they lack some basic information of e-learners wide picture. That's why it is important to start the analysis from open statistical databases available on-line. They are European Statistical Committee (Eurostat) statistical database [15] and Organization of Economic Cooperation and Development (OECD) database [16] which provide systematically gathered data on e-learning and e-learners.

Those data may not have all the needed details, but they are provided by all the member countries in a wide time period (from 10 to 15 years). All that makes it essential to use those databases, but the problem is a big dataset with lack of build-in tools for it analysis. All that highlights the problem of big dataset analysis with the means of analytical tools (in contrast to manual search).

III. MATHEMATICAL BASE

The mathematical justification of the European e-learner profiles' determination is based on the use of Bayes' theorem [17]. Let's introduce some notation:

V_i - i -th characteristic (variable) of the system under study.

The authors considered 8 characteristics of the system, 7 of which are basic with indices $i = 1..7$, an additional characteristic with the 8th index - for the study of anomalies in the system.

- V_1 - gender;
- V_2 - education;
- V_3 - age;
- V_4 - social status;
- V_5 - level of prosperity;
- V_6 - e-learning activities;
- V_7 - internet-use;
- V_8 - region.

Each of the characteristics is a discrete random variable, for each of which the set of all possible values is defined - D_i :

$$D_i = \{V_{i1}, \dots, V_{in}\}$$

Where V_{i1}, \dots, V_{in} - the set of all possible values of a variable V_i .

Let $H_{V_{ij}}$ denote the hypothesis of including the j -th value of the i -th characteristic of the system in the resulting portrait of the user of e-learning systems. The search result is a hypothesis of the form –

$$H_{V_{1j}} | H_{V_{2j}} | H_{V_{3j}} | H_{V_{4j}} | H_{V_{5j}} | H_{V_{6j}} | H_{V_{7j}}$$

The powerset of hypotheses can be determined by the formula:

$$W = |D_1 \times \dots \times D_i|$$

The Bayes formula was used to determine the desired portrait.

$$\max(P(H_{V_{1j}} | H_{V_{2j}} | H_{V_{3j}} | H_{V_{4j}} | H_{V_{5j}} | H_{V_{6j}} | H_{V_{7j}}))$$

Determination of the maximum probability value gives an exact solution in the choice of the hypothesis, and accordingly in the determination of the values of the characteristics included in the system.

Anomaly was investigated with the 8th characteristic included in the system.

IV. DATA MINING FOR E-LEARNER PROFILE IDENTIFICATION

A. Data mining for e-learners analysis

The list of used systems and technologies was determined by the need for a functional and friendly tool for conducting analytical research on data taking into account the complexity of the task [18]. They should be described at two levels: conceptual approaches and technical solutions.

At the conceptual level there were chosen: data mining, OLAP and ROLAP.

Data mining is known as the process of "extracting valid, unknown and wide ranging data from a data-warehouse, making it possible to organize decision-making processes" [19]. Data mining can be defined in several ways, which differ primarily in their focus on different aspects of data mining. One of the earliest definitions is the following: "the non-trivial extraction of implicit, previously unknown, and potentially useful information from data" [20].

The idea of data-mining usage in order to provide learner analytics is not entirely new [21–24], though they are primarily are focused on behavior analysis of e-learners while using some specific learning management system. But in our case we are going to use data mining in order to retrieve meaningful data about European e-learners from big dataset of open statistical databases.

OLAP is a technology of multidimensional tables that can be accessed for user requests. The second name for multidimensional tables is multidimensional cubes (hypercubes). The tables are based on raw and aggregated big-sized data.

ROLAP as a consequence of the previous one – the original data are stored in relational databases or in 2-dimension local tables on the file server. Aggregate data can be placed in service (temporary) tables of the same database. Data transformation from a relational database to multidimensional cubes are performed by the request of the OLAP tool (fig. 1).

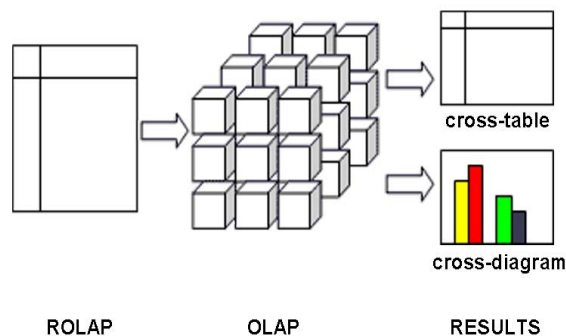


Fig. 1. Data conversion mechanism.

At the technical level there were chosen:

MySQL (a relational database for storing data and processing SQL queries),

PHP (scripting programming language for the server part of the application),

HTML (markup language for presentation of results).

The criteria of open source, high performance and universality have determined the use of these development kit.

B. The developed mechanism

The process of achieving the goal can be divided into several stages:

- 1) Select of the data source (Eurostat and OECD).
- 2) Data preparation (export of 26 tables with data in CSV format).

3) Data import into MySQL database (additional manipulations with data to ensure the process of correct import were held).

4) Design SQL queries (creation of crosstab queries).

5) Design a user interface.

6) Development of specialized classes and modules for working with data.

7) Merge scripts to process analyst queries into the database and data representation mechanisms.

The complexity of the implementation is caused by such factors as the disparity of data structures, large amounts of information (more than $0.25 \cdot 10^6$ records in tables) and, as a consequence, the need for a large amount of disk space for their storage and RAM for processing queries (the problem is solved by reconfiguring the MySQL server and related programs), the lack of PIVOT queries in MySQL for building a slice of OLAP hyper cubes (the problem is solved using the technology of subqueries, aggregate functions and conditional statements inside the body of the query) (fig. 2). The obtained data, although stored in relational (flat) tables, however represented a hypercube of data of 6-dimensional measurement.

```
SELECT
    'IND_TYPE',
    round(avg(IF('TIME'="2007", `Value`, NULL)),2) AS "2007",
    round(avg(IF('TIME'="2008", `Value`, NULL)),2) AS "2008",
    round(avg(IF('TIME'="2009", `Value`, NULL)),2) AS "2009",
    round(avg(IF('TIME'="2010", `Value`, NULL)),2) AS "2010",
    round(avg(IF('TIME'="2011", `Value`, NULL)),2) AS "2011",
    round(avg(IF('TIME'="2012", `Value`, NULL)),2) AS "2012",
    round(avg(IF('TIME'="2013", `Value`, NULL)),2) AS "2013",
    round(avg(IF('TIME'="2014", `Value`, NULL)),2) AS "2014",
    round(avg(IF('TIME'="2015", `Value`, NULL)),2) AS "2015",
    round(avg(IF('TIME'="2016", `Value`, NULL)),2) AS "2016",
    round(avg(IF('TIME'="2017", `Value`, NULL)),2) AS "2017"
FROM
    `bde15cua-y16_75`
GROUP BY
    'IND_TYPE'
ORDER BY
    'IND_TYPE'
```

Fig. 2. Example of cross-table result.

As a rule the header of rows was the field "TIME" (the year period) and the header of columns the field "IND_TYPE" (the subtypes of observed facts). The value of cells was the field "VALUE" (the value of the indicator – in all cases it was the percentage of individuals), the group operation mostly was the arithmetic average. An example of a query is shown in fig. 2. The part of a query with text like "round(avg(IF('TIME'="YYYY", `Value`, NULL)),2) AS "YYYY"" was generated automatically by PHP script.

The result is a flat (summary) table (fig. 3). The user is allowed to select the hypercube axes (the columns and rows of the resulting table) independently, and one of the 5 group operations can be applied to the values of the data cells.

INDIC_IS	2010	2011	2012	2013	2014	2015	2016	2017
Individuals, 15 years old or less	0.80	0.71	0.83	0.25	0.33	0.00	1.00	1.00
Individuals, 16 to 24 years old	3.46	3.70	4.22	5.17	6.08	5.36	5.80	7.08
Individuals, 25 to 34 years old	3.03	4.03	4.43	5.03	5.70	5.14	5.43	5.89
Individuals, 35 to 44 years old	2.80	3.09	4.11	4.39	5.00	3.83	4.23	5.16
Individuals, 45 to 54 years old	1.60	2.12	2.65	2.64	3.62	2.67	3.03	3.63
Individuals, 55 to 74 years old	0.69	0.82	0.89	1.03	1.32	0.97	1.09	1.50
Individuals, 75 years old or more	0.00	0.00	0.14	0.00	0.00	0.00	0.00	0.25

Fig. 3. Example of cross-table result.

From the cross table (fig. 3) it is evident that the age profile of European e-learner is mostly unchangeable: the most active e-learners are 16–34 old (or 16–44 years old as a wider margin). Also e-learning is gaining more popularity among these groups from year to year.

V. IMPLEMENTATION OF THE DEVELOPED DATA-MINING ALGORITHM FOR E-LEARNER'S PROFILE IDENTIFICATION

In order to make basic identification of European e-learner there were extracted and prepared the set of corresponding datasets from statistical databases of Eurostat and OECD. They contained the data as follows: set of specific e-learner activities (looking for information about education, training and courses offers; working with online learning material, doing an online course; communicating with instructor and/or students using educational websites or portals), their age (multiple division on age, i.e. see fig. 3), sex, education (low, middle, high formal level), status (active labor force, student, manual worker, etc.), income (four quartiles) by country and some mixture of the mentioned above (fig. 4). The 10 years range was taken for the analysis (2007–2017). Also some extra activities were taken for gaps exclusion. The implementation of developed data mining tools gave the opportunity to identify the following specific features of European e-learners:

according to the education attainment the e-learners mainly are with high formal education (no matter of gender), but in a group of individuals with high formal education the females rates are lower than males;

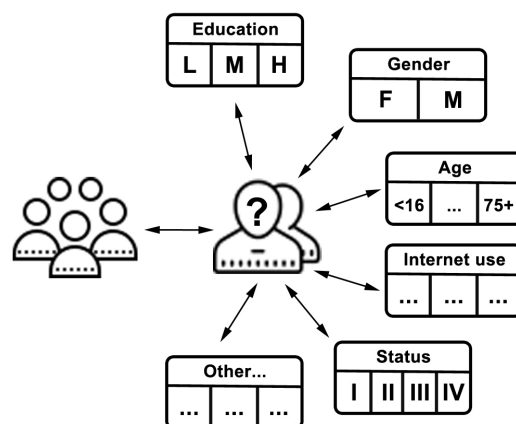


Fig. 4. The data structure in statistical databases.

the exceptionally high level of e-learning activities are in Iceland – 71% of average among all individuals, 79% among

males with high formal education, 83% of females with high formal education (all data are of 2017 year);

for some of Scandinavian countries (Denmark, Finland, Norway, Sweden) the level of e-learning activity doesn't depend on education attainment: the difference between groups of education attainment is present though much lower than in average in European Union;

the differentiation of e-learning activity by country seems to be based on average income per individual and density of population, though some exceptions (like Estonia and Iceland) prove the need of future in depth analysis;

the ICT professionals were considerably more active in e-learning activities of all kinds in all the countries (i.e. in Portugal in 2017 the average rate of e-learning activity was 21% among all the individuals, 69% among ICT professionals, 22% among non ICT professionals);

according to status the most active part of population are (42% of European students take part in e-learning activities); it is proved also by age groups analysis – the most active are individuals 16–19 years old and 20–24 years old (35% and 31% in 2017 respectively);

the differentiation of e-learning activities by income group are not evident and further analysis is needed (it seems that individuals of second and third quartiles are more active than others, though the pattern is not supported by all countries);

European e-learners are mostly mobile internet users.

VI. CONCLUSIONS

The detailed analysis of European e-learners was held. The data sources for it were open statistical databases of Eurostat and OECD. In order to process those data on the basis of data mining algorithm the data conversion mechanism was developed and implemented, so that it gave the opportunity to retrieve the following hidden information from data: in order to meet European e-learners' needs Ukrainian universities should propose the e-learning products to younger audience (individuals aged 16–24) on mobile e-learning platforms (or with build-in opportunities for mobile devices). Those courses should be oriented on ICT professionals, so they must be up-to-date and technologically rich. The developed big data conversion mechanism gives an opportunity to get the cross selection of OLAP hypercube. It might be implemented to analyze data from any sphere of application (not only e-learning). The mechanism retrieves the information to the analyst according to the chosen group operations and settings. The further research is needed in identifying the preferable countries and income groups for proposing of Ukrainian e-learning products though there more hidden factors should be taken into account and more data sources taken.

REFERENCES

- [1] F. Levy, "How technology changes demands for human skills," OECD Education working papers, no. 45, OECD publishing, Paris, 2010. <http://dx.doi.org/10.1787/5kmhds6c zqzq-en>.
- [2] S. Kavun, I. Sorbat and V. Kalashnikov, "Enterprise Insider Detection as an Integer Programming Problem," In: J. Watada, G. Phillips-Wren, L.C. Jain and R.J. Howlett (eds.), *Advances in Intelligent Decision Technologies*, SpringerVerlag Series "Smart Innovation, Systems and Technologies", vol. 12, 2012, pp. 820-829.
- [3] J. P. Batalla-Busquets and C. Pacheco-Bernal, "On-the-job e-learning: worker's attitudes and perceptions," *The International Review of Research in Open and Distance Learning*, vol. 14, no. 1., pp. 40–64, 2013. <http://www.irrodl.org/index.php/irrodl/article/viewFile/1304/2444>.
- [4] S. Kavun, I. Mykhalchuk, N. Kalashnykova, and A. Zyma, "A Method of Internet-Analysis by the Tools of Graph Theory," In: Watada, J., Phillips-Wren, G., Jain, L.C., and Howlett, R.J. (Eds.), *Advances in Intelligent Decision Technologies*, SpringerVerlag Series "Smart Innovation, Systems and Technologies", vol. 15, part 1, 2012, pp. 35-44.
- [5] *E-Learning market trends & forecast 2017 – 2021 report*, A report by Docebo, 2016. London : Docebo S.p.A, 2016.
- [6] *Report and Forecast to 2017-2022*, A report by Orbis Research, Dallas, Texas, June 15th, 2017. <http://www.orbisresearch.com/contacts/request-sample/226355>.
- [7] *Role of e-learning in Higher Education in Latin America / Iner-American Dialogue*, April 12th, 2016. <http://www.thedialogue.org/2016/04/role-of-e-learning-in-higher-education-in-latin-america>.
- [8] M. Gaebel, V. Kupriyanova, R. Morais and E. Colucci, *E-Learning in European Higher Education Institutions. Results of a mapping survey conducted October–December 2013*. European University Association. Brussels : EUA, 2014.
- [9] *A. Sursock Trends 2015: Learning and Teaching in European Universities*. European University Association. Brussels: EUA, 2015.
- [10] J. Bichel *The state of e-learning in higher education: An eye toward growth and increased access (research report)*. CO: EDUCAUSE Center for Analysis and Research. Louisville: EDUCAUSE, 2013.
- [11] *NVAO MOOCs and Online HE. A survey conducted June 2014*, Accreditation Organisation of the Netherlands and Flanders (NVAO). The Hague: NVAO, 2014.
- [12] I. El Haddioui and M. Khaldi, "Learner Behavior Analysis on an Online Learning Platform," *International Journal Of Emerging Technologies In Learning (IJET)*, vol. 7(2), pp. 22–25, 2012. doi:<http://dx.doi.org/10.3991/ijet.v7i2.1932>.
- [13] Kun Liang, Yiyin Zhang, Yeshen He, Yilin Zhou, Wei Tan, and Xiaoxia Li, "Online Behavior Analysis-Based Student Profile for Intelligent E-Learning," *Journal of Electrical and Computer Engineering*, vol. 2017, Article ID 9720396, 7 pages, 2017. doi:[10.1155/2017/9720396](http://dx.doi.org/10.1155/2017/9720396).
- [14] S. A. Hosseinil, A. H. Tawil, H. Jahankhani and M. Yarandi, "Towards an ontological learners' modelling approach for personalised e-learning," *International Journal Of Emerging Technologies In Learning (IJET)*, vol. 8(2), pp. 4–10, 2013. doi:<http://dx.doi.org/10.3991/ijet.v8i2.2476>.
- [15] Eurostat database, <http://ec.europa.eu/eurostat/data/database>.
- [16] OECD database, <http://stats.oecd.org>.
- [17] D. Kahneman, P. Slovic, A. Tversky (eds.), *Judgment under Uncertainty: Heuristics and Biases*. Cambridge University Press, 1982.
- [18] A. Zamula, and S. Kavun, "Complex systems modeling with intelligent control elements," *International Journal of Modeling, Simulation, and Scientific Computing*, vol. 8, no. 1, 2017. doi:<http://dx.doi.org/10.1142/S179396231750009X>.
- [19] O. Pushkar, and O. Zavgorodnia, *Decision support systems*. Kharkiv: Publishing House of KhNUE, 2012.
- [20] R. Niesbert, J. Elder, G. Miner, *Handbook of statistical analysis and data mining applications*. London: Elsevier Inc., 2009.
- [21] D. Monk, "Using Data Mining for e-Learning Decision Making," *The Electronic Journal of e-Learning*, vol. 3 iss. 1, pp 41–54, 2005.
- [22] S. Suhirman, J. M. Zain, H. Chiroma and T. Herawan, "Data Mining for Education Decision Support: A Review," *International Journal Of Emerging Technologies In Learning (IJET)*, vol. 9(6), pp. 4–19, 2014. doi:<http://dx.doi.org/10.3991/ijet.v9i6.3950>.
- [23] R. R. Zhou, "Education Web Information Retrieval and Classification with Big Data Analysis," *Creative Education*, vol. 7, pp. 2868-2875, 2016. <http://dx.doi.org/10.4236/ce.2016.718265>.
- [24] S. Kavun, Y. Daradkeh and A. Zyma, "Safety Aspects in the Distance Learning Systems," *Creative Education*, vol. 3, pp. 84–91, 2012. doi:[10.4236/ce.2012.31014](http://dx.doi.org/10.4236/ce.2012.31014).

High-Performance Data Stream Mining by Means of Embedding Hidden Markov Model into Reproducing Kernel Hilbert Spaces

Galyna Kriukova
Faculty of Computer Sciences
National University of Kyiv-Mohyla Academy
Kyiv, Ukraine
kriukovagv@ukma.edu.ua

Mykola Glybovets
Faculty of Computer Sciences
National University of Kyiv-Mohyla Academy
Kyiv, Ukraine
glib@ukma.kiev.ua

Abstract—Hidden Markov models (HMMs) are a well-known probabilistic graphical model for time series of discrete, partially observable stochastic processes. We consider method to extend the application of HMMs to non-Gaussian continuous distributions by embedding the belief about the state into a reproducing kernel Hilbert space (RKHS). Corresponding regularization techniques are proposed to reduce tendency to overfitting and computational complexity of algorithm, specifically, Nyström subsampling for feature and kernel matrices and general regularization family. This method may be applied to various statistical inference and learning problems, including classification, clustering, prediction, identification, and as an online algorithm it may be used for dynamic data mining and data stream mining. We investigate, both theoretically and empirically, regularization and approximation bounds. Furthermore, we discuss applications of the method to real-world problems, comparing the approach to several state-of-the-art algorithms.

Index Terms—hidden Markov model, data stream mining, reproducing kernel Hilbert space, online algorithm, regularization

I. INTRODUCTION

Development of proper models for time series of stochastic semi-observable processes is crucial for solving a wide variety of problems in learning theory. Most of the observed data from system does not depict the true states but rather noisy variates of them. Moreover, the observed state space is generally only a subset of the true state space, as the sensory equipment of most systems is limited.

Hidden Markov models (HMM) are applied to various learning problems, including prediction, classification, clustering, identification, segmentation, reinforcement learning, pattern recognition, time series change point detection, and as an online algorithms they are widely used for dynamic data stream mining [1], [2]. Basic assumption for HMM is that to obtain the current hidden state we need only a fixed number of preceding hidden states (Markovian property for transition model), and an observation depends conditionally on its corresponding hidden state (observation model). Accordingly, HMM has a bunch of disadvantages, among which

Galyna Kriukova thanks the International Charitable Foundation for Renaissance of the Kyiv-Mohyla Academy for financial support of her research

large number of unstructured parameters, limitations caused by Markov property for first order HMMs, and the most critical is that only a small portion of distributions may be represented by HMM due to the assumption of discrete number of hidden states.

A well established concept that extends the ideas of HMMs to continuous domains is the Kalman filter (KF), which assumes linear system dynamics and represents the state as a Gaussian random variable. Considering of non-linear system dynamics by means of its sequential linearization leads to Extended Kalman filter (EKF), notwithstanding, assuming zero mean multivariate Gaussian noises for transition and observation models. As further step to address complex problems with non-linear models and non-Gaussian noise, the particle filter has been proposed. Particle filter is a technique for implementing recursive Bayesian filter by Monte Carlo sampling representing the posterior density by a set of random particles with associated weights, thereby, estimates are computed based on these samples and weights. Despite of its undoubted ability to represent arbitrary densities and deal with non-Gaussian noise, there is a list of disadvantages of particle filter, among which high computational complexity, difficulties while determining optimal number of particles, number of particles increase with increasing model dimension, a vital role of proper importance density choice, and necessity of resampling to avoid potential risk of degeneracy and loss of diversity. Various modifications of particle filter have been proposed, nevertheless, there is still research challenge to develop optimal algorithm with reduced complexity.

In our study we consider a nonparametric HMM that extends traditional HMMs to structured and non-Gaussian continuous distributions by means of embedding HMM into Reproducing Kernel Hilbert Space (RKHS). Much recent progress has been made for Hilbert space embedding in probabilistic distributions and their application to HMM [3], [4], [5], [6], [7]. Due to interference and ill-posedness of the inverse problem arising at learning of embedded HMM into RKHS, regularization is required. Proposed training algorithms [3], [6], [4] use L_1 , L_2 and truncated spectral regularization

to invert the corresponding kernel matrix. In our research, we consider more general regularization techniques [8], specifically Nyström-type subsampling [9]. Moreover, simultaneous regularization by means of Nyström-type subsampling and improved optimization technique enable us to use this approach for online algorithms.

This paper is organized as follows. In Section II we develop and study the basic structure and theoretical background of the method. Section III describes the experimental framework used to evaluate the performance, pros and cons of the method.

II. EMBEDDING HMM INTO RKHS

In the standard type of HMM, there is a hidden random variable $x(t)$, $x(t) \in \{x_1^t, x_2^t, \dots, x_n^t\}$ and random variable $y(t)$ of the corresponding observation at time t . HMM is defined by emission probabilities $P(y(t)|x(t))$, transition probabilities $P(x(t)|x(t-1))$ and initial state probabilities. To train HMM generally Viterbi training or expectation-maximization (EM) algorithm are used. A priori assumptions on distribution model (i.e. Gaussian mixture) lead to narrowing of suite of considered probability densities. Employment of RKHS embedding for probability distributions allows the generalization of machine learning methods to arbitrary probability densities, not only Gaussian ones, by providing a uniform representation of functions and, consequently, probability densities as elements of a RKHS.

Here we briefly remind method for RKHS embedding for distributions and HMM described in [5], [10], [11].

Definition 1: RKHS is a Hilbert space of functions $f : \Omega \rightarrow \mathbb{R}$ with a scalar product $\langle \cdot, \cdot \rangle$ that is implicitly defined by Mercer kernel function $k : \Omega \times \Omega \rightarrow \mathbb{R}$ as $\langle \varphi(x), \varphi(y) \rangle = k(x, y)$, where $\varphi(x)$ is a feature mapping into space corresponding to the kernel function. According to reproducing property $\forall x \in \mathcal{X}, \forall f \in \mathcal{H} \langle f, k(\cdot, x) \rangle_{\mathcal{H}} = f(x)$ we have for any element f from RKHS $f(y) = \sum_{i \in I} \alpha_i k(x_i, y)$, $\alpha_i \in \mathbb{R}$.

Kernel functions have been thoroughly explored since initiative paper [12], and they have been defined on various structured objects, such as strings and graphs, although standard Gaussian radial basis function kernel is widely used as well.

Joint and conditional distributions may be embedded into a RKHS and manipulate the probability densities, by means of the chain, sum and Bayes' rule, entirely in Hilbert space.

Remark 1: Given a set of feature mappings $\Phi = [\varphi(x_1), \dots, \varphi(x_m)]$ any distribution $q(x)$ may be embedded as a linear combination $\hat{\mu}_q = \Phi \beta$, with weight vector $\beta \in \mathbb{R}^m$. The mean embedding of a distribution can be used to evaluate expectation of any function f in the RKHS, e.g. if $f = \Phi \alpha$, then

$$\mathbb{E}_q[f(x)] = \langle \hat{\mu}_q, f \rangle = \langle \Phi \beta, \Phi \alpha \rangle = \beta^\top \Phi^\top \Phi \alpha = \beta^\top K \alpha,$$

where $K = \Phi^\top \Phi$ is Gramian matrix, $K_{ij} = k(x_i, x_j)$.

Theorem 1 ([5]): Assume $k(x, x')$ is bounded. With probability $1 - \delta$

$$\|\hat{\mu}_q - \mu_q\|_{\mathcal{H}} = O\left(m^{-1/2} \sqrt{-\log \delta}\right).$$

Now we are ready to consider RKHS embedding for HMM.

Definition 2: Assuming RKHS \mathcal{F} with kernel $k(x, x') = \langle \varphi(x), \varphi(x') \rangle_{\mathcal{F}}$ defined on the observations, and RKHS \mathcal{G} with kernel $l(h, h') = \langle \phi(h), \phi(h') \rangle_{\mathcal{G}}$ defined on the hidden states, *observable operator* $\mathcal{A}_x : \mathcal{G} \rightarrow \mathcal{G}$ is defined as

$$\mathcal{A}_x \phi(h_t) = p(X_t = x | h_t) \mathbb{E}_{H_{t+1} | h_t} [\phi(H_{t+1})].$$

The *observation operator* is defined as a conditional operator $\mathcal{C}_{X_{t+1} | H_{t+1}} = \mathcal{C}_{X_t | H_t}$ that maps distribution function over hidden states embedded into \mathcal{G} to a distribution function over emissions embedded in \mathcal{F} .

Straightforward from Theorem 1 we have

Corollary 1: Assume $k(x, x')$ and $l(x, x')$ are bounded. Then with probability $1 - \delta$

$$\|\hat{\mathcal{C}}_{XY} - \mathcal{C}_{XY}\|_{\mathcal{F} \otimes \mathcal{G}} = O\left(m^{-1/2} \sqrt{-\log \delta}\right).$$

For conditional embedding operator use of regularization is needed. Thus, for Tikhonov regularization and given regularization parameter λ we have

Corollary 2: Assume $k(x, x')$ and $l(x, x')$ are bounded. Then with probability $1 - \delta$

$$\|\hat{\mu}_{Y|x} - \mu_{Y|x}\|_{\mathcal{G}} = O\left(\sqrt{\lambda} + \sqrt{\frac{-\log \delta}{\lambda m}}\right).$$

Appropriate value of regularization parameter λ may be selected by means of classical approaches, such as Morozov's discrepancy principle, or using Linear Functional Strategy considered in [13]. Moreover, other regularization techniques may be successfully applied for regularization, such as Nyström subsampling [9] or regularization family $\{g_\lambda\}$ [14].

Definition 3 ([14], Definition 2.2): A family $\{g_\lambda\}$ is called a regularization on $[0, a]$, if there are constants $\gamma_{-1}, \gamma_{-1/2}, \gamma_0$ for which

$$\begin{aligned} \sup_{0 < t \leq a} |1 - t g_\lambda(t)| &\leq \gamma_0, \\ \sup_{0 < t \leq a} |g_\lambda(t)| &\leq \frac{\gamma_{-1}}{\lambda}, \\ \sup_{0 < t \leq a} \sqrt{t} |g_\lambda(t)| &\leq \frac{\gamma_{-1/2}}{\sqrt{\lambda}}. \end{aligned}$$

It is clear that by taking $g_\lambda(t) = (t + \lambda)^{-1}$ we get widely used regularized matrix inversion. Note, that $g_\lambda(t) = \frac{1}{t}$ for $t \geq \lambda$, and 0 otherwise, corresponds to the regularization by means of spectral cut-off scheme. For details on regularization families and corresponding approximation bounds we refer to [14].

Nyström subsampling is a learning scheme applied in RKHS setting for matrix inversion, where the kernel matrix is replaced with a smaller matrix obtained by column subsampling [15], [16]. Note, that arbitrary regularization family may be applied after Nyström subsampling.

This approximation-preserving reduction allows us to extend Corollary 2 to general regularization scheme and gives us an estimation of the emission probability distribution with an approximation error of order $O(\lambda^{1/2} + (\lambda m)^{-1/2})$.

In order to evaluate transition probability distribution, we use Algorithm 1 and Theorem 1 from [5] extended for general regularization scheme, that gives us bound for

$$\|\mu_{X(t+1)|\{X(1),X(2),\dots,X(t)\}} - \hat{\mu}_{X(t+1)|\{X(1),X(2),\dots,X(t)\}}\|_{\mathcal{F}}$$

of order

$$O(t(\lambda^{1/2} + (\lambda m)^{-1/2})).$$

In order to reduce the computational complexity, for each of kernel matrices used in Algorithm 1 [5] we apply Nyström subsampling, considering instead of $m \times m$ matrix $m' \times m$ for $m' \ll m$, that reduces kernel matrix construction complexity from quadratic to subquadratic preserving approximation bounds. Note also, that embedding into RKHS reduces training and application of HMM to linear operations for kernel and feature matrices for fixed sampling basis, which consequently reduces computational complexity.

Note, that for some regularization methods, such as Tikhonov regularization, Moore-Penrose pseudo-inverse is used. Combining this approach with Nyström subsampling allows to adopt unlabeled samples to kernel matrix construction, enabling semi-supervised learning for suitable kernel.

III. APPLICATIONS

Need for online denoising and data stream segmentation occurs in various real-life problems. For various health-care problems it becomes vital, i.e. for nocturnal hypoglycemia prevention for diabetes patients either wearing continuous glucose monitoring devices, or self-monitoring glucometers [17], [18]. In [11] applications of embedded HMM in RKHS to robot vision, slot car inertial measurement and audio event classification were shown as exceeding previous state-of-the-art solutions, including ordinary HMM as well. We conducted sets of experiments to evaluate the effectiveness of learning embedded HMMs into RKHS for real-world prediction and filtering tasks.

A. Map Matching

Widely used applications such as traffic sensing, routing time prediction and recommendations require reliable online localization algorithm. Due to various errors of sensors, imprecise measurements and imperfect maps, state-of-the-art online map-matching algorithms employ HMMs [19]. Most existing approaches use Viterbi algorithm, using various sliding windows to improve performance. In HMM-based map-matching algorithms, candidate paths are sequentially generated and evaluated on the basis of their likelihoods. When a new trajectory point is acquired, past hypotheses of the map-matched route are extended to account for the new observation. One of the advantages of the approach is corresponding likelihood estimation, which can be considered as a way for uncertainty quantification. In this setting, to meet the requirement of HMM on limited state space, for each observation point only a fixed number of position candidates are considered. It leads to considering only one candidate on each map-graph edge (as a rule, the closest one to observation point). Moreover, emission

TABLE I
PERFORMANCE OF BASELINE ALGORITHM (BL) AND IMPLEMENTED ALGORITHM (EHMM) IN TERMS OF ACCURACY AND RMSE

No	Accuracy		RMSE	
	BL	EHMM	BL	EHMM
min	0.01	0.07	3.71	0.07
max	0.73	0.98	19.58	16.94
mean	0.26	0.66	13.14	6.08
std	0.18	0.21	3.88	4.34

and transition probabilities for HMM are predetermined by the authors, and tuning of corresponding parameters is required. It respectively implies distinctive drawbacks, such as cutting-off the angle at crossroads, extra U-turns, back-and-forth jumps on road segment, etc.

For our experiments we need accurate dataset with noised and ground-truth positions. To generate it we used traffic simulator SUMO — Simulation of Urban MObility [20] and OpenStreetMap data [21]. It enables us to generate accurate map ground-truth positions (taking into account road network information from OpenStreetMap), corresponding exact GPS positions and model other observations. Then we added noise to these accurate observations. Noise was modelled according to our assumptions and evaluated on known dataset [22]. Online algorithm was set as a sequence of trajectory reconstructions for sliding window of up to 50 previous observation points (to limit number of layers in corresponding HMM). Performance was measured in terms of accuracy (hitting the correct road-segment) and RMSE for point-to-point correspondence of map-matched and ground-truth positions. As a baseline we used [22] algorithm (with heuristics for probability distribution assumptions). For our algorithm we used 50 trajectories for training HMM, and then applied it for the following trajectories. For training we applied RKHS with Gaussian kernel with various values of $\sigma \in \{0.5, 1, 5, 10\}$, and applied Linear Functional Strategy [13] while converting kernel matrices. For baseline and proposed algorithm outputs we compared accuracy (hit rate, value from 0 to 1) and RMSE for point-to-point correspondences in meters. Results are presented in Table I. We observe dramatic improvement of performance for proposed algorithm. Here we have to notice, that baseline's performance suffers because of fixed heuristics applied for determining emission and transition probability distribution of HMM. Re-trained HMM by means of EM algorithm (Gaussian mixture) shows better performance, although implemented algorithm (EHMM) outperforms it as well, apparently it is because of implicit ensemble of Gaussian kernels in the implemented solution.

B. Seizure prediction on Electroencephalography signal

The electroencephalography (EEG) is crucial tool for monitoring brain activity in various clinical applications. The typical EEG data contains a set of signals measured with electrodes placed on the human scalp. Brain state recognition from EEG signals requires specific signal processing and pre-processing, features extraction, and classification tools. We

TABLE II
BEST PRECISION AND RECALL FOR BENCHMARKED METHODS, FOR BOTH SEMANTIC SEGMENTATION AND SPEAKER SEGMENTATION TASKS [3]

	Semantic segmentation		Speaker segmentation	
	Precision	Recall	Precision	Recall
KFDR	0.72	0.63	0.89	0.90
MMD	0.71	0.58	0.76	0.73
KCD	0.65	0.63	0.78	0.74
HMM	0.73	0.65	0.93	0.96

evaluate our approach on seizure prediction on Electroencephalography (EEG) signal. For more details on experiment setting we refer to [23]. As hidden variable we consider seizure risk, and observation is given by EEG signal and processed cumulative features for given sliding time-interval. In [23] we applied ranking algorithm for seizure risk prediction, and achieved successive prediction rate approximately 83% for time horizon up to 1 minute. That algorithm required a lot of pre-processing and calibration, therefore could not be considered as real-time application. In our current investigation we applied embedding of HMM in RKHS for corresponding hidden state and observation model. We’ve achieved the same accuracy for the same time horizon (see Figure III-B), whereas performance increased dramatically.

C. Temporal Audio Segmentation

Embedding of HMM into RKHS with Tikhonov regularization was studied in [3], although without implicit naming. In [3] experimental results were presented both on segmentation of whole audio track from TV show and on speaker diarization withing the interview segments. Namely, two soundtracks of the French 1980s entertainment TV-shows (“Le Grand Echiquier”) of approximately three hours each, labelled with characteristic, i.e. “applause”, “movie”, “music”, “speech”, “speaker turns”. After data preprocessing, every 10 ms of audio where matched to 13-dimensional vector. The experimental results from [3] are presented in Table II, where the following methods were compared: *regularized kernel Fisher discriminant ratio* (KFDR), which is basically embedded HMM with truncated spectral regularization, *Maximum Mean Discrepancy* (MMD), *Kernel Change Detection* (KCD) algorithms and standard supervised HMM.

Authors mention in [3], that HMM outperforms all the algorithms, but it is explained by rather unrealistic training procedure, as all speakers and possible labels involved are explicitly modelled beforehand in the speech sections, whereas proposed method demonstrated competitive performance with a completely unsupervised approach.

Unfortunately, we didn’t manage to find the mentioned dataset to reproduce the results. Therefore, we applied the same preprocessing technique to dataset [24]. The Free Spoken Digit Dataset consists of 1500 audio records in wav files at 8kHz of English pronunciations of digits by 3 speakers (50 of each digit per speaker). Corresponding digit are easily

TABLE III
PRECISION AND RECALL FOR AUDIO SEGMENTATION TASK

	Speaker segmentation	
	Precision	Recall
HMM	0.82	0.71
EHMM	0.89	0.78
EHMMN	0.88	0.81

labelled, as each file is named in the following format: $\{\text{digitLabel}\}_{\text{speakerName}}_{\text{index}}.wav$.

We split preprocessed dataset into training and testing set, as it was proposed by its owners, namely 10% (records with indices 0–4 inclusive) of recordings for training set, and 90% for testing (indices 5–49).

In this setting, we compare HMM trained with EM algorithm, embedded HMM, and embedded HMM with regularization by means of Linear Functional Strategy over regularized with Nyström subsampling solutions. Corresponding results are presented in Table III.

IV. CONCLUSION

We consider a Hibert space embedding of HMMs as an extension of traditional HMMs to continuous observation distributions. In this setting we apply more advanced regularization techniques comparing to Tikhonov regularization. Simultaneous regularization by means of Nyström-type subsampling and improved optimization technique enable us to use this approach for online data stream mining. Combining Nyström-type subsampling and Linear Functional Strategy apparently reduce error and its variation, presumably, due to boosting effect, although detailed investigation is needed. As further steps of research we consider multi-penalty regularization for multi-dimensional observation models. Note, that combining modern kernel methods, regularization techniques and graphical models significantly improves well-known algorithms, preserving the advantages of each one.

ACKNOWLEDGMENT

Galyna Kriukova would like to thank Prof. Dr. Sergei Pereverzyev, Johann Radon Institute for Computational and Applied Mathematics (RICAM) of the Austrian Academy of Sciences, for sharing his wisdom and support.

REFERENCES

- [1] A. Bargi, R. Y. D. Xu, and M. Piccardi, “Adon hdp-hmm: An adaptive online model for segmentation and classification of sequential data,” *IEEE Transactions on Neural Networks and Learning Systems*, vol. PP, no. 99, pp. 1–16, 2018.
- [2] J. Kohlmorgen and S. Lemm, “A dynamic hmm for on-line segmentation of sequential data,” in *Proceedings of the 14th International Conference on Neural Information Processing Systems: Natural and Synthetic*, ser. NIPS’01. Cambridge, MA, USA: MIT Press, 2001, pp. 793–800. [Online]. Available: <http://dl.acm.org/citation.cfm?id=2980539.2980642>
- [3] Z. Harchaoui, F. Vallet, A. Lung-Yut-Fong, and O. Cappe, “A regularized kernel-based approach to unsupervised audio segmentation,” in *2009 IEEE International Conference on Acoustics, Speech and Signal Processing*, April 2009, pp. 1665–1668.

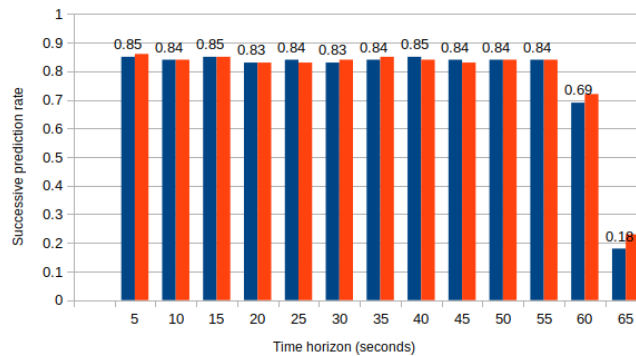


Fig. 1. Accuracy of seizure prediction depending on prediction of time-horizon for baseline algorithm (blue) and proposed method (red)

- [4] K. Muandet, K. Fukumizu, B. Sriperumbudur, and B. Schölkopf, "Kernel mean embedding of distributions: A review and beyond," *Foundations and Trends in Machine Learning*, vol. 10, no. 1-2, pp. 1–141, 2017. [Online]. Available: <http://dx.doi.org/10.1561/22000000060>
- [5] A. Smola, A. Gretton, L. Song, and B. Schölkopf, "A hilbert space embedding for distributions," in *Algorithmic Learning Theory*, M. Hutter, R. A. Servedio, and E. Takimoto, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2007, pp. 13–31.
- [6] L. Song, B. Boots, S. M. Siddiqi, G. Gordon, and A. Smola, "Hilbert space embeddings of hidden markov models," in *Proceedings of the 27th International Conference on International Conference on Machine Learning*, ser. ICML'10. USA: Omnipress, 2010, pp. 991–998. [Online]. Available: <http://dl.acm.org/citation.cfm?id=3104322.3104448>
- [7] L. Song, J. Huang, A. Smola, and K. Fukumizu, "Hilbert space embeddings of conditional distributions with applications to dynamical systems," in *Proceedings of the 26th Annual International Conference on Machine Learning*, ser. ICML '09. New York, NY, USA: ACM, 2009, pp. 961–968. [Online]. Available: <http://doi.acm.org/10.1145/1553374.1553497>
- [8] F. Bauer, S. Pereverzev, and L. Rosasco, "On regularization algorithms in learning theory," *Journal of Complexity*, vol. 23, no. 1, pp. 52 – 72, 2007. [Online]. Available: <https://doi.org/10.1016/j.jco.2006.07.001>
- [9] G. Kriukova, S. P. Jr, and P. Tkachenko, "Nystrm type subsampling analyzed as a regularized projection," *Inverse Problems*, vol. 33, no. 7, p. 074001, 2017. [Online]. Available: <http://stacks.iop.org/0266-5611/33/i=7/a=074001>
- [10] L. Song, J. Huang, A. Smola, and K. Fukumizu, "Hilbert space embeddings of conditional distributions with applications to dynamical systems," in *Proceedings of the 26th Annual International Conference on Machine Learning*, ser. ICML '09. New York, NY, USA: ACM, 2009, pp. 961–968. [Online]. Available: <http://doi.acm.org/10.1145/1553374.1553497>
- [11] L. Song, B. Boots, S. M. Siddiqi, G. Gordon, and A. Smola, "Hilbert space embeddings of hidden markov models," in *Proceedings of the 27th International Conference on International Conference on Machine Learning*, ser. ICML'10. USA: Omnipress, 2010, pp. 991–998. [Online]. Available: <http://dl.acm.org/citation.cfm?id=3104322.3104448>
- [12] G. Kimeldorf and G. Wahba, "Some results on tchebycheffian spline functions," *Journal of Mathematical Analysis and Applications*, vol. 33, no. 1, pp. 82 – 95, 1971. [Online]. Available: <http://www.sciencedirect.com/science/article/pii/0022247X71901843>
- [13] G. Kriukova, O. Panasiuk, S. V. Pereverzyev, and P. Tkachenko, "A linear functional strategy for regularized ranking," *Neural Networks*, vol. 73, pp. 26 – 35, 2016. [Online]. Available: <http://www.sciencedirect.com/science/article/pii/S0893608015001756>
- [14] S. Lu and S. V. Pereverzev, *Regularization theory for ill-posed problems: selected topics*, ser. Inverse and Ill-Posed Problems Series. Berlin: De Gruyter, 2013. [Online]. Available: <http://cds.cern.ch/record/1619978>
- [15] A. J. Smola and B. Schölkopf, "Sparse greedy matrix approximation for machine learning," in *Proceedings of the Seventeenth International Conference on Machine Learning*, ser. ICML '00. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2000, pp. 911–918. [Online]. Available: <http://dl.acm.org/citation.cfm?id=645529.657980>
- [16] C. Williams and M. Seeger, "Using the nystrm method to speed up kernel machines," in *Advances in Neural Information Processing Systems 13*. MIT Press, 2001, pp. 682–688.
- [17] G. Kriukova, N. Shvai, and S. V. Pereverzyev, "Application of regularized ranking and collaborative filtering in predictive alarm algorithm for nocturnal hypoglycemia prevention," in *2017 9th IEEE International Conference on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications (IDAACS)*, vol. 2, Sept 2017, pp. 634–638.
- [18] S. Fong, J. Fiaidhi, S. Mohammed, and L. Moutinho, "Real-time decision rules for diabetes therapy management by data stream mining," *IT Professional*, vol. PP, no. 99, pp. 1–1, 2017.
- [19] C. Y. Goh, J. Dauwels, N. Mitrovic, M. T. Asif, A. Oran, and P. Jaillet, "Online map-matching based on hidden markov model for real-time traffic sensing applications," in *2012 15th International IEEE Conference on Intelligent Transportation Systems*, Sept 2012, pp. 776–781.
- [20] D. Krajzewicz, J. Erdmann, M. Behrisch, and L. Bieker, "Recent development and applications of SUMO - Simulation of Urban MOBility," *International Journal On Advances in Systems and Measurements*, vol. 5, no. 3&4, pp. 128–138, December 2012.
- [21] OpenStreetMap contributors, "Planet dump retrieved from <https://planet.osm.org>," <https://www.openstreetmap.org>, 2017.
- [22] P. Newson and J. Krumm, "Hidden markov map matching through noise and sparseness," in *Proceedings of the 17th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, ser. GIS '09. New York, NY, USA: ACM, 2009, pp. 336–343. [Online]. Available: <http://doi.acm.org/10.1145/1653771.1653818>
- [23] O. Sudakov, G. Kriukova, R. Natarov, V. Gaidar, O. Maximyuk, S. Radchenko, and D. Isaev, "Distributed system for sampling and analysis of electroencephalograms," in *2017 9th IEEE International Conference on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications (IDAACS)*, vol. 1, Sept 2017, pp. 306–310.
- [24] Z. Jackson, C. Souza, J. Flaks, and H. Nicolas, "Jakobovski/free-spoken-digit-dataset v1.0.7," Jan. 2018. [Online]. Available: <https://doi.org/10.5281/zenodo.1136198>

Forecasting the Oil Price with a Periodic Regression ARFIMA-GARCH Process

Daniel Ambach
Department for Data Science
smava GmbH
Berlin, Germany
daniel.ambach@smava.de

Oleksandra Ambach
FirmenCenter Gründung und Nachfolge
Berliner Sparkasse
Berlin, Germany
oleksandra.ambach@berliner-sparkasse.de

Abstract—This article provides a new periodic time series model to predict the oil price. Moreover, the approach discusses short-term forecasting of the oil price. Hence, we discuss the model fit and the out-of-sample performance. Finally, we derive further enhancements and improvements for further research.

Index Terms—long-memory, forecasting, oil-price, ARFIMA, periodic model

I. INTRODUCTION

Supply and demand are two major forces that drive the equilibrium on the market. Foreign exchange rates, interest rates, stock prices are the instruments that play an important role in this process. Though in the modern world there is one more important variable that determines the equilibrium on the market - the oil price. An oil price is a commodity traded on a global market. The crucial thing about it is that in contrast to the interest rates, for example, which have mostly an economic influence, the oil price affects us all being the principle source of the energy and pricing coal and natural gas.

Moreover, as we know the oil price is captured in the oil future contracts. An oil futures contract is described as a binding agreement which provides a the right to purchase a barrel oil at a predefined price on a predefined date in the future. Therefore, the buyer and the seller are obliged to make a deal according to the contract. It means that that the oil prices are manually agreed and accordingly predicted. Things become more difficult with the speculations on the oil market and cyclical trends. It turns out that regardless of how the price on the market is determined, based on its use in fuels and countless consumer goods, the oil price is inevitably in high demand for the future [18].

The aim of our paper is related to the successful application of a periodic regression model and residual process that follows a autoregressive fractional integrated moving average process with generalised autoregressive conditional heteroscedasticity (ARFIMA-GARCH) process to predict the oil price. In recent literature we find on the one-hand different ideas to model the oil price and on the other-hand we find different applications of the ARFIMA-GARCH model. Therefore, we use the properties of our proposed process to model and predict the oil price as good as possible. [3] analyse inflation by the fractionally integrated ARFIMA-GARCH model.

They consider the application of long-memory processes to describing the inflation for ten countries. It is proved that for three high inflation economies there is evidence that the mean and the volatility of inflation interact in a way that is consistent with the Friedman hypothesis. [17] use the ARFIMA-GARCH Model and apply it to the realized volatility and the continuous sample path variations constructed from high-frequency Nikkei 225 data. [20] consider a periodic seasonal Reg-ARFIMA-GARCH model for daily electricity spot prices. This approach depicts periodic extensions of dynamic long-memory regression models with autoregressive conditional heteroscedastic errors. There model is accurate to analyse and predict the daily spot prices. [21] sufficiently explore the heart rate variability data with its stationary characteristics, long range correlations and instantiations volatility with the help of ARFIMA GARCH model. Another application is presented by [19]. They assess the persistence dependence of rainfall time series of Chui Chak, a station in Peninsular Malaysia that observed the highest rainfall event for the period 1975-01-01 to 2008-12-31.

The theory and the application of the periodic regression with ARFIMA-GARCH process is discussed in this paper. There is an evidence in the literature that there exist characteristics in the data which enables the use of the aforementioned model. For example we find cyclic behaviour throughout the trading year, high autocorrelation which is related to the seasonal behaviour and use of oil, conditional heteroscedasticity related to the volatility of the price process and heavy tailed residuals modelled by a t-distribution. Therefore, we will discuss the theoretical background and the modelling process. After that we will apply the model to the in-sample data, and if it fits well, we will try to forecast the out-of-sample results.

This article starts with a description of an ARFIMA-GARCH in Section II. Hereafter, the theory of long memory process, GARCH process, model fit and model diagnostics is presented. Section III discusses the application of the aforementioned model to the data set and presents in-sample results. Section IV provides the out-of sample results and discusses improvements and finally V concludes.

II. THEORETICAL BACKGROUND

The oil price (P_t) is from the economic sense an important indicator. Figure 1 shows all observations from 1986-01-02 to 2017-01-09 of the West Texas Intermediate (WTI) Cushing oil price. The oil price incorporates different properties. It is reasonable to assume that the oil price has periodic as well as autoregressive disturbances. Moreover, the variance process will show conditional heteroscedasticity. Therefore, we will apply the following model to the oil price

$$P_t = \mu + Periodic^{Reg} + \epsilon_t, \quad (1)$$

where μ is an intercept and ϵ_t is the residual process that follows an autoregressive fractional integrated moving average process with generalised autoregressive conditional heteroscedasticity (ARFIMA-GARCH). Furthermore, we model the periodic or seasonal structure by periodic sine and cosine functions which are

$$Periodic^{Reg} = \left(a_{\cos} \cos\left(\frac{2\pi t}{P}\right) + a_{\sin} \sin\left(\frac{2\pi t}{P}\right) \right), \quad (2)$$

where P is the period which is the basis of the periodic functions. Here we suggest that the period could be related to an annual cycle.

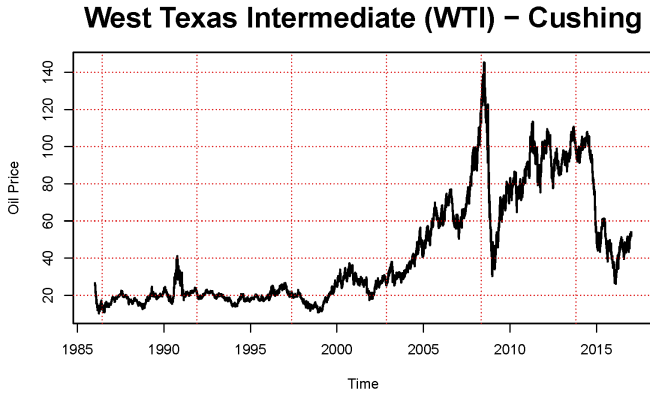


Fig. 1. Price of the West Texas Intermediate (WTI) Cushing, full sample (left)

A. the long-memory process

We introduce the ARFIMA-GARCH model to describe the autocorrelation and conditional heteroscedasticity. The residual process $\{\epsilon_t\}$ foremost as ARFIMA(p,d,q)-GARCH(P,Q) model in the following way

$$\epsilon_t \equiv \phi(B)\nabla^d X_t = \theta(B)\eta_t \quad (3)$$

$$\eta_t = \sigma_t \tau_t, \quad (4)$$

$$\sigma_t^2 = \alpha_0 + \sum_{i=1}^q \alpha_i \eta_{t-i}^2 + \sum_{j=1}^p \gamma_j \sigma_{t-j}^2, \quad (5)$$

$$\tau_t \sim F \quad \text{here we take the t-distribution.} \quad (6)$$

Obviously equation (3) describes a typical ARFIMA(p,d,q) process and (4) depicts a GARCH(P,Q) process. Quite frequently long memory processes are considered while analyzing environmental time series [16]. In contrast an ARMA process denoted as U_t is considered as a short-memory process because the covariance of U_1 and U_{1+k} is decreasing fast as k converges to ∞ . [7] explain that the autocorrelation function of U_t is geometrically bounded and therefore shows a short dependence structure. In contrast a long memory process has an autocorrelation function for which $\rho(k) \sim Ck^{2d-1}$ as $k \rightarrow \infty$, where C is a constant with $C \neq 0$ and $d < 0.5$ [7]. [16] use the ARFIMA model to predict the wind speed in Ireland.

Defining for any real number of d and $d > -1$, the difference operator $\nabla^d = (1-B)^d$ by the binomial expansion:

$$\nabla^d = (1-B)^d = \sum_{k=0}^{\infty} \frac{(k-d-1)!}{k!(-d-1)!} B^k,$$

where B is the backward shift operator $B^v X_t = X_{t-v}$.

In the continuous case [22] introduce the fractional Brownian motion. In the discrete setting the fractional integrated noise are the following difference equations $\nabla^d X_t = Z_t$, where $\{Z_t\} \sim WN(0, \sigma^2)$ and $d \in (-0.5, 0.5)$ (see [7]).

The ARFIMA(p,d,q) and GARCH(P,Q) process is a combination of an ARFIMA and a GARCH model. The contribution of a process which has a fractionally integrated conditional mean and generalized autoregressive conditional heteroscedasticity is to cover the effects that are not probably modelled by short time memory and a constant variance. An initial approach is given in [15] wherein ARMA-ARCH model explains macroeconomic time series. [2] used an ARFIMA-GARCH model to describe inflation rates. Moreover, they develop an approximate maximum likelihood estimate of an ARFIMA-GARCH process. The ARCH model [10] and GARCH model [4] are invented to model financial time series. The combined ARFIMA-GARCH model is given by

After discussing different time series models for the residual process we have to apply them to the data set and proceed with the parameter estimation and diagnostic checking. The next Section provides such results as well as the estimation of the complete univariate time series model.

B. Parameter estimation and model selection

The estimation of the complete model could be done in two different ways. On the one-hand it is possible to use a two step approach, where first the mean and hereafter the variance is estimated. Hence, on the other-hand the whole model is estimated in a single maximum likelihood approach. However, the exact likelihood is unknown, therefore we consider the conditional (quasi) maximum likelihood. For deriving the quasi maximum likelihood we can use the explanations and derivations of [7] and [26].

Finally, we assume that the residual process $\{\tau_t\}$ has to be independent and equally distributed with $\{\tau_t\} \sim t$. The t-distribution provides better results related to the tail behaviour of the residuals. The estimated periodic regression model with

ARFIMA-GARCH residual process is evaluated by means of the in-sample and the out-of-sample performance. [11] describes the R-square R^2 as a possible criterion for the in-sample performance of a regression model. Unfortunately, while using correlated data, we do not obtain an unbiased value of R^2 . Hence, we can not interpret the R^2 so easily. One alternative is related to the goodness of the model. Therefore, we calculate the autocorrelation function ACF and the partial autocorrelation function.

The autocorrelation function can be used for two purposes. The first reason is to detect non-randomness in data. Moreover, we identify an appropriate time series model if the data are not random. If we want to identify the correct model order of an autoregressive model and figure out whether we observe further autocorrelation, we have to consider the partial autocorrelation function as well. These functions are also useful for the model diagnostic, if they provide evidence for further autocorrelation, we observe that the goodness of our model is not sufficient.

[1] proposes to measure the goodness of fit for a certain model by balancing the error of the fitted model against the number of parameters in the estimated model. The Akaike information criterion (AIC) is given by [1]

$$AIC(k) = \ln \hat{\sigma}(k)^2 + k \frac{2}{n}, \quad (7)$$

where $\hat{\sigma}(k)^2$ is the estimated variance, k is the number of all parameters in the complete model and n is the sample size. Another information criterion is the Bayesian information criterion (BIC). [24] propose the BIC which has a larger punishing term. The BIC is given by

$$BIC(k) = \ln \hat{\sigma}(k)^2 + k \frac{\ln n}{n} \quad (8)$$

Both information criteria are appropriate to choose the best model order of our ARFIMA-GARCH model, but they could select a different model order in the end. The reason for such a result is related to the punishing terms $k \frac{2}{n}$ and $k \frac{\ln n}{n}$, which are different. We calculate the information criterion for each model order of the ARFIMA-GARCH process. Hence, the AIC and the BIC can be derived by means of the likelihood function instead of $\hat{\sigma}(k)^2$. The optimal model order minimizes the AIC and BIC.

The optimal model provides necessary information for the in-sample performance, but in addition, we are able to predict different unobserved values. [7] discusses the prediction of ARMA, ARFIMA and GARCH model.

III. ANALYSING THE OIL PRICE AND THE GOODNESS OF FIT

The regression model with ARFIMA-GARCH residual process with t-distribution given by (1) - (3) is applied to our oil price data set of the West Texas Intermediate (WTI). Figure 2 shows a part of the data set which is modelled with the aforementioned approach

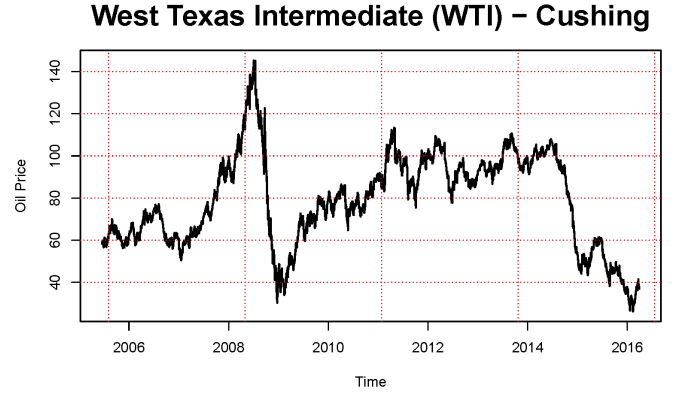


Fig. 2. Price of the West Texas Intermediate (WTI) Cushing

The data set is provided by “U.S. Energy Information Administration” and the investigated time horizon reaches from 2005-06-17 till 2016-03-29. The regressors of the model are given by periodic variables.

$$P_t = \mu + \text{Periodic}^{Reg} + \epsilon_t, \quad (9)$$

$$\text{Periodic}^{Reg} = \left(a_{\cos} \cos \left(\frac{2\pi t}{252} \right) + a_{\sin} \sin \left(\frac{2\pi t}{252} \right) \right) (10)$$

where the period is 252, which is related to an average trading year of 252 days. The price process P_t has to be analysed according to the autocorrelation structure and to find heteroscedastic effects. Figure 3 provides a huge autocorrelation structure and the PACF selects a positive autocorrelation for the first lag and some more autocorrelation for the following lags.

Furthermore, Figure 5 depicts a high presence of conditional heteroscedasticity. From the aforementioned findings, we may assume that the proposed periodic regression model with ARFIMA-GARCH residuals are appropriate to capture the main characteristics of the WTI oil price. Subsequently, we fit the model to our data set. Table III shows the model estimation results. The derived model order is related to the smallest AIC and BIC and significance of the parameters. The majority of the parameters are significant, but some GARCH parameters are not. Moreover, the periodic regressor which is modelled with the sine function is only significant to a significance level of 0.1, which is acceptable in practice.

The obtained estimation results seem to be sufficient, but we have to discuss the goodness of fit. The data set has to be uncorrelated, homoscedastic and the remaining residuals should follow the t-distribution. The ACF of the residuals does not provide any correlation structure. Furthermore, we detect only for the first lag of the squared standardized residuals a remaining presence of autocorrelation. In addition we can observe that the Ljung and Box test points in the same direction. The test is applied to the residuals as well as to the squared standardised residuals. Moreover, a weighted ARCH-LM test for detecting further conditional heteroscedasticity is

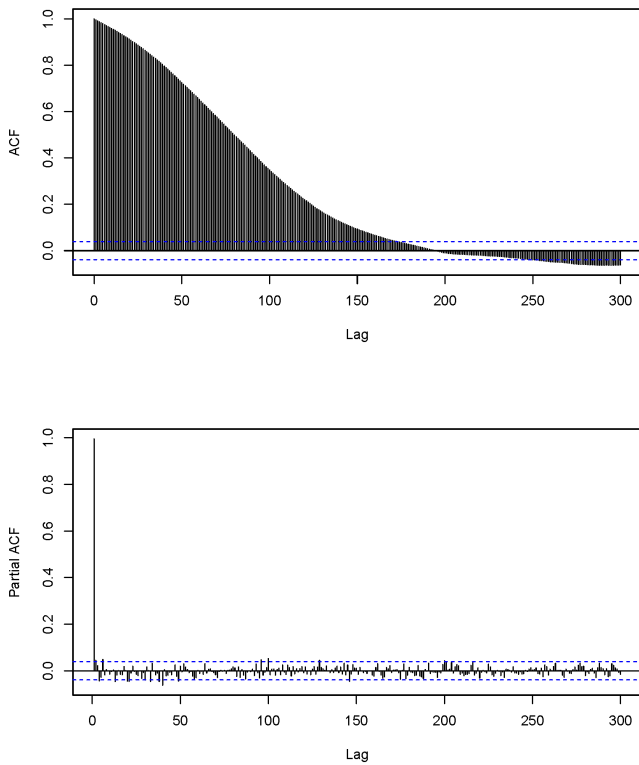


Fig. 3. Autocorrelation function and partial autocorrelation function of the data set.

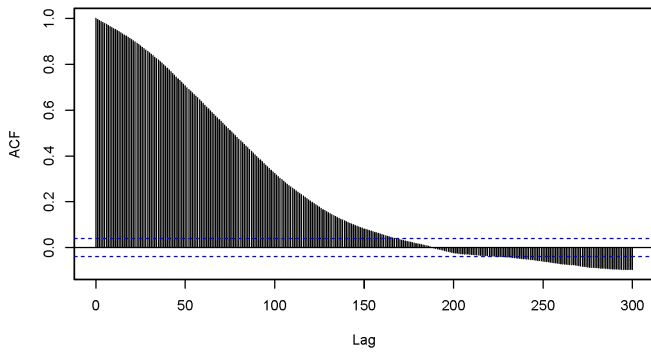


Fig. 4. Autocorrelation function of the squared observations.

applied. [13] propose the test which is much better for the distribution of the statistics of the values from the estimated models. The ARCH-LM test is a weighted portmanteau test. Under the null hypothesis it is assumed that the ARCH process is fitted accurately. The weighted ARCH-LM test does not reject the null hypothesis and thus we are able to conclude, that there is no further improvement of the ARFIMA-GARCH model order.

Figure 6 depicts the Quantile-quantile (Q-Q) plot, which

TABLE I
PARAMETER ESTIMATION OF THE OIL PRICE, WHERE THE BOLDDED VALUES PROVIDE SIGNIFICANCE GIVEN A SIGNIFICANCE LIMIT OF $\alpha = 0.05$

	Estimate	Std. Error	t value	Pr(> t)
Regression coefficients				
μ	55.301296	1.768863	31.26376	0.000000
a_{cos}	3.342449	1.544597	2.16396	0.030467
a_{sin}	2.566762	1.509088	1.70087	0.088967**
ARFIMA parameters				
d	0.099415	0.031782	3.12802	0.001760
ϕ_1	1.741240	0.000271	6419.64884	0.000000
ϕ_2	-0.672959	0.000128	-5247.05016	0.000000
ϕ_3	-0.053276	0.003060	-17.41137	0.000000
ϕ_4	-0.015083	0.003355	-4.49510	0.000007
θ_1	-0.876162	0.017276	-50.71414	0.000000
GARCH parameters				
α_0	0.043254	0.016544	2.61447	0.008937
α_1	0.081730	0.017577	4.64975	0.000003
γ_1	0.311167	0.269917	1.15282	0.248983
γ_2	0.217017	0.285986	0.75884	0.447949
γ_3	0.372788	0.174081	2.14146	0.032237
df	9.221080	1.410838	6.53589	0.000000
Diagnostic checking			Statistics	
AIC			3.6526	
BIC			3.6851	

shows that our considered t-distribution is a good choice for the heavy tailed residuals. This argument is supported by the Pearson chi-squared goodness of fit test.

Finally, we are able to sum up the in-sample goodness of fit. It turns out, that the model provides good in-sample performance. The model assumptions are fulfilled, Only the first lag of our squared standardized residuals shows a minor presence of autocorrelation. The next step is related to the out-of-sample forecast.

IV. FORECASTING PERFORMANCE OF THE MODEL

The Figure 7 below shows the oil prices in 2016 and our forecast up to 21 trading day. The quality of the forecast is not completely satisfying as the forecast does not reflect the cycle behaviour. Even the in-sample results are not completely satisfying. We observe that the 99%-confidence intervals are too narrow. Moreover, the realised values in future and our forecasts are different. Thereby the mean changes, but the periodic or cyclic behaviour is not captured accurately. The 99%-prediction interval is relatively narrow, which is a good result, but the observations lie outside the bounds. Nevertheless, the shortcomings of the forecast do not provide a good model result. It is obvious, that we need further improvements of the applied model. One possibility is related to the periodic behaviour which has to be captured by different and more periodic regressors. Furthermore, we can conclude that the process includes different changes which are suitable captured by a non-linear model.

V. CONCLUSION

The oil price is a very important indicator for the economy in the world. We need oil for everything so the development of the price is very important. We can find a direct correlation between the cost of gasoline or airplane fuel to the price

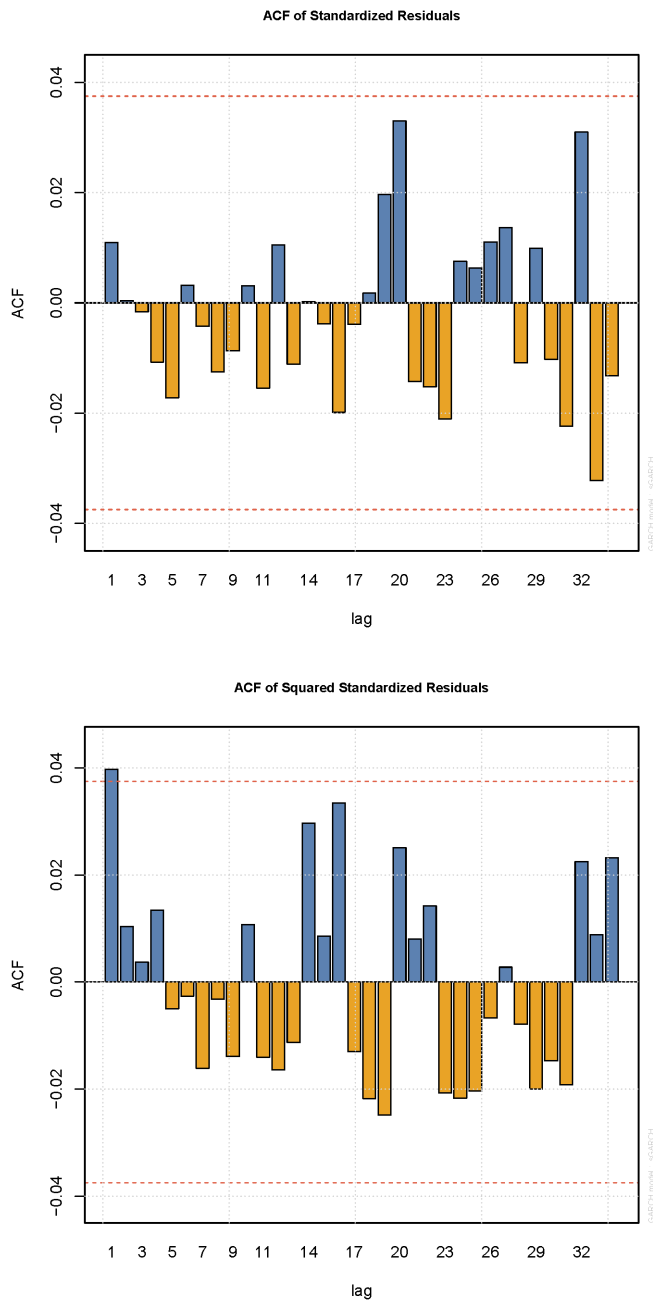


Fig. 5. Autocorrelation function of the standardised residuals and of the squared residuals.

of transporting goods and people. A drop in fuel prices means lower transport costs and cheaper airline tickets, lower transportation costs for apples from Italy or furniture from China. Therefore, we discuss the application of a periodic regression model with ARFIMA-GARCH residual process to model and predict the oil price.

The aforementioned model captures the long-memory and the conditional heteroscedasticity. In addition we include two periodic regressors. The model provides some advantages, but

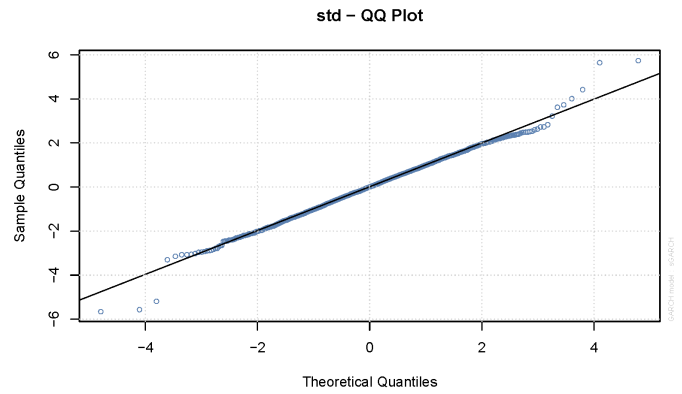


Fig. 6. Quantile-quantile plot of the residuals and the theoretical t-distribution.

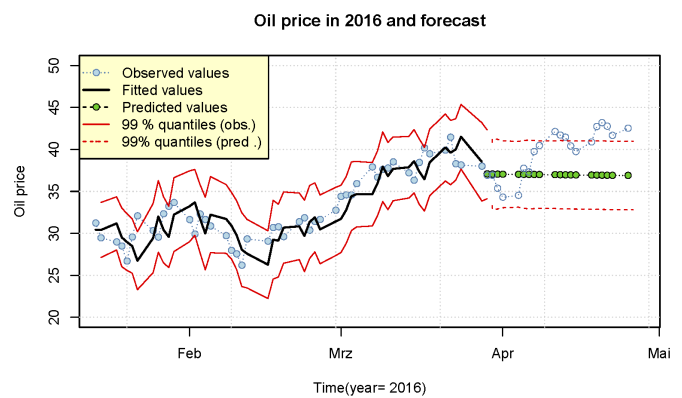


Fig. 7. Oil price forecasting for 21 future observations.

it fails to capture the periodicity in a good way. Moreover, the model shows for the first lag of the squared standardised residuals a remaining presence of correlation, which is not satisfying at all.

The model should be extended. The volatility process could be modelled by an asymmetric or threshold GARCH model. Moreover, it might be useful that the power of the volatility is changed, so that extreme values in the volatility do not have such a big impact. The periodic regression part of the model might be extended by a stacked regularised model that captures the non-linear disturbances. The use of different periodic functions and different periods as basis has a huge influence on the structure of the model. Finally, we detect a suitable model for the oil price which provides much space for further research.

REFERENCES

- [1] Akaike, H. (1974), A new look at the statistical model identification, *Automatic Control, IEEE Transactions on*, 19(6), pp. 716–723.
- [2] Baillie, R.T. (1996), Long memory processes and fractional integration in econometrics, *Journal of econometrics*, 73(1), pp. 5–59.
- [3] Baillie, R.T., Chung, C.F., and Tieslau, M.A. (1996), Analysing inflation by the fractionally integrated ARFIMA-GARCH model, *Journal of applied econometrics*, pp. 23–40.

- [4] Bollerslev, T. (1986), Generalized autoregressive conditional heteroskedasticity, *Journal of econometrics*, 31(3), pp. 307–327.
- [5] Box, G.E. and Pierce, D.A. (1970), Distribution of residual autocorrelations in autoregressive-integrated moving average time series models, *Journal of the American statistical Association*, 65(332), pp. 1509–1526.
- [6] Breusch, T.S. and Pagan, A.R. (1979), A simple test for heteroscedasticity and random coefficient variation, *Econometrica: Journal of the Econometric Society*, pp. 1287–1294.
- [7] Brockwell, P.J. and Davis, R.A. (2009), *Time series: theory and methods*, Springer, New York.
- [8] Brockwell, P.J. and Davis, R.A. (2013), *Time series: theory and methods*, Springer Science & Business Media.
- [9] Durbin, J. and Watson, G.S. (1951), Testing for serial correlation in least squares regression. II, *Biometrika*, 38(1/2), pp. 159–177.
- [10] Engle, R.F. (1982), Autoregressive conditional heteroscedasticity with estimates of the variance of United Kingdom inflation, *Econometrica: Journal of the Econometric Society*, pp. 987–1007.
- [11] Fahrmeir, L., Kneib, T., and Lang, S. (2007a), *Regression: Modelle, Methoden und Anwendungen*, Springer-Verlag.
- [12] Fahrmeir, L., Künstler, R., Pigeot, I., and Tutz, G. (2007b), *Statistik: Der Weg zur Datenanalyse*, Springer-Verlag.
- [13] Fisher, T.J. and Gallagher, C.M. (2012), New weighted portmanteau statistics for time series goodness of fit testing, *Journal of the American Statistical Association*, 107(498), pp. 777–787.
- [14] Goldfeld, S.M. and Quandt, R.E. (1965), Some tests for homoscedasticity, *Journal of the American statistical Association*, 60(310), pp. 539–547.
- [15] Greenwald, B.C., Stiglitz, J.E., and Weiss, A. (1984), Informational imperfections in the capital market and macro-economic fluctuations.
- [16] Haslett, J. and Raftery, A.E. (1989), Space-time modelling with long-memory dependence: Assessing Ireland’s wind power resource, *Applied Statistics*, 30(1), pp. 1–50.
- [17] Ishida, I., Watanabe, T., et al. (2009), Modeling and Forecasting the Volatility of the Nikkei 225 realized Volatility using the ARFIMA-GARCH model, *Global COE Hi-Stat Discussion Paper*, 32.
- [18] Kalkman, J., Pfeiffer, W., and Pereira, S. (2013), Are we running out of oil?
- [19] Kane, I.L. and Yusof, F. (2013), Assessment of Risk of Rainfall Events with a Hybrid of ARFIMA-GARCH, *Modern Applied Science*, 7(12), p. 78.
- [20] Koopman, S.J., Ooms, M., and Carnero, M.A. (2007), Periodic seasonal Reg-ARFIMA–GARCH models for daily electricity spot prices, *Journal of the American Statistical Association*, 102(477), pp. 16–27.
- [21] Leite, A., Rocha, A., and Silva, M. (2009), Long memory and volatility in HRV: an ARFIMA-GARCH approach, *Computers in Cardiology*, 2009, IEEE, pp. 165–168.
- [22] Mandelbrot, B.B. and Van Ness, J.W. (1968), Fractional Brownian motions, fractional noises and applications, *SIAM review*, 10(4), pp. 422–437.
- [23] Palm, F.C. (1996), 7 GARCH models of volatility, *Handbook of statistics*, 14, pp. 209–240.
- [24] Schwarz, G. et al. (1978), Estimating the dimension of a model, *The annals of statistics*, 6(2), pp. 461–464.
- [25] Shapiro, S.S. and Wilk, M.B. (1965), An analysis of variance test for normality (complete samples), *Biometrika*, 52(3/4), pp. 591–611.
- [26] Shumway, R.H. and Stoffer, D.S. (2010), *Time series analysis and its applications: with R examples*, Springer Science & Business Media.
- [27] White, H. (1980), A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity, *Econometrica: Journal of the Econometric Society*, pp. 817–838.

Improvement of Character Segmentation using Recurrent Neural Networks and Dynamic Programming

Valentyna Volkova

Samsung R&D Institute Ukraine (SRK)

Kyiv, Ukraine

v.volkova@samsung.com

Ivan Deriuga

Samsung R&D Institute Ukraine (SRK)

Kyiv, Ukraine

i.deriuga@samsung.com

Vadym Osadchyi

Samsung R&D Institute Ukraine (SRK)

Kyiv, Ukraine

vad.osadchiy@samsung.com

Olga Radyvonenko

Samsung R&D Institute Ukraine (SRK)

Kyiv, Ukraine

oradivonenko@gmail.com

Abstract—A common characteristic of all the existing online handwritten text recognition algorithms is that the character segmentation process is closely related to the recognition process. There are different approaches to segment data but all of them don't give absolutely correctly segmentation results due to specifics of handwriting data input. In this paper, we present a new approach for character segmentation improvement in online handwriting recognition which is based on using recurrent neural networks and dynamic programming. Due to online handwritten text is a sequence of points we propose to use Bidirectional Long Short-Term Memory (BLSTM) for classification of decoder outputs and dynamic programming for interpretation of classification results. Experimental evaluation shows the effectiveness of a proposed approach in increasing of segmentation quality.

Index Terms—online handwriting recognition, character segmentation, recurrent neural networks, dynamic programming

I. INTRODUCTION

Handwriting recognition applications are popular and useful in business, education and other because allow to make quickly handwritten notes and convert them into printed text. Also, handwriting input is an alternative to using keyboards for pen-based, touch-based tablets and smartphones [1] – [2].

Character segmentation is an important part of online handwriting recognition process. The goal of character segmentation is to partition of a handwritten text into segments, each containing an isolated and complete character.

Distinguish several types of segmentation for handwriting recognition: line, word and character segmentation. Each type is important and has a big influence on recognition result in general. Especially it influences on recognition such entities as formulae where the quality of relations between symbols is very important. Also, it is useful in solving of the ink beautification problem where every handwritten stroke in a text has to be modified in a better way to beautify the general text representation. Such correction depends not only on geometric

characteristics of a stroke but also on a quality of character segmentation [3].

This work considers a character segmentation task as an important step in the recognition process and designs the method of segmentation improvement based on using of recurrent neural networks (RNN) and dynamic programming.

The paper has next structure: introduction, related publications with short overview represented in section 2. Section 3 describes the proposed model architecture, feature selection, and proposed dynamic programming. Experimental results are presented in section 4. The conclusion of the paper is given in Section 5.

II. BACKGROUND

Character segmentation is an operation that attempts to decompose a sequence of strokes into subsequences of individual strokes. It is one of the crucial processes in a system for online handwriting recognition.

There are exist different segmentation algorithms [4] – [6]. And all of them don't give absolutely correctly segmentation results or aimed at solving a word, line segmentation problems.

For character segmentation usually are used classifiers such as SVM, BLSTM, decision trees and dynamic programming [7] – [12]. These methods are applied after removing delayed strokes from the handwritten text and potential breakpoints. Breakpoints are detected after the shape analysis of the stroke trajectory to find the best segmentation point for each character. These approaches allow finding a global optimal path of segments.

The difficulties in segmenting of handwritten text arise due to the following factors:

- 1) characters in cursive writing usually are connected;
- 2) character shaping depends on its position in the word and what characters are next;
- 3) neighboring characters in a word may overlap;

- 4) delayed strokes (the dot of "i", "j" or the crossing "f", "t", "H", "F", "E");
- 5) the variance of writing styles.

Most of considered approached use a pre-segmentation step for decoding improvement. Specifically, character segmentation is used to speed-up the decoding by pruning word lattice on only segmentation points. The goal of this paper is to improve the character segmentation after decoding without affecting of the recognition result.

III. PROPOSED APPROACH

A. Model architecture

To improve the quality of character segmentation it is proposed to use RNN which corrects the segmentation points received from the output of the decoder. Segmentation point is a transition between recognition entities (letters, digits etc.). The goal of this RNN is to classify input strokes into N classes: $N - 1$ classes are potentially character segments and "0" class reserved for marking of delayed strokes. A number of classes can be different. It is better to use one class for delayed strokes and three or more for segments.

In this work we use three classes for segments and experiments have shown that such amount is sufficient to obtain a good quality of character segmentation. As result, the input sequence of points is marked as four classes. The class change in this sequence shows the beginning of a new character (Fig. 1).

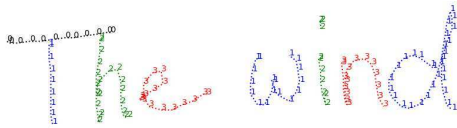


Fig. 1. Example of marked by classes segments.

A general scheme of the proposed algorithm is shown in Fig. 2. Due to input handwritten data is represented as a sequence of points we propose to use BLSTM [13] for classification of decoder outputs.

According to machine learning principles, an input dataset have to be divided into train, validation and test sets and preliminary preprocessed.

To preprocess data we propose to use the following algorithms: size normalization, density normalization of points (interpolating for missing points and resampling if there is overmuch of points), Bezier smoothing, slant and skew correction [14]. Multilines have to be separated into single lines.

Preprocessed samples are fed to a handwriting recognition neural network. As already said for online text recognition we propose to use a BLSTM neural network which is well-proven in solving such type of problems.

From preprocessed samples, there are extracted features which fed to the input of recognition BLSTM.

As recognition BLSTM in this work is used decoder from the RNNLIB library [15]. RNNLIB is a recurrent neural network library for sequence learning problems which has proven particularly effective for speech and handwriting recognition.

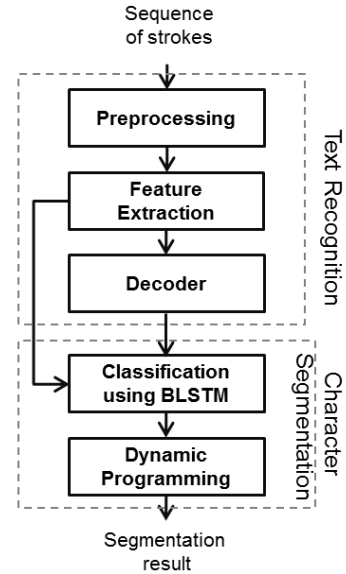


Fig. 2. General scheme of proposed approach.

After the recognition training is done we mark the obtained sequence of points according to the output of decoder (decoding results) using segmentation BLSTM (Fig. 3) as it was described earlier. Decoding results can be approved by implementation of the token passing algorithm [17]. The proposed segmentation classifier contains two BLSTM hidden layers. Every layer consists of 20 cells.

As inputs of segmentation BLSTM are four features. Three of them are extracted on the text recognition step and the last one the output of decoder (the recognition BLSTM) which represents preliminarily segmented by frames characters. More detailed description of this features is given in Section 3.2.

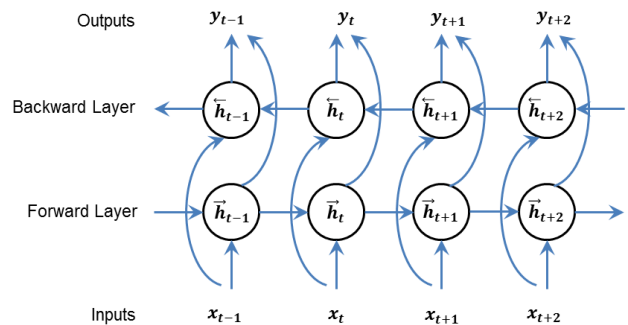


Fig. 3. Segmentation BLSTM.

The network is trained to minimize the cross-entropy error of the targets using a softmax output layer. For interpretation of obtained results, we apply the dynamic programming described in Section 3.3.

B. Feature extraction

The common set of features for character segmentation is given in the Table I.

TABLE I
FEATURES FOR CHARACTER SEGMENTATION

Feature	Description
$f_{\Delta x}$	delta x coordinates
$f_{\Delta y}$	delta y coordinates
f_{upd}	pen-up/pen-down. Means the points in the sequence when the individual strokes end
f_{output}	output of the previous stage (sequence of points marked according to output of decoder after recognition)

At the input of the segmentation neural network are fed four features: f_{output} which represents decoding results and $f_{\Delta x}$, $f_{\Delta y}$, f_{upd} obtained on the feature extraction stage for the recognition BLSTM training.

Features $f_{\Delta x}$, $f_{\Delta y}$ and f_{upd} are extracted from the input handwritten sequence of strokes on a online text recognition stage. These characteristics are classic for online handwriting recognition using RNN and are used in the recognition training.

As the fourth feature can be used a raw decoder output when each frame is marked with a letter code. Using of such feature is possible because a codebook for concrete language is limited (less than 256 characters) and thus it can be normalized. Such feature gives some segmentation improvement but is sensitive to a language.

To achieve character segmentation which doesn't depend on language codebook there can be used the binarization of decoding output. In this case segmentation points (frames) are marked as "1" and all other regular frames are marked as "0". This feature gives a smaller improvement comparing to the previous one but can be used for different language groups.

In this paper as fourth feature we propose to use the specific representation of a decoder output with its initial segmentation which has to be further improved. For this we select N segmentation classes for ordinary char segments and one special for delayed strokes ("0" class). All frames which belong to one character segment are marked with the same class. All delayed stroke frames are marked with class "0". Each character segment is marked in ordered way starting from "1" class and following class numbers, when the next non-delayed stroke segment has a next class number by module N .

Using of this feature gives better segmentation results than previous one and in this case segmentation is language independent.

C. Dynamic programming

Neural network output has to be properly interpreted. Outputs of the segmentation RNN discussed in this work can be considered as probabilities $\tilde{P}(\omega_t)$ of a frame belonging t (point) to a segment class ω . By analogy with the output feature there was defined four segment classes: from "1" to "3" are for regular segments and "0" class is for delayed strokes. Then a character segmentation can be formulated as a classical maximization likelihood decoding problem:

$$\begin{aligned} L(\{\omega\}_{t=1, \overline{T}}; \{(x, y)\}_{t=1, \overline{T}}) &= \\ &= P(\{\omega\}_{t=1, \overline{T}} | \{(x, y)\}_{t=1, \overline{T}}) = \\ &= \prod_{t=1}^T P(\omega_t | \{\omega\}_{t' < t}; \{(x, y)\}_{t=1, \overline{T}}) = \prod_{t=1}^T \tilde{P}(\omega_t), \end{aligned} \quad (1)$$

$$\{\omega_t^0\}_{t=1, \overline{T}} = \operatorname{argmax}_{\{\omega_t\}_{t=1, \overline{T}}} L(\{\omega\}_{t=1, \overline{T}}; \{(x, y)\}_{t=1, \overline{T}}), \quad (2)$$

where $\{\omega_t^0\}_{t=1, \overline{T}} = \omega_1^0, \dots, \omega_T^0$ is a sequence of segment classes with a special restriction for segment numbers K .

From one side there is known an expected segment number. It is equal to a length of a character sequence output of decoder without counting of spaces. In this case, one character corresponds to one segment. From another side we can count a character segments number K from a sequence $\{\omega\}_{t=1, \overline{T}}$ in a next way:

$$K = |S(\{\omega_t^0\}_{t=1, \overline{T}})| = |\{\tilde{\omega}_t^0\}_{t=1, \overline{T}'}| = T', \quad (3)$$

where S is a sequence transformation which suppose removing of $\omega_{t_0}^0$ from $\{\omega_t^0\}_{t=1, \overline{T}}$ if $\omega_{t_0}^0 = 0$ or $\omega_{t_0-1}^0 = \omega_{t_0}^0$. After such transformation we obtain a reduced sequence $\{\tilde{\omega}_t^0\}_{t=1, \overline{K}}$ with a length equal to a character segment count.

Previously many researchers noted relation between a maximum likelihood decoding and DTW algorithm [17]. Following this idea there was decided to apply dynamic programming with a logarithmic likelihood estimation function:

$$\begin{aligned} \log L(\{\omega\}_{t=1, \overline{T}}; \{(x, y)\}_{t=1, \overline{T}}) &= \\ &= \sum_{t=1}^T \log \tilde{P}(\omega_t) = \sum_{t=1}^T c_t(\omega_t), \end{aligned} \quad (4)$$

where $c_t(\omega_t) = \log \tilde{P}(\omega_t)$.

$$\{\omega_t^0\}_{t=1, \overline{T}} = \operatorname{argmax}_{\{\omega_t\}_{t=1, \overline{T}}} \log L(\{\omega\}_{t=1, \overline{T}}; \{(x, y)\}_{t=1, \overline{T}}). \quad (5)$$

In terms of dynamic programming there were defined conventions (Table II).

TABLE II
CONVENTIONS.

Designation	Description
$dp_t(k, s)$	the cost function of traversed path to segment k of time t , where $t = \overline{1, T}$
T	a general number of frames
k	a number of current segment, $k = \overline{1, K}$
K	a total number of segments
s	a binary state which represents a delayed stroke with values $s = \{0, 1\}$

Relation between segment number k and segment class can be described as follows:

$$\omega = k \bmod N + 1 \text{ if } s = 1 \text{ and } \omega = 0 \text{ if } s = 0, \quad (6)$$

where N is a number of segment classes, in this paper $N = 3$. There is a special extra class for delayed strokes and thus total number of segment classes is $N + 1$.

Also, there were applied the following restrictions for dynamic programming related to segment classes possible transitions:

- 1) the first class can be changed only by the second or by the delayed class, or not changed;
- 2) the second class can be changed only by the third or the delayed classes, or not changed;
- 3) the third class can be changed only by the first class or the delayed classes, or not changed.
- 4) a delayed stroke can be only the stroke as a whole;
- 5) the zero class can be changed by the previous one.

Possible state transitions between classes are presented in Fig. 4.

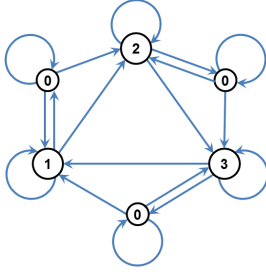


Fig. 4. Transitions between segmentation classes.

Taking into account described restrictions the traversing path for proposed dynamic programming is given in Fig. 5

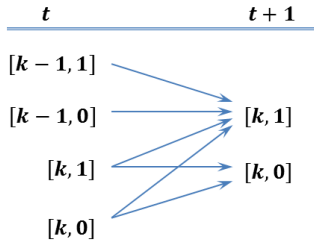


Fig. 5. Traversing path.

Recurrent equations for cost functions can be represented as follows:

Initial state:

$$t = 0$$

$$k = 1$$

$$s = 1$$

$$dp_0(1, 1) = c_0(1)$$

General step:

$$dp_{t+1}(k, 1) = \begin{cases} dp_t(k, 1) + c_{t+1}(k \bmod N + 1); \\ dp_t(k - 1, 1) + c_{t+1}((k - 1) \bmod N + 1); \\ dp_t(k - 1, 0) + c_{t+1}((k - 1) \bmod N + 1) \end{cases} \quad (7)$$

if the stroke begins in current frame;

$$dp_{t+1}(k, 1) = \begin{cases} dp_t(k, 0) + c_{t+1}(k \bmod N + 1) \end{cases} \quad (8)$$

if the stroke begins in current frame,

$$dp_{t+1}(k, 0) = \max \begin{cases} dp_t(k, 0) + c_{t+1}(0); \\ dp_t(k, 1) + c_{t+1}(0); \end{cases} \quad (8)$$

Next state is given by

$$(k_{t+1}, s_{t+1}) = \underset{(k, s)}{\operatorname{argmax}} (dp_{t+1}(k_t, 0), dp_{t+1}(k_t, 1), dp_{t+1}(k_{t+1}, 1)) \quad (9)$$

IV. EXPERIMENTS

Experiments were conducted on IAM online database (IAMonDo) [16] and handwritten dataset (HWRD) collected for training and testing of the proposed solution.

Both datasets consist of pen trajectories collected from different writers using a whiteboard (IAMonDo) and pen on a smartphone or hand on a tablet. The HWRD contains 11297 sequences for the English language. IAMonDo contains about 3859 sequences for the English language, a number of sequences with 100% recognition accuracy is 2043. For the experiment evaluation was selected 3170 sequences with the same length and which give 95% of recognition accuracy.

Datasets were divided into train, validation and test sets and preliminary preprocessed. From preprocessed samples there were extracted features $f_{\Delta x}$, $f_{\Delta y}$ and f_{upd} (Table I). After that, they are fed to the input of recognition BLSTM which consist of two hidden layers and 100 cells in every layer. The training was accelerated using sequence bucketing and data parallelization [19]. After recognition training is done we obtain the sequence of points marked according to an output of decoder (decoding results) which represent the f_{output} feature.

As inputs of segmentation BLSTM are four features: f_{output} which represents decoding results and $f_{\Delta x}$, $f_{\Delta y}$, f_{upd} obtained on feature extraction stage for the recognition BLSTM training.

The structure of trained segmentation neural network is next: an input layer, two forward layers, one layer which combines outputs of forward layers, two backward layers, one layer which combines outputs of backward layers and an output layer. Total segmentation network consists of 7 layers.

Trained segmentation BLSTM contains 14164 weights. Training data (HWRD dataset) has 6848 sequences and 931343 frames in sequences. The average ratio of frames per sequence is 136.00. Validation data (HWRD dataset) contains 1488 sequences, 226969 frames in sequences, the average ratio of frames per sequence is 152.53.

On 1st epoch of the training classification error was equal to 20.08%, cross entropy error was equal to 76.03%. The lowest classification error was obtained on 55 epoch and is equal to 0.48%, cross entropy error is 2.22%.

Fig. 6 illustrates an example of character segmentation by RNNLIB and proposed approach.

Train and validation errors are shown in Fig. 7.

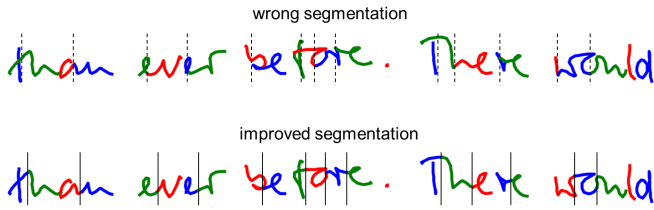


Fig. 6. Example of segmentation.

Here the first line shows segmentation results of text recognition system based on RNNLIB and the second one shows results of the proposed approach. The vertical line shows the beginning of a new segment for cases with wrong segmentation. From the figure follows that unlike the RNNLIB in proposed approach crossing characters were segmented more correctly. Such cases can be fixed using rule-based approach, but it is difficult to find all cases of wrong segmentation and to cover them by rules.

Character segmentation results are given in the Table III.

TABLE III
SEGMENTATION RESULTS

Dataset	Initial Segmentation Accuracy, %	Final Segmentation Accuracy, %	Delta, %
IAMonDO	91.35	98.75	7.4
HWRD	89.2	98.81	9.61

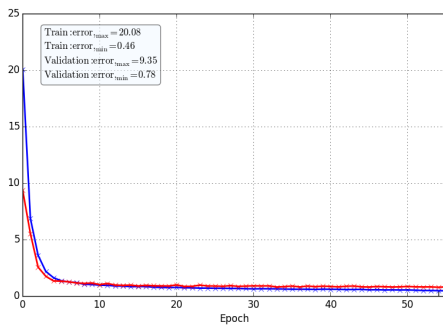


Fig. 7. Error curves.

V. CONCLUSION

In this paper, a new approach for character segmentation improvement of on-line handwritten text using recurrent neural networks and dynamic programming was proposed. Implementation of these methods to outputs of decoder gives improved segmentation results and allows to increase character segmentation accuracy after decoding without degradation of recognition results.

Experiments show that using of recurrent neural networks for classification segments gives an increase about 7% of

segmentation accuracy. Implementation of proposed dynamic programming approach gave 1% of performance improvement.

The proposed approach can be applied to the ink beautification problem, formulae recognition and for editing of handwritten text or other data represented as a sequence of points where a correct segmentation is needed.

REFERENCES

- [1] R. Plamondon and S. N. Srihari, "Online and off-line handwriting recognition: a comprehensive survey," in *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 1, pp. 63-84, 2000.
- [2] C. C. Tappert, C. Y. Suen, T. Wakahara, "The state of the art in online handwriting recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 12, no. 8, pp. 787-808, 1990.
- [3] P. Y. Simard, D. Steinkraus and M. Agrawala, "Ink normalization and beautification," *Eighth International Conference on Document Analysis and Recognition (ICDAR'05)*, vol. 2, pp. 1182-1187, 2005.
- [4] R.G. Casey and E. Lecolinet, "A Survey of Method and Strategies in Character Segmentation," *IEEE Trans. on PAMI*, vol. 18 (7), pp. 690-706, 1996.
- [5] C. T. Nguyen and M. Nakagawa, "An improved segmentation of online English handwritten text using recurrent neural networks," *2015 3rd IAPR Asian Conference on Pattern Recognition (ACPR)*, Kuala Lumpur, pp. 176-180, 2015.
- [6] S. P. Naeni, M. Khademi and A. Nikoogar, "A novel approach to segmentation of Persian cursive script using decision tree," *International Journal of Computer Theory and Engineering*, vol. 4 (3), p. 465, 2012.
- [7] I. Mayire, H. Askar and T. Dilmurat, "A Dynamic Programming Method for Segmentation of Online Cursive Uyghur Handwritten Words into Basic Recognizable Units," *Journal of Software*, vol. 10(8), pp. 2535-2540, 2013.
- [8] E. Kavallieratou, E. Stamatatos, N. Fakotakis and G. Kokkinakis, "Handwritten character segmentation using transformation-based learning," *Proceedings 15th International Conference on Pattern Recognition. ICPR-2000*, vol.2, pp. 634-637, 2000.
- [9] R. Ghosh, "Stroke segmentation of online handwritten word using the busy zone concept," *2013 International Conference on Soft Computing and Pattern Recognition (SoCPaR)*, pp. 54-59, 2013.
- [10] F. Naohiro, J. Tokuno and H. Ikeda, "Online character segmentation method for unconstrained handwriting strings using off-stroke features," *Tenth International Workshop on Frontiers in Handwriting Recognition, IWFHR-10*, pp. 361-366, 2006.
- [11] N. Bhattacharya and U. Pal, "Stroke segmentation and recognition from Bangla online handwritten text," *2012 International Conference on Frontiers in Handwriting Recognition*, pp. 740-745, 2012.
- [12] I. Mayire, H. Askar, T. Dilmurat, "A dynamic programming method for segmentation of online cursive Uyghur handwritten words into basic recognizable units," *Journal of Software*, vol. 8 (10), pp. 2535-2540, 2013.
- [13] A. Graves and J. Schmidhuber, "Framewise phoneme classification with bidirectional LSTM and other neural network architectures," *Neural Networks*, v.18 n.5-6, pp. 602-610, 2005.
- [14] A. Graves, M. Liwicki, S. Fernandez, R. Bertolami, H. Bunke and J. Schmidhuber, "A novel connectionist system for unconstrained handwriting recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31 (5), pp. 855-868, 2009.
- [15] A. Graves, "RNNLIB: A recurrent neural network library for sequence learning problems," <http://sourceforge.net/projects/rnnl/>, 2013.
- [16] E. Indermhle, M. Liwicki and H. Bunke, "IAMonDo-database: an online handwritten document database with non-uniform contents," *In Proc. Of Int. Workshop on Document Analysis Systems*, pp. 97-104, 2010.
- [17] S. J. Young, N. H. Russell, and J. H. S. Thornton, "Token passing: A simple conceptual model for connected speech recognition systems," *Tech. Rep. CUED/F-INFENG/TR38*, Cambridge University Engineering Department, 1989.
- [18] F. Chunsheng, "From Dynamic Time Warping (DTW) to Hidden Markov Model (HMM) Final project report for ECE742 Stochastic Decision," 2009.
- [19] V. Khomenko, O. Shyshkov, O. Radyvonenko and K. Bokhan, "Accelerating recurrent neural network training using sequence bucketing and multi-GPU data parallelization," *Proceedings of the 2016 IEEE First International Conference on Data Stream Mining & Processing*, pp. 100-103, 2016.

The Method of Cliodinamik Monitoring

Sergiy Golub

*Department of Intelligent Decision Making Systems
Cherkasy National University named after Bohdan Khmelnytsky
Cherkasy, Ukraine
fpkpkp@ukr.net*

Nataliia Khymytsia

*Social Communication and Information Science Department
Lviv Polytechnic National University
Lviv, Ukraine
nhymytsa@gmail.com*

Abstract — This article describes the process of using information technology of multi-level monitoring in the research of the similarity of historical periods. The results of inductive modeling are used to explain historical processes. The methods of clusterization and expert evaluation of the obtained results are combined. The results of experimental research of revealing similar periods of the history of Ukraine are presented.

Keywords— *history, monitoring information systems, modeling, historical processes, cliodinamik studies.*

I. INTRODUCTION

Monitoring is a technology for providing information about decision-making processes by providing continuous monitoring and processing of their results. The processing of the results of the observation is carried out in order to identify the properties of the objects and processes that are subject to research.

In accordance with the methodology for the establishment of multi-level monitoring information systems (MLM) [1], the processing of monitoring results is carried out through the consistent application of statistical processing methods, inductive and other methods for synthesizing models that solve local data transformation tasks at each of the levels of monitoring. Stratification of the structure of the multi-level monitoring information systems allow solving tasks of identifying functional dependencies, classifying, forecasting and some others by constructing multilayered models [2]. The hierarchical combination of these models form a global functional dependence (GFD) [3]. For the synthesis of multilayered models that form the structure of GFD, the basic algorithms of the GMDH [4], genetic and other evolutionary methods of synthesis of models, neural networks with various topologies, and hybrid algorithms, which are formed by combining several methods into a single process of designing the algorithm of model synthesis is used. Today MLM is used in many subject areas. In particular, information socioecological [5], sociogeogenic [6], medical monitoring [7] have already been created. Monitoring systems are used in public administration [8], in pedagogy [9, 10].

The variety of tools of historical science [11, 12] opens up wide opportunities for using information technology of multilevel monitoring based on inductive modeling methods.

Application of such methods allows realizing the prognostic function of historical science, which aims at predicting ways of development of historical processes, various variants of historical events. The prognostic function answers the question: what the historical reality will be and

when certain events will take place. The result of this function is, above all, the hypothesis, historical forecasting, which is based on objective scientific data.

II. ANALYSIS OF RESEARCH AND PUBLICATIONS

In practice, mathematical modeling in historical science has been used for more than 40 years. I. D. Kovalchenko, the founder of the Klimometric School in the USSR and the co-chairman of the International Commission on Quantitative History (INTERQUANT) widely applied mathematical modeling [13]. The scientist believed that this method makes it possible to analyze the historical process that could be implemented, but by virtue of some reasons has not happened. An American economist, one of the founders of the cliometriya and Nobel Laureate of the Year 1993, Robert William Fogel applied counter-factual modeling and proved that small innovations in industry contribute more to its evolution than large-scale technological discoveries [14]. The scientist conducted a fundamental study on what would be the US transport system if the railways were not invented.

At the beginning of the twenty-first century, especially during periods of social chaos, when the predictability of the course of history is significantly weakened and the possibility of unexpected variants of development increases, the relevance of multi-parametric modeling as a study of historical alternatives increases. Within this approach, the principle of synergetics as an interdisciplinary science that deals with the study of the processes of self-organization and the emergence, support of the stability and decomposition of structures (systems) of different nature is realized. Synergetics allows the historian to determine in which conditions small influences cause a complex system of large-scale changes and avalanche-like processes.

A new direction of modeling in historical research is the study of nonlinear processes. professor L. Borodkin introduces the direction - the cliodinamik that study the models of unstable historical processes [15, 16]. Under the guidance of L. Borodkin, the project was released which is devoted to the analysis of alternatives to historical development in 1929, when the so-called "big change" began in the USSR (the transition from the new economic policy to the beginning of the course on industrialization and collectivization).

Thus, for the identification and research of historical laws on the basis of the analysis of long-term social processes methods of cliodinamik are used by various scientific schools [17].

It is proposed to use data processing tools in the information technology of multi-level monitoring of cliodinamik.

The purpose of this study is to develop a method of cliodinamik monitoring, which will provide the identification of similar historical periods by clustering the vectors of their numerical features.

III. USING THE TEMPLATE

The method of cliodinamik monitoring combines the use of processes of clustering historical periods with their numerical features and expert justification of clustering results. It involves the following steps:

1. The list of features that are significant for making decisions based on the results of monitoring of historical processes is determined.
2. An array of numerical characteristics is formed for meaningful signs. The point of observation is numerical characteristics within one year.
3. The clustering of monitoring points by modeling results is carried out.
4. The hypothesis about the similarity of historical processes that took place over the years that formed separate clusters is proposed.
5. Examination of hypotheses is carried out by their expert justification using historical research methods. If an expert way has succeeded in substantiating the given hypotheses, the historical periods included in one cluster are considered to be similar. If this does not work - the following research is conducted and conclusions about the similarity of historical periods are not announced.

At the initial stage, it is necessary to solve the problem of clustering historical periods, which are presented in the form of vectors of their numerical features. Moreover, each historical period is represented by one vector of numerical characteristics of these attributes. The list and limits of historical periods, as well as the list of characteristics, are determined by expert means. The results are presented as an array of data (1):

$$\begin{pmatrix} x_{11} & x_{12} & \dots & x_{1n} & y_{11} & y_{12} & \dots & y_{1m} \\ x_{21} & x_{22} & \dots & x_{2n} & y_{21} & y_{22} & \dots & y_{2m} \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ x_{k1} & x_{k2} & \dots & x_{kn} & y_{k1} & y_{k2} & \dots & y_{km} \end{pmatrix} \quad (1)$$

where x_{ij} is the j -th independent index of the i -th historical period, y_{ij} is the j -th dependent index of the i -th historical period, k is the number of vectors; n - the number of indicators, m - the number of historical periods.

It is necessary to create a method of clustering historical periods, which allows you to obtain an array of input data in the form of a matrix (2):

$$\begin{pmatrix} x_{11} & x_{12} & \dots & x_{1n} & y_1 \\ \dots & \dots & \dots & \dots & y_2 \\ x_{k1} & x_{k2} & \dots & x_{kn} & y_k \end{pmatrix}, \quad (2)$$

which contains the sequence of vectors of the indices (X_i, y_j), $X_i \times X_k, i = 1, k$, where $X_i = \{x_{i1}, x_{i2}, \dots, x_{in}\}$ is the array of independent indicators of the i -th historical period

It is necessary to develop a method of clustering historical periods in the form of a function

$$a : (X_i, y_i) \rightarrow r, \quad (3)$$

where r is the cluster number that forms the set of clusters R .

The process of clusterization for each vector of the signs of the historical period sets the cluster number $r \in R$. In this case, the power of the set R is not known in advance.

Table I lists the features of the historical period that make up the vector.

TABLE I. LIST OF FEATURES.

Feature	Variable	Feature	Variable	Feature	Variable
Time of observation	x_1	Unemployed population (according to ILO methodology) at working age, ths	x_8	Budget (consolidated) UAH, billions. Income	x_{15}
Income population, billion UAH.	x_2	Economically inactive population of working age, ths	x_9	Budget (consolidated) UAH, billions. Expenditures	x_{16}
Average monthly salary, UAH	x_3	The level of economic activity, the percentage of the total population of the corresponding age group of working age	x_{10}	Financial result of ordinary activity before taxation, UAH billions.	x_{17}
The average monthly retirement pension, UAH	x_4	Number of births, ths.	x_{11}	Investments in fixed assets *, UAH, billions.	x_{18}
The average size of assistance to low-income groups of population per capita, monetary, UAH	x_5	Number of deceased, ths	x_{12}	Exports of goods and services, billion dollars. USA	x_{19}
The average size of assistance to low-income groups of population per capita, natural	x_6	Natural population growth, ths.	x_{13}	Imports of goods and services, billion dollars USA	x_{20}
Economically active population of able-bodied age, ths	x_7	Fixed assets *, UAH	x_{14}	Direct foreign investment in Ukraine (at the end of the year), mln. USA	x_{21}

To solve the clustering problem, the method proposed in [16] was used.

Socio-economic and demographic development of Ukraine was investigated during 1998-2012 based on the technology of multi-level monitoring.

The dependence of the gross domestic product of the country from the indicators presented in Table I was modeled. For synthesis of models, multi-row algorithm of GMDH [4] was used. A set of models was synthesized. After the tests, models were selected that met the criteria of accuracy, adequacy and stability, and the influence of the indicators from Table I that were included in the structure of these models was estimated. The impact of indicators was determined by the weighting factors calculated by the formula:

$$W_i = \frac{F'_{x_i}}{\sum_{i=1}^n F'_{x_i}}, \quad (4)$$

where F'_{x_i} is a partial derivative of the model on its i -th variable, n - number of indicators Table I that entered into the structure of the model.

The results of the study of these models are presented in Table II.

TABLE II. RESULTS OF CLUSTERIZATION

Indicator	Cluster 1 (2001-2012)		Cluster 2 (1999-2009)		Cluster 3 (1998-2002)	
	Weight coefficient, %	With indicator growth, GDP:	Weighty coef. %	With an increase in the indicator, GDP:	Weighty coef. %	With an increase in the indicator, GDP:
x ₁	99,88	Decreasing	0,00	unchanged	0,00	unchanged
x ₂	0,00	unchanged	20,60	decreasing	90,66	increasing
x ₃	0,06	increasing	19,50	increasing	0,00	unchanged
x ₄	0,00	unchanged	1,72	increasing	0,00	unchanged
x ₆	0,00	unchanged	22,86	decreasing	0,00	unchanged
x ₉	0,002	increasing	0,00	unchanged	0,00	unchanged
x ₁₄	0,00	unchanged	8,96	decreasing	0,00	unchanged
x ₁₅	0,00	unchanged	1,09	decreasing	0,00	unchanged
x ₁₆	0,00	unchanged	3,05	increasing	0,00	unchanged
x ₁₈	1•10 ⁻⁶	decreasing	20,75	increasing	0,00	unchanged
x ₂₀	0,06	decreasing	0,00	unchanged	0,00	unchanged
x ₂₁	0,00	unchanged	1,46	increasing	9,34	decreasing

Expert estimation of historical periods is as follows:

1 cluster of 2001-2012 : decrease of investments into fixed capital; strengthening the processes of industrialization; crisis in mechanical engineering, chemical industry; growth in agriculture, trade, transport; according to sources of

financing, bank loans and other loans play an essential role; negative balance of trade balance; lack of investment; the super-profits of the export industries fell into the pockets of the oligarchic circles in full; increase in the amount of social assistance; uncontrolled migration of the population; growth of demographic load.

2 clusters of 1999-2009 : the economic downturn, resulting in GDP declining to 40.8%; violation of macroeconomic equilibrium, correlation between wage level and gross income / mixed income in the structure of GDP, did not have the character of a stable trend; industry was the leading sector of the real economy, while retaining the largest share in its structure (in 2008, it accounted for 46.1% of the gross output of goods and services and 31.3% of gross value added); more than 2/3 of the total industrial output accounted for in the industry producing raw materials and energy resources; the share of products of social orientation is 1/5 of the total volume of industrial production. The light industry almost disappeared (1999: 1.6%, in 2008 - 0.9, in August 2009 - 0.8%); the machine-building industry even added in the rate of development during 1999-2007, but its share in the structure of industry through the crisis has fallen to 10.3%, which was 3-4 times lower than the level of developed countries.

3 clusters of 1998-2002 : growth of real incomes of citizens was 5-6 times higher than GDP; there was a rise in wages; the real GDP growth was 4.1-9%; recorded increase in production in industry; the growth rate of export of goods exceeded the growth rate of imports; increased competitiveness of Ukrainian goods; currency offer exceeded demand. Thus, we proved the closeness of the points included in each cluster. This means that during these years historical events in Ukraine were caused by close processes and research objects.

IV. CONCLUSIONS

A new method of cliodynamic research has been developed, which combines clustering and expert evaluation processes. Experimentally confirmed its effectiveness. The use of powerful means of synthesis of models in the cliometry allows us to draw conclusions based on the results of scientific research with formalized metrics. It helps to identify historical patterns and use them to justify future decisions, using the experience of past years with similar events.

In the difficult conditions of the economic crisis and social transformations, the value of qualitative historical knowledge is constantly increasing, as it allows us to explain the patterns of socio-economic events, identify trends, models of the future. It is necessary to develop a mechanism for the rational organization of social processes, taking into account the resources of political power, especially its components and forms, such as authority, control, influence, coercion, etc.

The simulation results obtained using the new method of cliodynamic monitoring should be used to identify the same types of periods of Ukrainian history (according to the mechanisms of influence on the economy that creates GDP).

REFERENCES

- [1] S. Golub, Multi-level modeling in environmental monitoring technologies. Cherkasy, Ed. from. Bohdan Khmelnytsky National University of Chernivtsi, 2007.

- [2] S. V. Golub, and I. V. Burley, "Multilayer data transformation in information systems of multi-level monitoring of fire safety," Collection of scientific works of Kharkiv University of Air Forces, Kharkiv University of Air Force named after Ivan Kozhedub, iss. 1 (38), pp. 246-251, 2014.
- [3] S. V. Golub, V. N. Rudnitsky, V. Yu. Dendarenko, and S. V. Pivneva, "Formation of the structure of the global function in information systems of multi-level monitoring of fire safety," The Science Vector of the Togliatti State University, no. 4 (22), p. 41-45, 2012.
- [4] A. G. Ivakhnenko, Inductive method of self-organization of models of complex systems. K: Naukova Dumka, 1981.
- [5] S. V. Golub, and P. O. Kolos, "Features of creation of information systems for crisis socioecological monitoring," Systems of information processing, iss. 4 (94), pp. 273-277, 2011.
- [6] S. V. Golub, and V. Yu. Nemchenko, "Adaptive formation of duplicate levels in the structure of hierarchical systems of multi-level sociohygienic monitoring," Inductive modeling of complex systems. Collection of scientific works. Editor V. Stepashko, Kyiv, International sciences center inform technologies and systems of NAS and MES of Ukraine, iss. 1, pp. 41-48, 2011.
- [7] S. V. Golub, and V. M. Dzholos, "Using of heuristic monitoring system for forecasting postinfarction complications," Visnyk of Zhytomyr State Technological University. Engineering, no. 4 (31), pp. 109-114, 2004.
- [8] S. V. Golub, N. Khymytsia, "The display of consolidated information of regional economic performance in the structure of multi-level models," Visnyk of Volodymyr Dahl East Ukrainian National University, no 8 (179), iss. 1, pp. 122-128, 2012.
- [9] S. V. Golub, and L. M. Semenenko, "Information modeling as the basis of the didactic system of training young scientists," Problems of general and pedagogical psychology. Collection of scientific works of the Institute of Psychology named after G.S. Kostyuk APS of Ukraine, Ed. S.D. Maximenko, vol. VII, iss. 7, 2005, p. 96 -100.
- [10] S. V. Golub, L. M. Semenenko, "Application of heuristic observation systems as a modeling pedagogical tool," Vesnyk Kharkiv National Automobile and Road University. Collection of scientific works, Kharkiv, KhNADU, iss.31, pp. 10-12, 2005.
- [11] S. Williamson, "Cliometrics history in the United States," (translation A. N. Polevaia), Economic History Review, Ed. V. I. Bovykin and L. I. Borodkin, vol. 1, 1996, pp. 75-107.
- [12] N. Khymytsia, S. Lisina, O. Morushko, and P. Zhezhnych, "Analysis of Computer-based Methods for Processing Historical Information," Advances in Intelligent Systems and Computing: Selected Papers from the International Conference on Computer Science and Information Technologies, CSIT 2017, September 5-9 Lviv, Ukraine, Shakhovska N. (Ed.), Springer International Publishing: 2017, vol. 1, pp. 365-367.
- [13] I. D. Kovalchenko, "Methods of historical research", (second edition), M: Nauka, 2003.
- [14] R. W. Fogel, Railroads and American Economic Growth: Essays in Econometric History. Baltimore: Johns Hopkins Press, 1964.
- [15] L. I. Borodkin, "Historical information science at the bifurcation point: the movement to Historical Information Science. Circle of ideas: algorithms and technologies of historical Informatics," IX conference of the Association "History and Computer", Under the editors L. I. Borodkin, V. N. Vladimirov, Moscow-Barnaul, 2005, pp.7-21.
- [16] L. I. Borodkin, Modeling of historical processes: from the reconstruction of reality to the analysis of alternatives. St. Petersburg: Aletya, 2016.
- [17] S. V. Golub, and I. V. Burlay, "Swear Increasing the efficiency of clustering by the results of simulation in the information technology of operational fire monitoring," Information Processing Systems, Kharkiv, iss. 2 (118), pp. 253-257, 2014.

CoLiTec Software for the Astronomical Data Sets Processing

Sergii Khlamov
*Laboratory of astrometry
Institute of Astronomy,
Kharkiv National University*
Kharkiv, Ukraine
sergii.khlamov@gmail.com

Artem Pohorelov
*Computer engineering and
management
Kharkiv National University of Radio
Electronics*
Kharkiv, Ukraine
artempogorelov@gmail.com

Vadym Savanevych
*Western Radio Technical Surveillance
Center
State Space Agency of Ukraine*
Mukachevo, Ukraine
vadym@savanevych.com

Vladimir Vlasenko
*Western Radio Technical Surveillance
Center
State Space Agency of Ukraine*
Mukachevo, Ukraine
vlasenko.vp@gmail.com

Olexander Briukhovetskyi
*Western Radio Technical Surveillance
Center
State Space Agency of Ukraine*
Mukachevo, Ukraine
izumsasha@gmail.com

Eugen Dikov
*Scientific Research
Design and Technology Institute of
Micrographs*
Kharkiv, Ukraine
endikov@gmail.com

Abstract—Nowadays, quick technological progress provokes creation of a big amount of the information that can be fed in different forms. There are a lot of different fields of science that use high dimensional data sets to analyze them in their researching. So, we need the data pre-processing methods and data reduction models to simplify input data sets by reducing unnecessary information. The paper deals with an approach of CoLiTec (Collection Light Technology) software to process in automated mode the different types of astronomical information which is fed online in the form of data sets or streams. Also the benefits of an usage of the OnLine Data Analysis System (OLDAS) was described. OLDAS helps with solving of the following Data Mining problems, such as clustering, classification and identification.

Keywords—data, set, stream, series of frames, processing, CoLiTec, OLDAS.

I. INTRODUCTION

The 21st century is closely connected with a huge revolution that characterized by tremendous technological progress. This progress provokes appearing of a big amount of different data sets, streams, but data growth is ahead of computing abilities of the existed machines. That's why it is very important to optimize data stream processing by using only necessary input data to allow improving the computing abilities of machines. So, data mining has become under the interest that has attracted a huge number of researches and experimentation to improve efficiency and productivity in different fields of interest.

What is the Data Mining? Data mining can be defined as a process during which previously unknown, nontrivial and hidden information will be collected in a larger dataset [1]. The data mining goal is extraction of the potentially useful information from the given large input online stream or data sets using necessary relationships, associations or patterns within the data, and transformation received data into subsets with understandable structure. These formed subsets will be used for the effective analysis and using in the future.

II. PROBLEM STATEMENT

The new networks of automated ground- and space-based observation systems and new surveys projects lead to a fast growing of the astronomical data sets. These sets can be fed in different forms, e.g. files stream, video stream, physical data saved on the different servers. For example, the Panoramic Survey Telescope and Rapid Response System (Pan-STARRS) [2] currently contains two telescopes with 1.8-m aperture that located at Haleakala in Hawaii (Fig. 1).



Fig. 1. Panoramic Survey Telescope and Rapid Response System

Each telescope has a field of view that equals to 3-degree. Both of them are equipped with the largest CCD-camera, which records about 1.400 millions of pixels per image. Each image requires about 2 gigabytes of storage and exposure times will be up to one minute. Also the time for computer processing (saving image to the storage) is equal to one minute or even more. Since images are taken on a continuous basis as an online data stream, more than 10 Terabytes of data are obtained every night.

Also the Large Synoptic Survey Telescope (LSST) [3] currently is under construction. It is a wide-field survey with reflecting telescope, which has a primary mirror with diameter 8.4 meters. The design of a telescope includes three mirrors. Both of them have very wide field of view that equals to 3.5-degree. Telescope uses a CCD-camera with resolution of 3.2-gigapixel (Fig. 2).

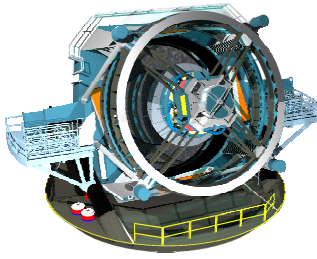


Fig. 2. Large Synoptic Survey Telescope (LSST)

LSST will take images of the full sky every few nights. There will be about 200 thousands of uncompressed images per year that equals to 1.28 petabytes. The managing and effective data mining and processing of the received from telescope data sets will be a very difficult part of the LSST project. Approximate requirements for the servers are about one hundred teraflops of power and about 15 petabytes of storage.

Such a big amount of the observational data requires a big volume of the hard disks or clusters. All these observational data can't be collected on a single universal local server. So, the disturbed systems as well as clouds technologies should be used for this goal, but how to process the big and high dimensional data streams in an efficient way?

III. TASK SOLUTION

The problems with classification and identification the data sets are related to Data Mining problems. They can be resolved in semi-automatic or automatic mode. Before the applying of data mining algorithms, intelligent methods to formed useful data, the pre-processing of it is recommended or even required. Because data mining can accidentally causes the misunderstanding when producing the significant results that cannot be used in the future to predict behavior and cannot be reproduced. Often, such results can appear during investigating of too many hypotheses and performing not properly testing of statistical hypotheses.

That's why the correct data selection is a necessary pre-processing step in the analyzing high-dimensional data sets or streams. It is considered to be a complex and time-consuming problem. So, the main goal of optimization of the data stream processing is an enhance the accuracy of classification and reducing the execution time.

During pre-processing step the anomaly detection and data cleaning are performed. Anomaly detection allows the identification of the records with an unusual data. Some of them except data errors can be interesting for the further investigation. Data cleaning removes noisy, irrelevant, redundant or missing data.

After pre-processing step the remaining useful information in data set will be categorized into clusters, based on the specific attributes. Clustering therefore breaks down data sets into subsets, whereby the different elements are assigned to the appropriate groups while the similar data are grouped together. Then created subsets will be classified by applying known structure to the new data.

Approach of CoLiTec (Collection Light Technology) software (<http://neoastrsoft.com>) [4-6] can resolve data

mining problems by using of the OnLine Data Analysis System (OLDAS) that covers different scientific and technological fields.

The main purpose of it is to bring together astronomical observation results of ground- and space-based observation systems, provide astronomers with the full-scope instruments for accessing and analyzing of the collected data.

IV. COLITec SOFTWARE

Software for data sets processing in automated mode is necessary for the most effective astronomical observations. This data set can be a survey given as series of frames. Modern astronomical systems and telescopes, e.g. Panoramic Survey Telescope and Rapid Response System (Pan-STARRS) [2] and Large Synoptic Survey Telescope (LSST) [3], allow taking a lot of frames of considerable sky area in one night.

This is a big amount of the raw data sets that should be processed. The main goal of this can be achieved using the Data Mining approach.

This approach is provided by CoLiTec software [5, 6] that allows processing of the input data sets or streams in real time. The visual confirmation of results after processing is also available.

With help of the CoLiTec software you can process the observation data that is continuously formed during observation (online stream). The processing pipeline includes corrupted data rejection, brightness equalization of frames, astrometry, photometry, detection of the moving objects and others.

The CoLiTec software guarantees not only a high efficiency of data sets processing but also a high accuracy of the data measurements [4, 5, 7]. We provided the comparison of statistical characteristics of positional CCD-measurements between CoLiTec and Astrometrica [8] software with the same set of CCD-frames. This comparison demonstrated that the limits for reliable positional CCD-measurements with CoLiTec software are wider than those with Astrometrica one for the area of extremely low signal-to-noise ratio [9].

The On-line Data Analysis System (OLDAS) is a very important part of the CoLiTec software. Using OLDAS you can process the data sets and streams as soon as they are successfully saved on the storage or uploaded to the server. This approach allows speeding up of the processing with preventing the collision. Also it provides the immediate notification about the emerging issues for user.

Also OLDAS provides ability to process the Big Data in real-time. For example, the data set that includes frames can be used for real-time photometry. The result of processing can be represented as light curves of the investigated variable stars. These light curves will be created and visualized on our server.

Some another CoLiTec software features are the following: intraframe processing (estimation of the objects position, astrometry and photometry reduction), interframe processing (detection of the moving objects and trajectories) and confirmation of the most interesting objects at the night of their preliminary discovery.

According to Data Mining approach CoLiTec software performs the following.

A. Pre-processing

During pre-processing step CoLiTec software in OLDAS mode starts processing with the input data set or stream as soon as they successfully received. These raw data will be moderated before using in computing process. At this stage unsupported and corrupted frames will be rejected. Only useful information from data set will be used in computing process.

B. Clustering

The remaining useful information in data set will be categorized into clusters with help of specified attributes. CoLiTec software uses the following attributes: equatorial coordinates, telescope, filter type, object under investigating and others. According to these attributes the appropriate input data from set will be separated into subsets with similar data.

C. Classification

Created after clustering subsets of data will be classified by applying known astronomical structure of the raw data that specified in Flexible Image Transport System (FITS) standard by NASA [10]. FITS is the most commonly used digital file format in astronomy. It is designed specifically for scientific data and includes as well as various astrometric, photometric or calibration information as the image metadata. After input data set classification the FITS files are sent to the processing pipeline.

D. Identification

While processing pipeline starts receiving of the classified FITS files it identify types of them. For example, is this FITS file a raw light frame or maybe it is service master-frame that is used in frame's calibration (e.g. bias, dark, darkflat, flat). If this is a raw light frame the processing pipeline starts computing process.

E. Processing

Computing process consists of two stages: intraframe and interframe processing. Intraframe processing is designed to estimate the position of all objects (stars, galaxies, asteroids, comets) in the frame at current moments. Also calibration, background alignment and brightness equalization are performed at this stage (Fig. 3). In OLDAS mode the brightness equalization process has the following workflow:

- online searching for the frames in specified directories;
- searching for the required additional frames (bias, dark, darkflat, flat) if they are not specified;
- creating of the appropriate master-frames;
- applying the inverse median filter.

Interframe processing is used to detect and estimate moving objects trajectories. The core of CoLiTec software based on the preliminary objects detection by the accumulation of statistics that performed by multi-valued transformation of the objects coordinates that corresponds to Hough space-time transformation.

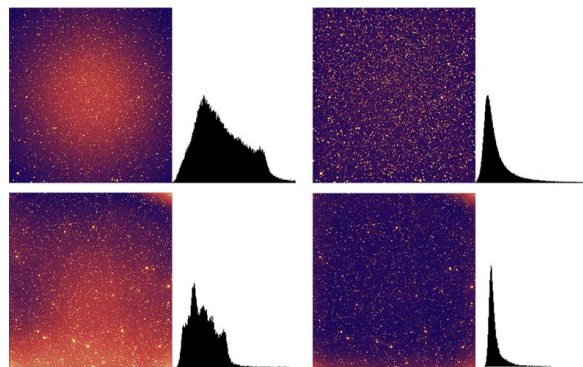


Fig. 3. Brightness equalization with histograms (before and after).

CoLiTec has abilities for detecting very slow, very fast objects and objects with near-zero apparent motion [5-7]. Range of velocities for detection is from 0.8 to 40 pixels per frame. For example, the fastest NEO is K12C29D (40 pixels per frame) or the object with near-zero apparent motion is ISON C/2012 S1 comet (0.8 pixels per frame) [11].

Also CoLiTec software has the following features: automated detection of the faint moving objects ($SNR > 2.5$); working with the huge field of view (< 10 degrees); automated calibration and brightness equalization; automated astrometric and photometric reduction; deciding system of the processing results allows adapting the user settings and inform user about correct results at the each stage of processing; automated objects rejection with bad or unclear observations; multi-threaded processing support; multi-cores systems support with the ability to manage individual treatment processes.

CoLiTec software has the system for monitoring processing messages with a detailed logging of handling process and tracking system of all running modules allows correct managing and terminating processes at any stage without data losses.

Also CoLiTec software includes pipeline for digital video processing. It is presented in form of the flexible platform for receiving and processing video in any resolution. Also the pipeline allows an easy integration of the different modules for improving the image quality and detection of the moving objects.

F. Summarization

After pipeline processing CoLiTec provides the different forms of data set representation, including results visualization and generation of report to different services.

CoLiTec software equipped with the modern viewer of processing results LookSky with a user-friendly GUI (Fig. 4).

LookSky viewer can be run without the main program. It can be used for independent review of the processing results by CoLiTec during data processing of the main program. With help of blinking method human can't carefully analyze all interesting objects in the frame.

The particularly serious difficulty for this is the frames analyzing from the wide-field telescopes with a huge aperture. Because about several tens of asteroids with slight shine can be present simultaneously in telescope's field of

view. That's why automatic series of frames processing is necessary to the modern astronomer.

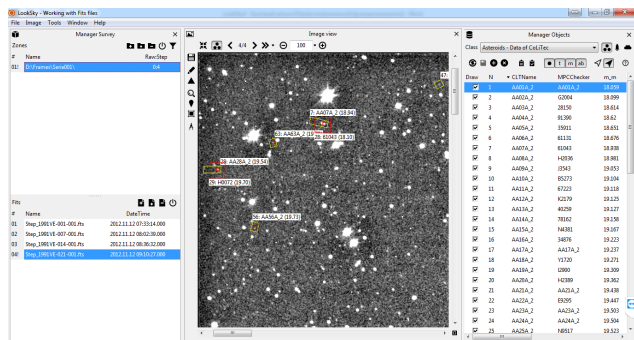


Fig. 4. LookSky viewer of processing results.

V. CONCLUSIONS

The quick technological progress, new surveys projects and networks of automated ground- and space-based observation systems lead to the fast growing of astronomical data sets that can be provided in the different form and volume. A huge amount of the raw data sets are needed to be processed, but trend of the data growth is ahead of computing abilities of the existed machines. So, we suggest using Data Mining approach for all goals that connected to the processing of a big amount of data. As described in the paper the Data Mining approach is very useful for the optimization of data stream processing. It allows using only necessary input data to improve the computing abilities of machines. The good example of using the data mining principles and stages of processing (anomaly detection, clustering, classification, identification, and summarization) is the CoLiTec software [4-6].

With help of the CoLiTec software more than 1,500 asteroids were preliminary discovered, including 5 NEO, 21 Trojan asteroids of Jupiter and 1 Centaur. CoLiTec software was used in about 700,000 observations, during which four comets (C/2010 X1 (Elenin), P/2011 NO1 (Elenin), C/2012 S1 (ISON) [11], P/2013 V3 (Nevski)) were discovered.

ACKNOWLEDGMENT

The authors thank observatories and institutes that have implemented CoLiTec software for observations. We express our gratitude to Mr. W. Thuillot, coordinator of the Gaia-FUN-SSO network [12], for the approval of CoLiTec as well-adapted software for the frames processing for all Gaia-FUN-SSO members (<https://gaiafunso.imcce.fr>).

CoLiTec software is available at the following website: <http://neoastrosoft.com>.

REFERENCES

- [1] D. Peralta, S. del Rio, S. Ramirez-Gallego, I. Triguero, J. Benitez, and F. Herrera, "Evolutionary feature selection for big data classification: A map reduce approach," *Mathematical Problems in Engineering*, vol. 2015, Article ID 246139, pages 11, 2015.
- [2] L. Denneau, R. Jedicke, Tommy Grav, Mikael Granvik, Jeremy Kubica, and Andrea Milani, "The Pan-STARRS Moving Object Processing System," *Publications of the Astronomical Society of the Pacific*, vol. 125, pp. 357-395, 2013.
- [3] M. T. Tuell, H. M. Martin, J. H. Burge, W. J. Gressler, and C. Zhao, "Optical testing of the LSST combined primary/tertiary mirror," *Proc. SPIE 7739, Modern Technologies in Space- and Ground-based Telescopes and Instrumentation*, 77392V, 23 July 2010.
- [4] V. E. Savanevych, O. B. Briukhovetskiy, N. S. Sokovikova, M. M. Bezdrovny, I. B. Vavilova, Yu. M. Ivashchenko, L. V. Elenin, S. V. Khlamov, Ia. S. Movsesian, A. M. Dashkova, and A. V. Pogorelov, "A new method based on the subpixel Gaussian model for accurate estimation of asteroid coordinates," *Monthly Notices of the Royal Astronomical Society*, vol. 451 (3), pp. 3287-3298, 2015.
- [5] V. E. Savanevych, S. V. Khlamov, I. B. Vavilova, A. B. Briukhovetskiy, A. V. Pohorelov, D. E. Mkrtychian, V. I. Kudak, L. K. Pakuliak, E. N. Dikov, R. G. Melnik, V. P. Vlasenko, and D. E. Reichart, "A method of immediate detection of objects with a near-zero apparent motion in series of CCD-frames," *A & A, Worldwide astronomical and astrophysical research*, vol. 609, A54, pages 11, 2018.
- [6] S. Khlamov, V. Savanevych, O. Briukhovetskiy, and A. Pohorelov, "CoLiTec software – detection of the near-zero apparent motion," *Proceedings of the International Astronomical Union: Cambridge University Press*, vol. 12(S325), pp. 349-352, 2017.
- [7] S. V. Khlamov, V. E. Savanevych, O. B. Briukhovetskiy, and S. S. Oryshych, "Development of computational method for detection of the object's near-zero apparent motion on the series of CCD-frames," *Eastern-European Journal of Enterprise Technologies*, vol. 2, iss. 9 (80), pp. 41-48, 2016.
- [8] H. Raab, "Astrometrica: Astrometric data reduction of CCD images" 2012, *Astrophysics Source Code Library*, record ascl:1203.012.
- [9] V. E. Savanevych, A. B. Briukhovetskiy, Yu. N. Ivashchenko, I. B. Vavilova, M. M. Bezdrovny, E. N. Dikov, V. P. Vlasenko, N. S. Sokovikova, Ia. S. Movsesian, N. Yu. Dikhtyar, L. V. Elenin, A. V. Pohorelov, and S. V. Khlamov, "Comparative analysis of the positional accuracy of CCD measurements of small bodies in the solar system software CoLiTec and Astrometrica," *Kinematics and Physics of Celestial Bodies*, vol. 31 (6), pp. 302-313, 2015.
- [10] D. C. Wells, E. W. Greisen, and R. H. Harten, "FITS - a Flexible Image Transport System", *A & AS*, vol. 44, p. 363, 1981.
- [11] Minor Planet Center, COMET C/2012 S1 (ISON). Available at: <http://www.minorplanetcenter.org/mpec/K12/K12S63.html>.
- [12] W. Thuillot, B. Carry, J. Berthier, P. David, D. Hestroffer, and P. Rocher, "Gaia-FUN-SSO: a network for ground-based follow-up observations of Solar System Objects," *Proceedings of the Annual meeting of the French Society of Astronomy and Astrophysics, SF2A-2014*, pp. 445-449, , 2014.

Piecewise-Linear Approach to Classification Based on Geometrical Transformation Model for Imbalanced Dataset

Anastasiya Doroshenko
ACS Department
Lviv Polytechnic National University,
Lviv, Ukraine
anastasia.doroshenko@gmail.com

Abstract— The article describes the method of cost-sensitive classification for imbalanced dataset based on neural-like structure of successive geometric transformations model using piecewise-linear approach to classification. The proposed method characterized by high learning speed and accuracy of classification.

Keywords—data mining, classification, imbalanced data, neural-like structure of successive geometric transformations model, NLS SGTm, cost sensitive classification.

I. INTRODUCTION

Most of the working companies today have a large amount of accumulated data with information about customers, sales, orders, and more. Such data is a source of hidden knowledge, the ownership of which can provide the company with further growth, profits and customers. These tasks are mainly formulated as Data Mining tasks and one of the most popular among them is the task of classification. These tasks are formulated daily, in such spheres of life as how target marketing, medicine, telecommunication, insurance, chemical industry, bioinformatics and others. Researchers use different methods to solve classification problems. The most effective today are classification by decision tree induction, Bayesian classification, neural networks, support vector machines (SVM), deep learning methods [1-5].

The main requirements for the data mining methods, besides the high accuracy of the classification, are their speed and the ability to process huge amounts of data. One of the methods that well matched to these requirements is neural-like structure of successive geometric transformations model (NLS SGTm) [9,13,14].

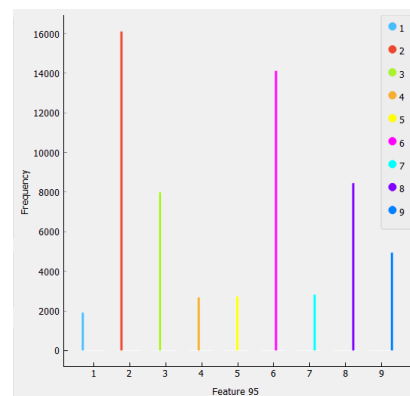
II. STATEMENT OF THE PROBLEME

When solving the classification problem, there are often additional restrictions caused by the peculiarities of the subject area or specific task.

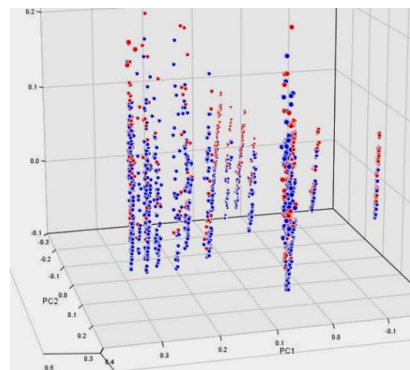
In addition, precisely when solving the classification problem for tasks that describe some kind of business or social processes, there are problems such as the imbalanced presentation of data and the different weight of classification errors.

A. Imbalanced representation of data

Imbalanced representation of data is a feature characteristic of such data mining task as classification. The number of instances of one class may differ by an order of magnitude from the number of instances of other classes, which greatly complicates the process of classification. In addition, the density of the distribution of instances of different classes in the space of signs may also be different (Fig.1). According to the hypothesis of compactness, it is assumed that objects belonging to the same class form certain clusters in the space of signs and can easily be separated by hyperplanes of a simple form. However, in many cases, in the systems under conditions of uncertainty, there is a mutual overlap of classes [1, 11-13].



a



b

Fig. 1. Data representation of the data mining task a) the frequency diagram that displays the number of objects in each class (for 9 classes) b) in the coordinates of three main components (for two classes)

B. A different weight of errors

Depending on the condition of the task to be solved, each type of error can have its own weight. Often such conditions exist for classification tasks. Accordingly, when solving the classification problem, it is necessary to consider not only the accuracy of the recognition of each of the classes, but also their interdependence - in order to ensure that the weight of the total number of errors was minimal.

In [6] described a comparative analysis of the solution of this problem using well-known methods such as Random Forest Learner, Logistic Regression, SVM and NLS SGTM. It has been demonstrated that NLS SGTM gives a highest accuracy of classification. This article suggests methods for further improving the accuracy of classification for this problem.

However, article [6] did not take into account the constraints caused by the subject area, namely the fact that various mistakes (TF or FT) are of different weight during the classification. These restrictions are formulated by the customer when setting the task depending on the priorities of their business (for example, depending on what is more important - to identify a potential fraudster or not to lose a potential client). In [6] this cost-sensitive problem was solved by using method of penalties and rewards. This article describes the combination of two methods: method of penalties and rewards and piecewise-linear approach to classification using NLS SGTM.

III. ARCHITECTURE AND LEARNING ALGORITHM OF NEURAL-LIKE STRUCTURE OF SUCCESSIVE GEOMETRIC TRANSFORMATIONS MODEL

A. Training NLS SGTM in supervisor mode

The supervisor mode is special because the components of the data vectors are divided into input and output, where the latter are known only for the elements of the training sample [12-14].

Accordingly to [9], the value of the coefficient $K_i^{(S)}$ required to execute the sequence of geometric transformations can be calculated only for the vectors of the training sample. In this case, it is sufficient to calculate the value of the coefficient $K_i^{(S)}$ based on

$$K_i^{(S)} = \frac{\sum_{j=1}^n (X_{i,j}^{(S)} \cdot X_{b,j}^{(S)})}{\sum_{j=1}^n (X_{b,j}^{(S)})^2} \quad (1)$$

where $X_{i,j}^{(S)}$ - elements of input vector of features, $X_{b,j}^{(S)}$ - elements of basic vector of features, n - the number of input components of the vector and represent the pseudo coefficient as the approximation of the dependence (2), which is given tabular

$$K_i^{(S)} = f(\overline{K_i^{(S)}}) \quad (2)$$

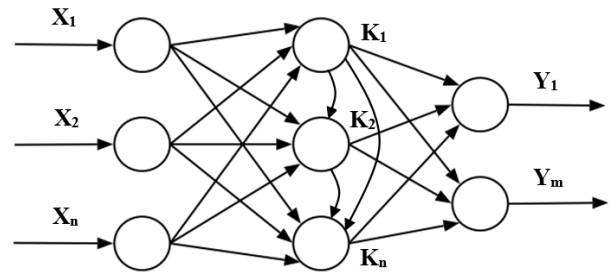


Fig. 2. The topology of NLS SGTM in a supervisor mode.

The topology of a NLS SGTM of this type is presented in Fig. 2, where linear or nonlinear neural elements can be used in the hidden layer, the function of which is the approximation of dependence (2).

For the case of a linear variant of the SNM, the function of activation of the neural elements has the form

$$K_i^{(S)} = \alpha \times \overline{K_i^{(S)}} \quad (3)$$

where α - the coefficient is calculated on the basis of the least squares criterion.

The use of neural-like structure of successive geometric transformations model in the classification mode involves the construction of a separating surface that provides the separation of given classes of objects. The construction of the hyperplane surface in the entire domain of the definition of the input variables has the following drawbacks and limitations:

- the number of vectors-implementations is too large for the simultaneous use of the entire training sample;
- the density of representation of the points of different classes in the realization space is approximately heterogeneous;
- separation of classes by a hyperplane does not take into account the essential nonlinearity of the tasks to be solved.

The use of linear classification methods for data mining is optimal in terms of the speed of solving the tasks, but they do not provide sufficient accuracy of the classification.

To solve this problem, a piecewise-linear approach for constructing separating surfaces based on a model of geometric transformations was developed. It allows for the nonlinearity of data mining tasks to be taken into account, but does not require a large amount of time to execute.

Another advantage of using the piecewise-linear approach of constructing separating surfaces is that by using it, by dividing the total sample into clusters, it is possible to process the entire sample for a reasonable time.

IV. DEVELOPMENT OF THE PIECEWISE-LINEAR APPROACH FOR THE CONSTRUCTION OF SEPARATING SURFACES BASED ON NEURAL-LIKE STRUCTURE OF SUCCESSIVE GEOMETRIC TRANSFORMATIONS MODEL

Let us consider the use of the piecewise-linear approach for constructing separating surfaces in combination with the

method of rewards and penalties for a problem with two classes.

It should be noted, that in order to evaluate the effectiveness of the method two data samples are necessary. The first it is the training dataset, which is used to training of NLS SGTM and for calculating coefficients $K_i^{(S)}$. The second dataset it is test sample whose data was not used during the construction of the model of geometric transformations, but which are necessary to assess the accuracy of the classification.

In general, objective of standard classification is to obtain highest accuracy of classification. However, for the data mining tasks, we have to organize cost-sensitive learning with formulated cost matrix, class probability estimate, misclassification costs and other types of costs involved in the learning process [10].

In order to take into account such a feature of Data Mining as a different weight of errors, we use the method of rewards and penalties. The matrix of rewards and penalties in this case will have the following form (Table I). As a rule, for each specific subject area and task, customer or business analytic forms such a matrix separately.

TABLE I. MATRIX OF REWARDS AND PENALTIES

Matrix of Rewards and Penalties	Values of Rewards and Penalties	
	The vector is recognized as class 0	The vector is recognized as class 1
The vector belongs to class 0	r_{11}	p_{12}
The vector belongs to class 1	p_{21}	r_{12}

To increase the accuracy of the solution to the classification problem, it is proposed to combine the use of the method of rewards and penalties [5,8] and the tree of division into classes (clustering).

Using the division tree into classes, we can combine data vectors with similar inputs into separate clusters and analyze them independently of each other (Fig.3). After receiving the penalty points for each of the clusters, they are summed up. This approach makes it possible significantly improve the overall accuracy of the classification.

Consider the method of combining the method of rewards and penalties and dichotomy to solve the classification problems in more detail.

The piecewise-linear approach to constructing separating surfaces for solving the problems of classification on the basis of NLS SGTM using the method of rewards and penalties [5]:

1. Initially, for NLS SGTM as a training set we used all training data. The accuracy of classification we estimated as amount of penalty points calculated according to the method of rewards and penalties.

2. After that, both training and test sets, are divided into 2 clusters: vectors recognized by the NLS SGTM as "class 1" and vectors recognized as "class 2".

3. After clustering into the training samples obtained for each cluster, we substitute the real values of the outputs and repeat from step 1 all action for both clusters separately.

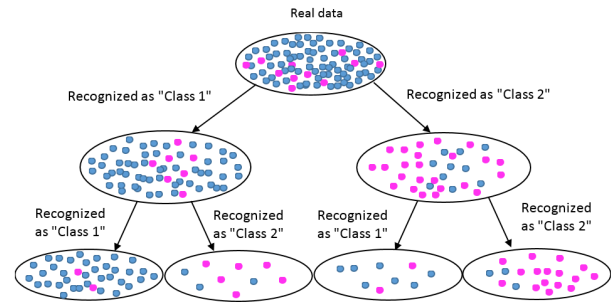


Fig. 3. The division tree into classes based on NLS SGTM

4. Repeat steps 1-3 until the amount of penalty points received for all lower-level clusters is not acceptable under the terms of the task. That is, the accuracy of the classification is sufficient.

5. When training of NLS SGTM for each cluster is finished and accuracy of the classification is sufficient we can use these networks for testing dataset.

So, due to the breakdown of data into clusters, we accelerate the implementation of the classification process, we can take into account a different weight of errors for a specific subject area and increase the accuracy of the classification.

V. EXPERIMENTAL RESULTS

This article describes the solving of classification task, which was formulated in [7]. The training sample describes the transactions carried out by credit card holders within two days and consists of 284,807 lines and 31 columns. For reasons of confidentiality, the dataset contains not original user data, but principal components obtained by the PCA method from the initial data (V_1, \dots, V_{28}). Only two features: 'Time' (the number of seconds passing through each transaction) and 'Amount' have not been transformed.

Also, the dataset contains one target feature 'Class', which shows the client's affiliation to one of two classes - frauds or ordinary clients. The main feature of the dataset is that the data set is highly unbalanced - only 492 transactions out of 284807 (0.172% of all transactions) have the value of the target field 1, that is, customers are fraudulent.

The dataset has been collected and analyzed during a research collaboration of Worldline and the Machine Learning Group (<http://mlg.ulb.ac.be>) of ULB (Université Libre de Bruxelles) on big data mining and fraud detection [7].

According to the subject area a matrix was formed. By analyzing this matrix, it can be seen that a properly classified vector that belongs to the "fraud" class has a much greater weight than a properly classified "ordinary client" vector. At the same time, the case where an ordinary customer is classified as a fraud has the highest number of penalty points (Table II) [8].

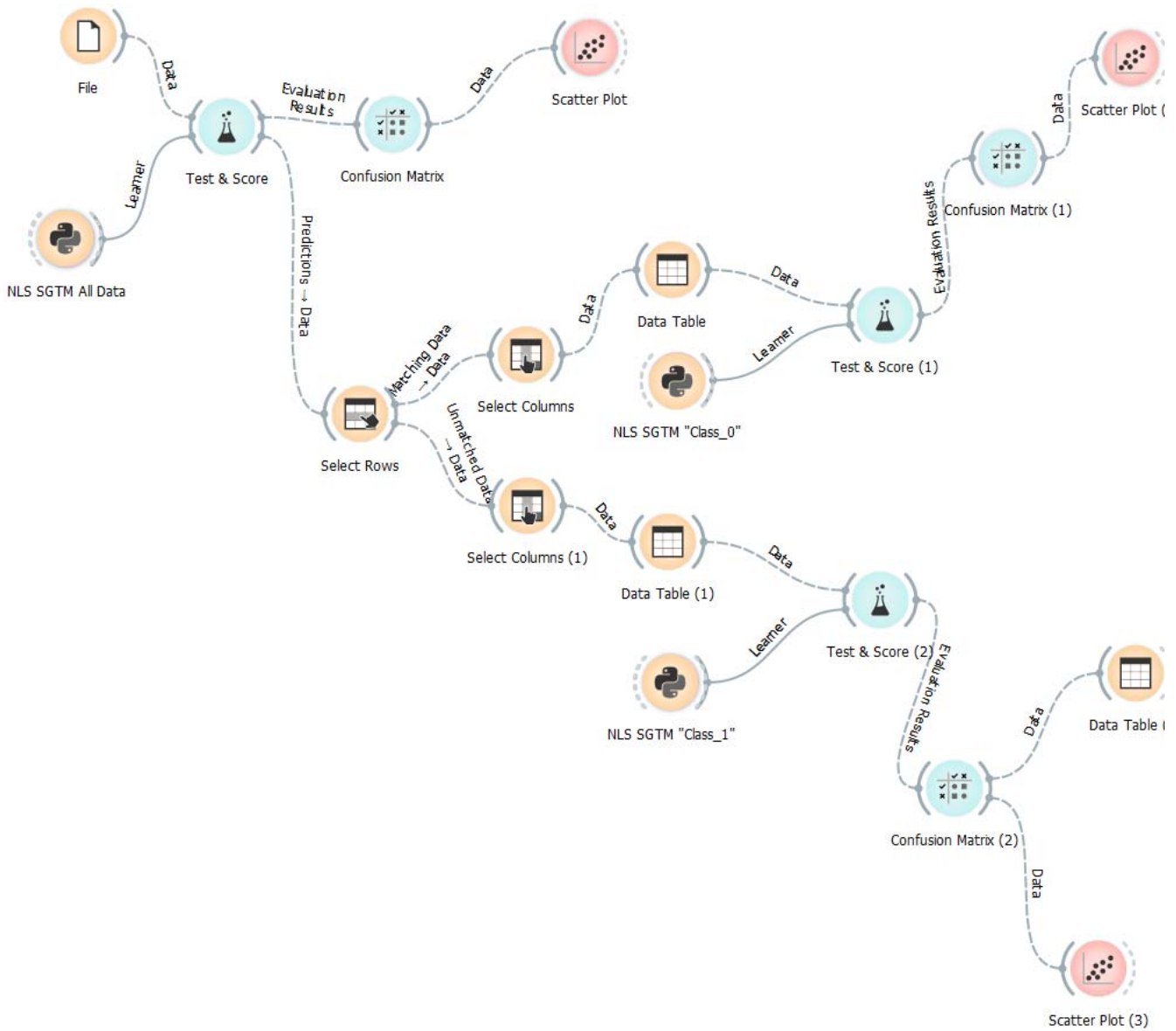


Fig. 4. The structure of the workflow in Orange

TABLE II. MATRIX OF REWARDS AND PENALTIES FOR SOLVING TASK

Matrix of Rewards and Penalties	Values of Rewards and Penalties	
	The vector is recognized as class 0	The vector is recognized as class 1
The vector belongs to class 0	1	-3
The vector belongs to class 1	-2	5

The structure of the workflow in Orange is shown in Fig.4.

Analyzing the results of the classification (Fig.5), we can see that despite the fact that for the sample as a whole the accuracy of the classification was very high in consequence of the application of the piecewise-linear approach accuracy further increased.

All Data		Predicted		
		1	2	Σ
Actual	1	284278	37	284315
	2	113	379	492
Σ		284391	416	284807

Cluster "1"		Predicted		
		1	2	Σ
Actual	1	284270	8	284278
	2	58	55	113
Σ		284328	63	284391

Cluster "2"		Predicted		
		1	2	Σ
Actual	1	21	16	37
	2	3	376	379
Σ		24	392	416

Fig. 5. Results of classification by clusters (in vectors)

All Data		Predicted		
		1	2	Σ
Actual	1	284278	-111	284167
	2	-226	1895	1669
Σ		284052	1784	285836

Cluster "1"		Predicted		
		1	2	Σ
Actual	1	284270	-24	284246
	2	-116	275	159
Σ		284154	251	284405

Cluster "2"		Predicted		
		1	2	Σ
Actual	1	21	-48	-27
	2	-6	1880	1874
Σ		15	1832	1847

Cluster "1" + Cluster "2" = 286252

Fig. 6. Results of classification by clusters according to the matrix of rewards and penalties (in points)

Fig. 6 shows that the sum of the points obtained for the correct classification by clustering is greater than for the initial sample. It is also worth noting that the initial data was very unbalanced and the number of instances of the lower class was only 0.17% of the total number of training vectors.

VI. CONCLUSIONS

The piecewise-linear approach to classification based on geometrical transformation model for imbalanced dataset is developed. This method realizes cost-sensitive classification for such tasks of data mining as fraud detection, home loan or credit applications, medical diagnostic, marketing research where some classes are rare but with big impact. The proposed method characterized by high learning speed and accuracy of classification.

REFERENCES

- [1] C. C. Aggarwal and P. S. Yu, "A Survey of Uncertain Data Algorithms and Applications," *IEEE Transactions on Knowledge and Data Engineering*, vol. 21, no. 5, pp. 609-623, May 2009.
- [2] A. J. Chamatkar and P. K. Butey, "Implementation of Different Data Mining Algorithms with Neural Network," 2015 International Conference on Computing Communication Control and Automation, Pune, pp. 374-378, 2015. doi: 10.1109/ICCUBEA.2015.78
- [3] D. Zhu, H. Jin, Y. Yang, D. Wu and W. Chen, "DeepFlow: Deep learning-based malware detection by mining Android application for abnormal usage of sensitive data," 2017 IEEE Symposium on Computers and Communications (ISCC), Heraklion, pp. 438-443, 2017. doi: 10.1109/ISCC.2017.8024568
- [4] Ye. Bodyanskiy, I. Perova, O. Vynokurova, and I. Izonin "Adaptive Wavelet Diagnostic Neuro-Fuzzy System for Biomedical Tasks," 14th International Conference on Advanced Trends in Radioelectronics, Telecommunications and Computer Engineering (TCSET), Lviv-Slavske, Ukraine, 299-303, February 20 – 24, 2018.
- [5] O. Riznik, I. Yurchak, E. Vdovenko and A. Korchagina, "Model of stegosystem images on the basis of pseudonoise codes," *VIIth International Conference on Perspective Technologies and Methods in MEMS Design*, Lviv, pp. 51-52, 2010.
- [6] R. Tkachenko, . Doroshenko, I. Izonin, Y. Tsymbal, and B. Havrysh, "Imbalance Data Classification via Neural-like Structures of Geometric Transformations Model: Local and Global Approaches," In: Hu, Z. B., Petoukhov, S., (eds) *Advances in Computer Science for Engineering and Education. ICCSEE2018. Advances in Intelligent Systems and Computing*. Springer, Cham, vol.754, pp.112-122, 2018. https://doi.org/10.1007/978-3-319-91008-6_12
- [7] A. D. Pozzolo, O. Caelen, R. A. Johnson and G. Bontempi, "Calibrating Probability with Undersampling for Unbalanced Classification," *IEEE Symposium Series on Computational Intelligence*, Cape Town, pp. 159-166, 2015. doi: 10.1109/SSCI.2015.33
- [8] J. Wang, P. Zhao and S. C. H. Hoi, "Cost-Sensitive Online Classification," *IEEE Transactions on Knowledge and Data Engineering*, vol. 26, no. 10, pp. 2425-2438, Oct. 2014.
- [9] R. Tkachenko and I. Izonin "Model and Principles for the Implementation of Neural-Like Structures based on Geometric Data Transformations", In: Hu, Z.B., Petoukhov, S., *Advances in Computer Science for Engineering and Education. ICCSEE2018. Advances in Intelligent Systems and Computing*. Springer, Cham (2018).
- [10] S. Ghosh, A. Ray, D. Yadav and B. M. Karan, "A Genetic Algorithm Based Clustering Approach for Piecewise Linearization of Nonlinear Functions," 2011 International Conference on Devices and Communications, Mesra, pp. 1-4, 2011.
- [11] H. He and E. A. Garcia, "Learning from Imbalanced Data," in *IEEE Transactions on Knowledge and Data Engineering*, vol. 21, no. 9, pp. 1263-1284, Sept. 2009. doi: 10.1109/TKDE.2008.239
- [12] R. Tkachenko, H. Cutucu, I. Izonin, A. Doroshenko, and Yu. Tsymbal "Non-iterative Neural-like Predictor for Solar Energy in Libya," In: Ermolayev, V., Suárez-Figueroa, M. C., Lawryniewicz, A., Palma, R., Yakovyna, V., Mayr, H. C., Nikitchenko, M., and Spivakovsky, A. (Eds.): *ICT in Education, Research and Industrial Applications. Proc. 14-th Int. Conf. ICTERI 2018. Volume I: Main Conference*. Kyiv, Ukraine, May 14-17, pp.35-45, 2018, CEUR-WS.org
- [13] U. Polishchuk, P. Tkachenko, R. Tkachenko and I. Yurchak, "Features of the auto-associative neurolike structures of the geometrical transformation machine (GTM)," 2009 5th International Conference on Perspective Technologies and Methods in MEMS Design, Zakarpattia, Ukraine, pp. 66-67, 2009.
- [14] R. Tkachenko, I. Yurchak and U. Polishchuk, "Neurolike networks on the basis of Geometrical Transformation Machine," 2008 International Conference on Perspective Technologies and Methods in MEMS Design, Polyana, Ukraine, pp. 77-80, 2008. doi: 10.1109/MEMSTECH.2008.4558743

Data Processing Methods for Mobile Indoor Navigation

Alexey Roienko
IT-Jim
Kharkiv, Ukraine
a.roienko@it-jim.com

Feliks Sirenko
IT-Jim
Kharkiv, Ukraine
sirenkofelix@gmail.com

Yevhen Chervoniak
IT-Jim
Kharkiv, Ukraine
eugenecher94@gmail.com

Ievgen Gorovyi
IT-Jim
Kharkiv, Ukraine
ceo@it-jim.com

Abstract—The competition on the world market of smartphones and tablets between the acknowledged leaders on one side and the numerous newcomers on the other makes them all look for new solutions that open additional opportunities for the developers and customers without the growth of the price. The progress with new opportunities turns possible when it happens simultaneously in the software and the hardware. The brightest one example of the above statement can be observed for the sensors of mobile devices. It is totally impossible to imagine modern smart devices having no sensors, as the progress of last decade (SLAM, face ID, OCR, pattern recognition etc.) was achieved thanks to considerable improvements of sensors and the algorithms for their processing. The paper addresses the questions of characteristics analysis of such mobile sensors as accelerometer, magnetometer and gyroscope from the point of view of their application in indoor navigation field. Signals of BLE beacons and their processing methods are investigated as well. The sensor fusion task is briefly discussed and several practical examples are given.

Keywords—*signal processing, sensor fusion, indoor navigation, BLE navigation, IMU navigation*

I. INTRODUCTION

Mobile sensors have been coming into our life more and more. Besides, being used widely in smartphones, they started a new branch in mobile industry – wearable fitness-trackers and smart watches. Almost all leading technology companies, like Apple, Samsung or Xiaomi produce such devices and integrate them into their infrastructure. Moreover, mobile sensors are the basis for such modern and greatly developing fields like Augmented Reality (AR) where they serve, for example, for improving the detection of markers on the scene, or Virtual Reality - for precise head pose tracking. Another popular and important scientific area of mobile sensors application is known as Simultaneous Localization and Mapping or SLAM. Finally, indoor navigation, which can be considered as GPS for indoor environments, relies greatly on smartphone inertial sensors [1] - [8]. Despite the fact that there are many different sensor types, some of them are more common and greatly applied in the mentioned applications. Namely, they are accelerometer (A) which measures the linear acceleration of a device, magnetometer (M) that detects the Earth magnetic field along three axis of device coordinate system, and gyroscope

(G) which measure angular velocity. The combination of these three sensors, known as Inertial Measurement Unit (IMU), is widely used for determination of device pose in global coordinate system. It is worth noting that mobile sensors can be made of different quality. Recently, the so-called Micro-Electro-Mechanical Systems (MEMS) sensors have been attracted great interest. They are widely applied in smartphones and different wearable devices. As it will be shown, they can provide rather noisy signals but their main benefit is their accessibility due to the low price.

Besides the IMU, Wi-Fi access points and Bluetooth Low Energy (BLE) beacons can also be considered as sensors. Numerous companies, like Estimote, Infsoft, Senion and many others, offer their Software Development Kits, which utilize the signals from BLE and/or Wi-Fi sensors as well as IMU data for finding a user position as a solution of Indoor Navigation task. BLE and Wi-Fi data signals differ greatly from the inertial sensor data. While the IMU signals possessing long-term drift (G) or variations (M and A) because of different metal things or internal sensor noise but are rather accurate for short period of time [13] - [14], BLE signals in opposite allow obtaining small positioning error in long-term perspective, but fluctuate greatly around its true values. As a result, modern indoor navigation systems provide fusion of BLE and IMU data for their mutual improvement. Such solutions are known as Hybrid Indoor Localization and Navigation (HILN) systems [17]. The HILN system is a drift-free, low-cost, light-weight, easy-to-integrate IPS, enabling ubiquitous navigation of pedestrians in buildings equipped with beacons or Wi-Fi. The article is organized as follows. At first, features of A, M and G sensor signals are analyzed as well as signal processing methods for IMU-based navigation system are discussed in Section 2. Next, BLE sensor signals are investigated and methods for their processing are discussed. Section 4 is devoted to the fusion technique based on a particle filter. Finally, practical results of three mentioned positioning approaches are provided for comparative analysis.

II. SIGNAL PROCESSING FOR THE IMU NAVIGATION

IMU as a combination of mainly three sensors, A, M and G, is a basis for inertial navigation system (INS) which is one of the important parts of Indoor Navigation systems. Usually INS determine a user position by implementing a Pedestrian Dead-Reckoning (PDR) algorithm [1] or its

numerous modifications (see for example [2]). At the first stage of the algorithm, a user step must be detected [3]. Next, the user step length is evaluated [6]-[8] followed by the attitude and heading estimation by one of the fusion algorithms. Finally, IMU readings are transformed from the local coordinate system to a global one using the determined device orientation angles [12]-[13]. Then the detected steps are summed up to get the user track in the building.

However, there are some challenges while designing INS. First of all the IMU readings are spoiled with noise. Let us show this fact by analyzing A and M sensor signals for the three test cases: #1. Smartphone lays on the table immovable; #2. User holds the device in hand and stays in static position; #3. User walks straight ahead. For comparative analysis, signals were harvested with Samsung Galaxy Note 5 (premium class smartphone) and Huawei P8 Lite 2017 (business class smartphone) and the mean square error (MSE) was calculated as an indicator of signal variations. It is worth noting that each sensor provides measurements as a 3D vector containing X, Y and Z axis component values of a parameter.

Results for A sensor data represented in Table I show that even in stationary cases there are fluctuations and their level depends on the smartphone class, i.e. the more expensive the smartphone, the better sensors are built in. The same results are obtained for M sensor data (see Table II). The difference between MSE values between test cases 1 and 2 is up to 5 times for Samsung Galaxy Note 5 and up to 7 times for Huawei P8 Lite 2017. It is worth noting that M readings suffer from numerous magnetic disturbances, which are available everywhere in modern buildings with their Wi-Fi spots and numerous obstacles made of steel.

TABLE I. ACCELEROMETER DATA MSE VALUES

grav. units	a_x	a_y	a_z
Test case 1			
Samsung Galaxy Note 5	0.0007	0.0007	0.0012
Huawei P8 Lite 2017	0.0139	0.0157	0.0067
Test case 2			
Samsung Galaxy Note 5	0.0176	0.0151	0.0199
Huawei P8 Lite 2017	0.0374	0.0619	0.0403

From practical point of view, the most important results are observed for the test case #3 when user moves. Accelerometer vector absolute values represented in Fig. 1 show that the A sensor components are corrupted by noise. In order to find ways how to cope with that, the A signal spectrum was investigated (Fig. 2). It is seen that there is a harmonic at about 2 Hz which corresponds to the signal oscillations due to the user steps. At the same time, there are additional harmonics at the region from 5 to 10 Hz caused by the influence of sensor noise. Note that the amplitude levels as well as the number of spurious spectral components depend on the phone model.

The other important facts can be found from the analysis of the influence of user movement on characteristics of M data. Results provided in Table II show that the MSE values of M signal components are reasonably increased for both analyzed devices. The difference is up to 3 times comparing to the stationary case #2. It is obvious that effective noise filtering is the first challenge to overcome on a way to an accurate user positioning and tracking. Especially it is important for A data because of its application for user step

detection. The task of A sensor signal denoising has numerous solutions. They are the application of the Butterworth, Bessel, Chebyshev, Savitzky–Golay, moving average, Total Variation [18] or Kalman [19] filter.

TABLE II. MAGNETOMETER DATA MSE VALUES

μT	m_x	m_y	m_z
Test case 1			
Samsung Galaxy Note 5	0.0127	0.0109	0.0315
Huawei P8 Lite 2017	0.0512	0.0418	0.0575
Test case 2			
Samsung Galaxy Note 5	0.0574	0.0637	0.0263
Huawei P8 Lite 2017	0.1696	0.2904	0.1021
Test case 3			
Samsung Galaxy Note 5	0.3913	0.2370	0.2641
Huawei P8 Lite 2017	0.5403	0.4350	0.3081

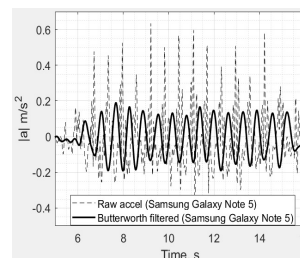


Fig. 1. The raw and filtered accelerometer signals for test case #3

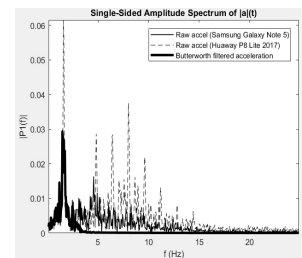


Fig. 2. Fourier spectrum estimates of the raw and filtered accelerometer signals for test case #3

The solutions differ from each other by the complexity, delay and quality of filtration. To our best knowledge there are no exact recommendations what algorithm must be used for the smartphone IMUs. Based on the held experiments, the highest performance for the A sensor was obtained from the Butterworth filter, which has the ‘smoothest’ frequency response in terms of having the most derivatives of its magnitude response being zero at the geometric center of the passband and the simplest transfer function in that the coefficients of the polynomial are easy to calculate. The result of filter application is shown in Figure 1 in temporal domain as well as in Figure 3 in spectral domain. It is clearly seen that the A signal becomes smoother due to the noise removal in high-frequency domain. It is known that MEMS gyroscope uses the Coriolis acceleration effect on a vibrating mass to detect angular rotation. The gyroscope measures the angular velocity, which is linear to rate of rotation. It responds quickly and accurately and the rotation can be computed by time-integrating the gyroscope output. But G sensor readings suffer from inertia. Without special treatment, the inertia causes the drift of the trajectory estimated by INS during the long tracks [13].

TABLE III. MSE OF GYROSCOPE VECTOR VALUES VARIATIONS

Rad. / sec.	g_x	g_y	g_z
Test case 1			
Samsung Galaxy Note 5	0.0001	0.0002	0.0001
Huawei P8 Lite 2017	0.0015	0.0025	0.0210
Test case 2			
Samsung Galaxy Note 5	0.0032	0.0027	0.0018
Huawei P8 Lite 2017	0.0078	0.0095	0.0103
Test case 3			
Samsung Galaxy Note 5	0.0204	0.0143	0.0131
Huawei P8 Lite 2017	0.0211	0.0277	0.0441

MSE values for G sensor signal components for three test cases represented in Table III show that there are G signal deviations which are caused by internal sensor noise (Case 1), user hand trembling (Case 2) and user movement (Case 3). As an example of inertia influence Fig. 3 shows the dependence of rotation angle around X axis on time. The data recorded by the smartphone laid on the table and rotated on 90 degrees anticlockwise at first and then backward to initial position. It is clearly seen that there is an integration error about $-3...-5$ degrees when device was returned to initial position. The same effects are observed for other components in case of corresponding rotations. Neither A and M sensors nor G can be considered error-free. While A signal is usually used for user step detection, G and M as well as A are mainly applied for determination of device orientation angles. Such fusion algorithms like Madgwick, Mahoney and Kalman fusion filters [14] allow neglect the drawbacks of each sensor and use their data for improving the estimate of orientation angles.

III. SIGNAL PROCESSING FOR THE BLE NAVIGATION

Another group of data processing methods deals with the signals from BLE beacons or Wi-Fi access points. They are known as Received Signal Strength Indicator (RSSI) signals. Practically, RSSI is a power of a received signal shifted by some offset value. Such signals do not suffer from the inertia, but are considerably influenced by fluctuations.

Note that the trilateration is a key method for user position estimation in beacons-based indoor navigation systems (**Error! Reference source not found.**). It requires at least three distances to the beacons near the user for its operation. Distances are obtained from the RSSI values received by application of the so-called, path-loss model [16], which represents the dependence of RSSI values on the distance between the smartphone and the corresponding beacon. Let's start the RSSI signal analysis with the investigation of number of received BLE packages. Test logs are recorded using Samsung Galaxy Note 5 device for the static test case when user stays immovable at the depicted (by diamond) point of the test room (Figure 5). There are nine installed BLE beacons, produced by Sensoro Comp.

The analysis of the so-called RSSI received package map (Figure 6a) shows that not all packages are received for each beacon (the missing packages are outlined with red rectangles). Moreover, the gaps in the RSSI signals result in misselection of the beacons.

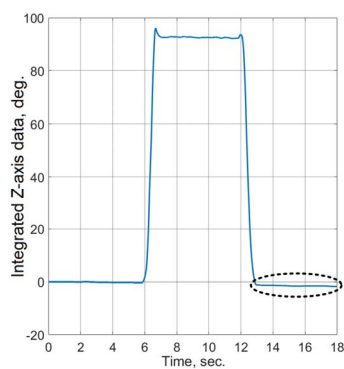


Fig. 3. The rotation angle around X device axis obtained by integrating G data values

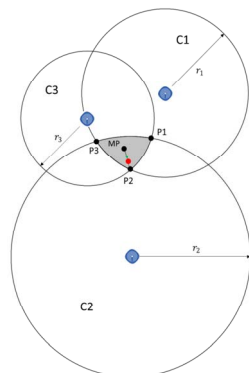


Fig. 4. Trilateration problem

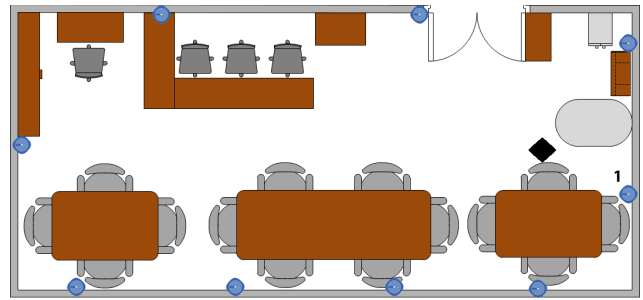


Fig. 5. Trilateration problem

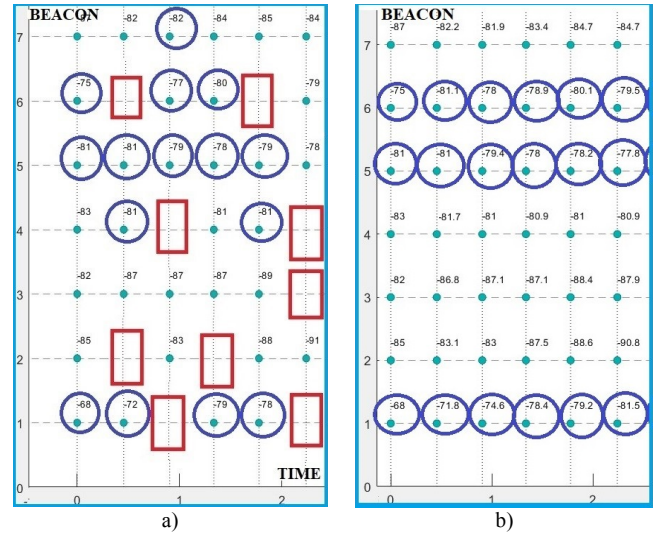


Fig. 6. RSSI package map before (a) and after Kalman filtering (b): \square - missing packets, \circ - beacons with RSSI that are selected for trilateration

For example, the selected beacons at the first step are the beacons with numbers 1 5 and 6, at the second – 1,5 and 4, and at the third – 5, 6 and 7, which makes no sense at all because the user is not moving and it is reasonable to expect the same selected beacons at each time moment. As a result, there is a discontinuous user positioning and trilateration can be not possible for some time samples.

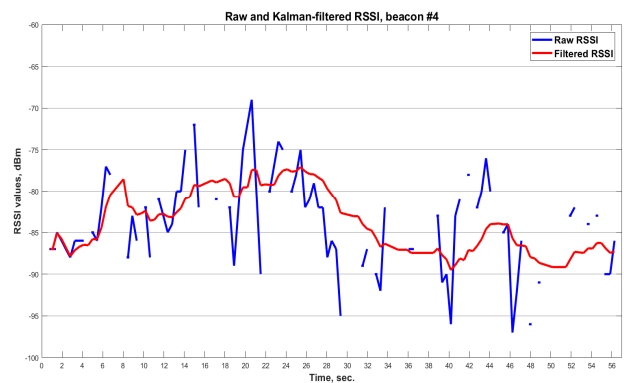


Fig. 7. Raw and Kalman-filtered RSSI signals for the beacon #4 in the test room

Next cornerstone of BLE signals is the great fluctuations of RSSI values on time for the given beacon. Figure 7 highlights this fact by an example of such a signal received from one of the beacons of the test room (Figure 5). As a result, the exact position can not be determined, and a user can be localized in some area only.

Both described challenges can be overcome by the Kalman filter for RSSI, described in [22]. The model of the process used in this case is

$$\begin{bmatrix} \text{RSSI}_i \\ \Delta \text{RSSI}_i \end{bmatrix} = \begin{bmatrix} 1 & \delta t_i \\ 0 & 1 \end{bmatrix} \begin{bmatrix} \text{RSSI}_{i-1} \\ \Delta \text{RSSI}_{i-1} \end{bmatrix} + \begin{bmatrix} v_i^{\text{RSSI}} \\ v_i^{\Delta \text{RSSI}} \end{bmatrix} \quad (1)$$

where RSSI_i and VRSSI_i are predicted values of RSSI and the RSSI's change rate for the i -th beacon; δt_i denotes time interval between current and previous received packages from the beacon; v_i^{RSSI} and $v_i^{\Delta \text{RSSI}}$ are random variables; and model of the measurement is

$$\text{RSSI}_i = \begin{bmatrix} 1 & 0 \end{bmatrix} \begin{bmatrix} \text{RSSI}_{\text{meas } i} \\ \Delta \text{RSSI}_{\text{meas } i} \end{bmatrix} + v_{\text{meas}}^{\text{RSSI}} \quad (2)$$

where $\text{RSSI}_{\text{meas } i}$ and $\text{VRSSI}_{\text{meas } i}$ are measured values of RSSI and the RSSI's change rate for the i -th beacon; $v_{\text{meas}}^{\text{RSSI}}$ is a random variable.

D. Gusenbauer [22] showed that this method (1)-(2) was able to smooth the RSSI readings from all beacons. However, we found out that this filter could also serve as a useful tool to restore the missing packets with high confidence. For doing this, the following algorithm's modification is applied – Figure 8. As you can see in Figure 6b, there are no missing packets after filtration, which is the first but not the major achievement. The beacons for the trilateration will be selected correctly, i.e. we expect to observe no weird beacon “swaps”. The last thing achieved by the Kalman filtration is the smooth signal with no intense fluctuations (Figure 7).

IV. SIGNAL PROCESSING FOR HIL NAVIGATION

No doubts, that the IMU component of a HILN system is very accurate for up to 1 minute navigation. However, this is not typical for the navigation to last only few minutes.

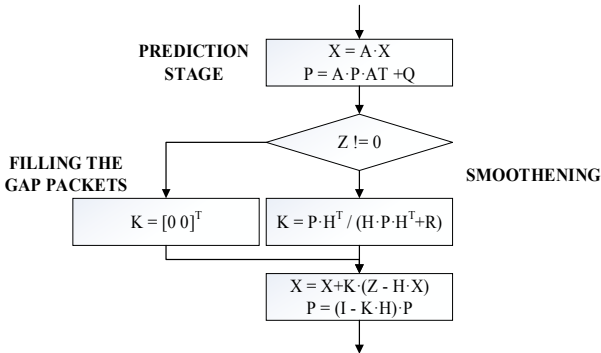


Fig. 8. The flow of Kalman filter for RSSI with restoring option:

$X = \begin{bmatrix} \text{RSSI}_i \\ \Delta \text{RSSI}_i \end{bmatrix}$ is a state matrix, $A = \begin{bmatrix} 1 & \delta t_i \\ 0 & 1 \end{bmatrix}$ is a state-transition model, P is the covariance of the process noise, $Z = \text{RSSI}_i$ is observation, $H = \begin{bmatrix} 1 & 0 \end{bmatrix}$ is the observation model, Q and R are the covariance of the process and observation noise respectively.

Hence, it sounds reasonable to expect the INS to have corrections from time to time to eliminate the accumulated drift. The data for the correction is typically obtained from the non-inertial systems, like, for example, BLE or Wi-Fi.

There are several known methods for fusing the inertial and non-inertial navigation systems. Kalman filtering is a very powerful fusion tool able to automatically determine the trust rates to different sources. However, the monolite structure of the filter makes its modification to be a very complex task. For example, if a user walks near the wall with the accumulated drift to the wall side, then sooner or later user will find himself in the wall. The considered case is typical for long (>25 meters) and narrow (<3 meters) corridors. As the accuracy of BLE is less than 3 meters, then the correction will have never be effective. Hence, the most prospecting method for the fusion is a particle filter. The filter was first proposed in 1996 in [20] and since that time takes a considerable portion of cases that relate to Markov processes. The filtering has three stages that happen every iteration and one stage that happens once (in the ideal case) or several times.

Figure 9 represents a comparison of three different approaches used for indoor navigation system design as well as the ground-truth marked with arrows. In Figure 9a you may observe a discontinuous change of user position and quite low positioning accuracy that corresponds to BLE-based system. There is a drift of the trajectory caused by the residual noise of the gyroscope (Figure 9b). This is typical situation for IMU-based indoor navigation systems. Finally, as was expected, the best performance is shown by HILN based on particle filter. It provides very accurate positioning with no visible track drifting in time. In all experiments held with HILN, the accuracy positioning error varied in the range 0.5 to 1 meter on an area of 15×6 meters, which is competitive to the leading commercial solutions.

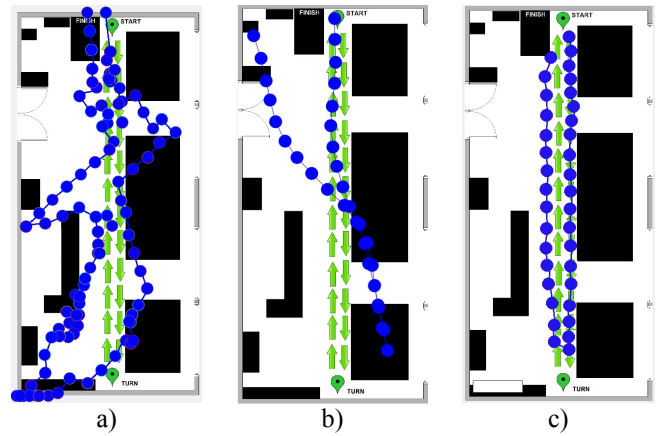


Fig. 9. Example of user position tracking performed by BLE-based system (a), IMU-based system (b) and HIL navigation system based on particle filter (c)

V. CONCLUSIONS

The indoor navigation is a prospective area for researching and commercializing the results of the researches as there is no unique opinion about the methods that must be used. The progress in the indoor navigation systems is provided the special techniques of signal processing, some of which were presented in current paper.

REFERENCES

- [1] A. Ali and N. El-Sheimy, “Low-Cost MEMS-Based Pedestrian Navigation Technique for GPS-Denied Areas,” *Journal of Sensors*, vol. 2013, Article ID 197090, 10 pages, 2013. doi:10.1155/2013/197090.

- [2] T. Zengshan, Z. Yuan, Z. Mu, and L. Yu, "Pedestrian dead reckoning for MARG navigation using a smartphone." *EURASIP J. Adv. Sign. Process.*, vol. 1, pp. 1–9, 2014.
- [3] A. Ali, and N. El-Sheimy, "Low-cost MEMS-based pedestrian navigation technique for GPS-denied areas," *Journal of Sensors*, 10 pages, 2013. 10.1155/2013/197090.
- [4] G. Trein, N. Singh, and P. Maddila, "Simple approach for indoor mapping using low-cost accelerometer and gyroscope sensors," *DOCPLAYER*, 2013.
- [5] H. Bao and W.-Ch. Wong "A Novel Map-Based Dead-Reckoning Algorithm for Indoor Localization," *Journal of Sensor and Actuator Network*, vol. 3, pp. 44-63, 2014. doi:10.3390
- [6] H. Weinberg, *Using the ADXL202 in Pedometer and Personal Navigation Applications*. Analog Devices, Inc.; Norwood, MA, USA: 2002.
- [7] Q. Tian, Z. Salcic, K.I.-K. Wang, and Y. Pan, "A Multi-Mode Dead Reckoning System for Pedestrian Tracking Using Smartphones," *IEEE Sens. J.*, vol. 16, pp. 2079–2093, 2016. doi: 10.1109/JSEN.2015.2510364.
- [8] N.-H. Ho, P. H. Truong, and G.-M. Jeong, "Step-Detection and Adaptive Step-Length Estimation for Pedestrian Dead-Reckoning at Various Walking Speeds Using a Smartphone," *Sensors*, Basel, Switzerland, vol. 16(9), pp. 1423, 2016.. <http://doi.org/10.3390/s16091423>
- [9] S.O.H. Madgwick, *An efficient orientation filter for inertial and inertial/magnetic sensor arrays*. 2010.
- [10] R. Mahony, T. Hamel, and J.-M. Pflimlin, "Complementary filter design on the special orthogonal group SO(3)," *Proceedings of the 44th IEEE Conference on Decision and Control, and the European Control Conference 2005 Seville, Spain, December 12-15, 2005*.
- [11] S. Mau, "What is the Kalman Filter and How can it be used for Data Fusion?" *Robotics Math*, pp. 16-811, December 2005.
- [12] M. Pedley, "Tilt Sensing Using a Three-Axis Accelerometer," *Freescale Semiconductor*, AN3461, 2013.
- [13] R. Zhi, *A Drift Eliminated Attitude & Position Estimation Algorithm In 3D*. Graduate College Dissertations and Theses, University of Vermont, 2016.
- [14] F. Abyarjoo, A. Barreto, J. Cofino, and F. R. Ortega, "Implementing a Sensor Fusion Algorithm for 3D Orientation Detection with Inertial/Magnetic Sensors." In: Sobh T., Elleithy K. (eds) *Innovations and Advances in Computing, Informatics, Systems Sciences, Networking and Engineering*. Lecture Notes in Electrical Engineering, vol 313. Springer, Cham. 2015, pp. 305-310.
- [15] F. Zafari, I. Papapanagiotou, M. Devetsikiotis, and T. Hacker "An iBeacon based Proximity and Indoor Localization System," arXiv:1703.07876v2 [cs.NI] 24 Mar 2017.
- [16] K. Vadivukkarasi, R. Kumar and Mary Joe, "A Real Time Rssi Based Novel Algorithm to Improve Indoor Localization Accuracy for Target Tracking in Wireless Sensor Networks," *ARNP Journal of Engineering and Applied Sciences*, vol. 10, no. 16, pp. 7015-7023, SEPTEMBER 2015.
- [17] T. Qinglin et al. "A Hybrid Indoor Localization and Navigation System with Map Matching for Pedestrians Using Smartphones." Ed. Kourosh Khoshelham and Sisi Zlatanova. *Sensors* (Basel, Switzerland) 15.12 (2015): 30759–30783. PMC. Web. 14 Mar. 2018.
- [18] A. Masse, S. Lefèvre, R. Binet, S. Artigues, G. Blanchet, and S.Baillarin, "Denoising Very High Resolution Optical Remote Sensing Images: Application and Optimization of Nonlocal Bayes method," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 11:3, pp. 691-700, 2018.
- [19] T. Singhal, A. Harit, and D. N. Vishwakarma, "Kalman Filter Implementation on an Accelerometer sensor data for three state estimation of a dynamic system," *International Journal of Research in Engineering and Technology (IJRET)*, vol. 1, no. 6, 2012. ISSN 2277 – 4378
- [20] P. Del Moral, "Non Linear Filtering: Interacting Particle Solution". *Markov Processes and Related Fields*. 2 (4) pp. 555–580, 1996.
- [21] B.I. Ahmad, J. Murphy, P.M. Langdon, and S. J. Godsill, "Filtering perturbed in-vehicle pointing gesture trajectories: Improving the reliability of intent inference", *Machine Learning for Signal Processing (MLSP) 2014 IEEE International Workshop on*, pp. 1-6, 2014.
- [22] D. Gusenbauer, C. Isert, and J. Krosche, "Self-contained indoor positioning on off-the-shelf mobile devices," *International Conference on Indoor Positioning and Indoor Navigation (IPIN)*, pp. 1-9, 2010.

Data Mining in Scientometrics: Usage Analysis for Academic Publications

Olesya Mryglod
*Laboratory for Statistical Physics
of Complex Systems Institute for Condensed
Matter Physics, NAS of Ukraine,
Lviv, Ukraine*
*L⁴ Collaboration & Doctoral College for the
Statistical Physics of Complex Systems*
Leipzig-Lorraine-Lviv-Coventry, Europe
kriukovagv@ukma.edu.ua

Yurij Holovatch
*Laboratory for Statistical Physics
of Complex Systems Institute for Condensed
Matter Physics, NAS of Ukraine,
Lviv, Ukraine*
*L⁴ Collaboration & Doctoral College for the
Statistical Physics of Complex Systems*
Leipzig-Lorraine-Lviv-Coventry, Europe
hol@icmp.lviv.ua

Ralph Kenna
*Applied Mathematics Research
Centre, Coventry University
Coventry, CV1 5FB, England*
*L⁴ Collaboration & Doctoral College for the
Statistical Physics of Complex Systems*
Leipzig-Lorraine-Lviv-Coventry, Europe
csx267@coventry.ac.uk

Abstract—We perform a statistical analysis of scientific-publication data with a goal to provide quantitative analysis of scientific process. Such an investigation belongs to the newly established field of scientometrics: a branch of the general science of science that covers all quantitative methods to analyze science and research process. As a case study we consider download and citation statistics of the journal “Europhysics Letters” (EPL), as Europe’s flagship letters journal of broad interest to the physics community. While citations are usually considered as an indicator of academic impact, downloads reflect rather the level of attractiveness or popularity of a publication. We discuss peculiarities of both processes and correlations between them.

Index Terms—scientometrics, usage metrics, citation analysis, data analysis

I. INTRODUCTION

Research, as any other kind of human activity, can be considered as a source for large sets or streams of multidimensional data, starting from trivial statistics of the number of researchers or the number of papers they produce [1], [2] and ending with records of communication through social networks on special topics in real time [3]. The opportunity to record, to store, and to analyze scientific-related data itself has become a trigger for the extremely fast development of scientometrics [4] — a branch of the general science of science that covers all quantitative methods to analyze science and research processes. The majority of scientometric tasks are highly applicable in practice. This is especially true today, when the proportion of researchers has become large enough to create a competitive environment and to become an object of careful review by governments [2]. It is always good to present

some new knowledge about the universe or some phenomenological findings, but nowadays the researcher is facing more mercantile questions such as: what is the practical impact, how the new results can be implemented into the technology, etc. Therefore, numerous quantitative indicators and metrics related to science have become candidates for immediate implementation in assessment procedures on different scales. This is not bad thing in general, but the important caveats should be kept in mind: (i) it is extremely hard to formalize any human activity in general, (ii) the work of scientist is not mechanical, but to large extent it is a creative process and (iii) science is far from being homogeneous — different disciplines should be evaluated in completely different ways. While the controlling and assessment of research activity seems to be the requirement of time, it should be performed in a non-trivial way and taking into account the multidimensionality of data used.

While journal publication remains the most common way to present the results of research, starting from the middle of past century citations became a sort of ‘currency’ in science [5]. Citation data are currently used for calculation of different indicators of research efficiency despite numerous drawbacks and caveats. Even not considering the peculiarities of citation counting, the meaning of a single citation is ambiguous. The motivation to cite a particular source can be various [6]. Therefore, recently it is usually considered as the unit of *impact*. The impact can be positive or negative; it can represent the usefulness of a methodology, the novelty of subject or the informativeness of a review presented in a publication; it can

be a form of acknowledgement; and so on.

The increasing popularity of different virtual social media, blogs and services such as managers of references, bookmarking, etc. caused the appearance of new source for scientometric data. New categories of metrics, called altmetrics or informetrics, were presented some time ago [3]. The latter obviously contain completely different information about publications. At this point one can think about such special characteristics as popularity. This can be already dependent not solely on scientific value of results, but also on the way how they are presented and promoted. One of the motivations behind is that even a good result often should be well positioned in such competitive environments. Even a very special ‘drop in an ocean’ can remain unseen, not to make an impact, to stay uncited and, therefore, to be underestimated. There are the kind of data which can say something about the attractiveness of the paper even before it is read. The statistics of downloads are referred to as one of the so-called usage metrics. Even if downloading never guarantees that the paper will be fully or even partially read, it can be considered as the unit of reader’s interest.

Considering downloads and citations — two dimensions of data characterizing research papers — one would initially speculate about the correlation between them. On the one hand, it is intuitively acceptable that, in the main, the wider the readership (downloads), the greater the probability that a document will be cited (citations). On the other hand, it has already been shown that downloads may influence citations and vice versa [7], [8]. The correlation coefficients between downloads and citations reported in the literature seem to be different not only for different disciplines, but for different journals and even for different types of publications [8]. This brings us to the clustering problem - the papers (journals?) characterized by similar patterns in downloading and citing should be grouped and analyzed separately in order to understand the nature of both processes and the relationship between them. In this paper we discuss a case study containing the results of the analysis of statistics of full-text download and corresponding citations of publications in *EPL* (formerly known as “Europhysics Letters”) — Europe’s flagship peer-reviewed letters journal of broad interest to the physics community. It is published by EDP Sciences, IOP Publishing and the Italian Physical Society on behalf of the European Physical Society and 17 other European physical societies [9], [10]. The rest of the paper is organized as follows: *EPL* download statistics is analysed in the next Section II and is compared with the citation statistics in Section III. We end by conclusions and outlook in Section IV.

II. EVOLUTION OF DOWNLOAD PROCESS

Scientific citations have been an object of careful consideration starting from the middle of past century. Even if collecting the citations is a highly non-trivial process, there are few comparatively reliable sources of the corresponding data (Web of Science [11], Scopus [12]). The downloads, on the other hand, have become a topic of interest later [13].

Download statistics are not publicly available — these have to be provided by publishers or, say, by owners of repositories. Therefore, the analysis of downloading process is important itself as well.

What is the downloading ‘profile’ of any scientific periodical? Does some typical downloading pattern for individual paper exist? Is it possible to perform clustering of publications according to their attractiveness for readers? Some clues to answer these questions can be found in this section containing the results of analysis of downloads for *EPL* [14]. The data on the full-text downloads for papers, published in *EPL* between January 2007 to June 2013 with one month resolution provided by *IOPscience*¹ were used for this purpose. The downloads are counted on an IP-address basis with multiple requests made from the same address considered as separate downloads. Only full-text downloads from the *IOPscience* web-pages are counted. The data are automatically cleansed of suspicious and robot activity, and are COUNTER compliant, see [15]. The data set comprises 377 open access (OA) papers (freely available via web-page for unregistered users) and 4986 non-OA papers (require a payment).

We used two approaches to analyse the time-involving sequences of data: (i) to fix the period of real time (say, a specific month, year, etc.) and to analyse the corresponding downloading statistics for papers of different ages² — the so-called *synchronous* method; or (ii) to consider the statistics of downloads of individual paper in its personal timescale, which starts at paper’s publication online — the *diachronous* approach. The latter allows one to compare the individual processes for separate publications in order to find patterns. Cumulative numbers of downloads d_i^{cum} vs. corresponding paper’s ages are plotted in Fig. 1. Each curve here represent a given paper and the value (x, y) means that this paper has got y downloads during x months after its publication.

Having the family of curves, there is a temptation to get some typical curve by averaging all values of d_i^{cum} for each age. Since the distribution of these values is skewed, median instead of averages should rather be used, see Fig. 1. The root-mean-square deviation (RMSD) of d_i^{cum} from the median values can be used to gauge the extent to which the latter can be considered as typical ones. The value $RMSD_c = 66$ is empirically determined as critical for non-OA papers ($RMSD_c = 105$ for OA papers) to detect the individual curves most close to median values. This allows us to determine the core of typical papers: about 60% of non-OA and 50% of OA papers. The fast accumulation of downloads seems to be typical during the first couple of months after publication, while the process is getting slower afterwards. The rest of the papers — let us call them atypical to distinguish from the first group — are basically characterized by the curves of a similar smooth shape, but different rates of downloads accumulation. However, some exceptions can be visually seen in Fig.1:

¹*IOPscience* is the online service for the journals of the *Institute of Physics (IoP)*

²Here the paper’s age is calculated as the number of months passed since it has been published on the web-page.

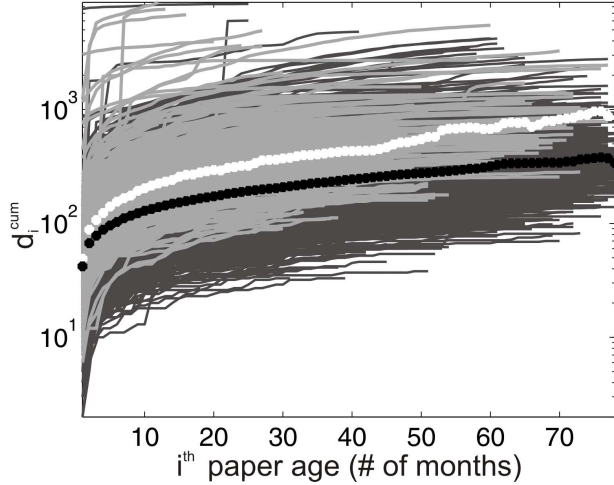


Fig. 1. Cumulative number of downloads for each paper individually vs. its age online: dark lines represent non-OA papers while light lines represent OA papers. The corresponding median values for papers of the same age are indicated by the circles (black for non-OA papers and white for OA papers).

unusual bursts can be noticed for particular papers. In order to capture such bursty behaviour the standard deviations can be used. Let $\sigma_i(T)$ be the standard deviation of i th paper over its entire history to month T . The median values are considered as benchmark to be compared to. $\sigma(T)$ is the standard deviation calculated for median values but only over the first T months. Each paper is then characterized by a value of $\Delta_i(T)$:

$$\Delta_i(T) = |\sigma_i(T) - \sigma(T)|.$$

The noticeable bursts can be flagged by large values of $\Delta_i(T)$ comparing to the average $\langle \Delta_i(T) \rangle$. The investigation of ratio $\Delta_i(T)/\langle \Delta_i(T) \rangle$ lead us to its critical value equals to 5. In this way, 96 non-OA papers (2% of all) and 10 OA papers (2.7%) demonstrating noticeable bursts were detected. The position and the nature of bursts can be the issue of separate analysis. E.g., some of bursty papers can be labelled as ‘sleeping beauties’ due to its delayed recognition (the downloading bursts occurred later than 6 months after publication for 35 non-OA and for 2 OA papers), see Table II.

Continuing with atypical non-bursty papers, we can also analyse their overall attractiveness. The cumulative curves of the more than 1/3 of these papers lie entirely over the median values. Another 1/3 of them are of persistently lower attractiveness comparing to medians. The rest are characterized by cumulative curves which cross medians curve, see Table II.

Another way to investigate the downloading rates for non-bursty papers is to use the notion of half-life M_i^{50} which is the number of months by which a paper achieve 50% of its current downloads. If this value falls between the percentiles P25 and P50, the corresponding paper is considered as usual, i.e. demonstrating the normal typical rate of downloads accumulation. The M_i^{50} value smaller than P25 means that 50% of downloads were acquired by i th paper faster than usual — we called such papers ‘flashes in the pans’. In contrast,

the ‘delayed’ papers are characterized by M_i^{50} values greater than P50. The exact numbers of *EPL* publications within these categories are shown in Table II.

TABLE I
CATEGORISATION OF *EPL* PAPERS ACCORDING TO DOWNLOADS (THE DATA FOR OA PAPERS ARE IN BRACKETS) [14].

Categorisation by burstiness		
4 986 (377) papers	98% are “non-bursty” papers	
	2% (3%) are “bursty” papers	1% (1%) are “sleeping beauties” 1% (2%) burst early
Categorisation by overall attractiveness		
4 890 (367) non-bursty papers	60% (50%) have typical overall attractiveness	
	40%(50%) are atypical	18% (22%) are more attractive
		12% (15%) are less attractive 10% (13%) the rest
Categorisation by half-lives (ageing of attractiveness)		
4 890 (367) non-bursty papers	62% (65%) exhibit usual ageing behaviour	
	18% (17.5%) are flashes-in-the-pan	
	20% (17.5%) exhibit delayed activity	

Coming back to *synchronous* approach, we can get the downloading patterns of the entire journal on a calendar basis — i.e., for different time periods. In this way 3-months observation periods were chosen to compare the statistics of downloads versus paper ages for three journals: *EPL* (OA and non-OA papers), *Tetrahedron Letters* and *Condensed Matter Physics*. The second curve is the digitized version of Fig. 1 from ref. [7] which corresponds to download data for the journal *Tetrahedron Letters*. The third one is based on the publisher’s downloading data containing 1201 records with one month resolution. Each symbol (x, y) in Fig. 2 means that $y\%$ of all downloads accumulated within given time period were to papers not older than x months.

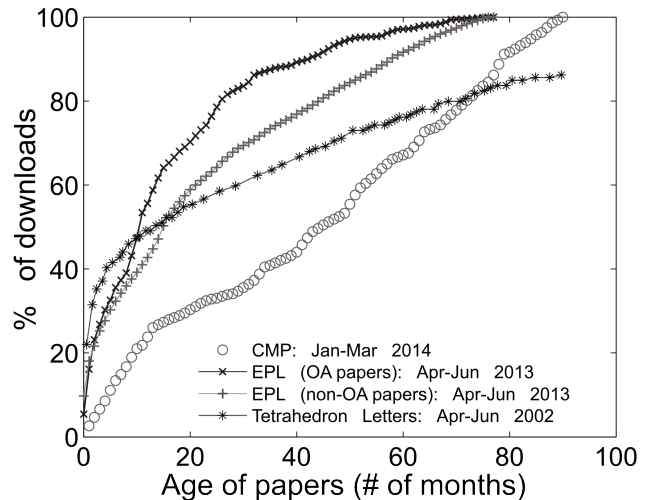


Fig. 2. Relative cumulative distributions of ages of downloaded papers. The “+” symbols represent non-OA *EPL* papers, “x” represent OA *EPL* papers [14], circles correspond to CMP data, and “*” represent data for *Tetrahedron Letters* [7].

As one can see in Fig. 2, the downloading process looks different for all three journals. This is rather an expected

finding since they differ by discipline (Physics for *EPL* and *CMP*, Organic Chemistry for *Tetrahedron Letters*), dominant kind of publications (rapid short publications in *EPL* and *Tetrahedron Letters*, and original regular papers in *CMP*), frequency (*Tetrahedron Letters* publishes one issue per week; *EPL* — per two weeks; *CMP* — per three months) and the rules for online access (full open access in *CMP* and hybrid scheme in the rest two). However, careful consideration reveals similar features: a rather faster accumulation of downloads during couple of months immediately after publication online and further slowing the process afterwards. In ref. [7] a critical point of three months was discussed, as a demarcation between two different regimes characterised by different slopes (see also [16]). A two-factor model was proposed in [7] to describe the corresponding two regions which supposedly can be explained by two initial motivations of users: to download more recent papers because of their novelty and older ones for archiving, background reading or similar. According to this model, two exponents can be used in order to fit non-cumulative data (we plot the density of downloads ρ – defined as mean numbers of downloads per paper – against their age see Fig. 3):

$$\rho(t) = \rho_0 [A \exp(-b_1 t) + (1 - A) \exp(-b_2 t)], \quad (1)$$

$$0 \leq A \leq 1, \quad b_1 > 0, \quad b_2 > 0,$$

where A and $(1 - A)$ are relative weights of the two factors (two different motives for downloads) and ρ_0 is the density of downloads which corresponds to the newest papers (published in the month of downloading). The parameters b_1 and b_2 are exponential decay constants corresponding to early and later download patterns.

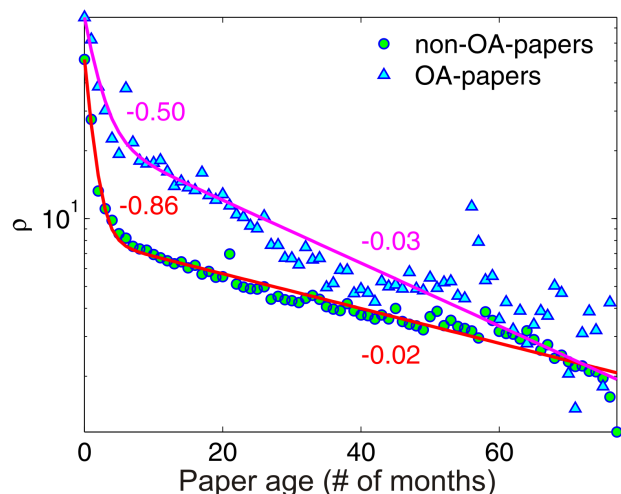


Fig. 3. Density of downloads per paper (ρ) versus papers' ages. The solid curves show the model (1) predictions and the corresponding exponential decay constants are indicated.

Using nonlinear-curve least-squares fitting, we obtain the estimates for non-OA papers $A \approx 0.84$, $b_1 \approx 0.86$, $b_2 \approx 0.02$ and for OA papers $A \approx 0.71$, $b_1 \approx 0.50$, $b_2 \approx 0.03$ (parameters close to the estimates of [7] for *Tetrahedron Letters* which

are: $A \approx 0.92$, $b_1 \approx 0.50$, $b_2 \approx 0.014$). The OA downloads are more concentrated on the first months after publication online. The model allows one to make predictions for long-term behaviour: e.g., typically 50% of non-OA downloads collected during by one month are to papers 25 months old or less, see [14].

III. DOWNLOADS VS. CITATIONS

Probably the hottest topic of interest is connected with analysis of correlation between downloads and citations. Due to the current applicability of citations, the downloads are usually considered as a potential proxy for future citations. A number of studies were already done in this direction. It seems that such desirable correlations are in fact dependent on different factors such as discipline, type of publications or the peculiarities of a specific journal [8], [17]. Still, there is a lack of clear understanding of relation between the two processes. While for a sake of statistical reliability aggregated data for several journals considered as a sort of super-journal are sometimes analysed, some specific features of separate journals can be missed. Therefore, any useful idea towards revealing clusters of journals (papers) with similar downloads-citations patterns is required. This case study is another contribution in this direction.

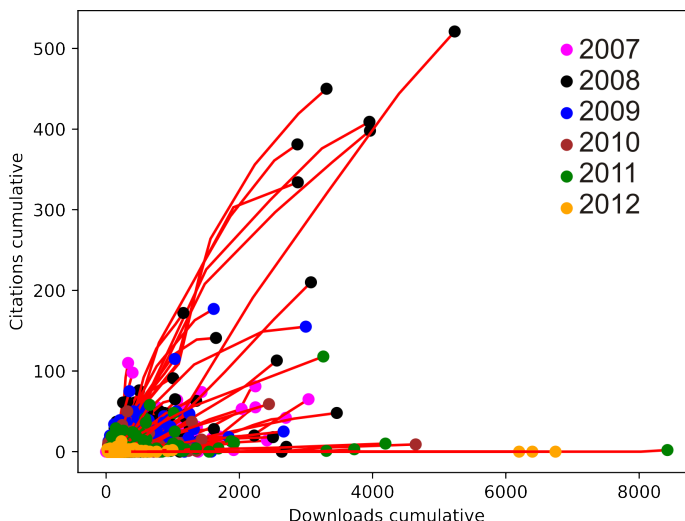


Fig. 4. Cumulative downloads vs. cumulative citations. The growing of annual values is shown by line while the final cumulative values are marked by circles. The colors of circles correspond to different publication years.

The citation data for 4894 *EPL* papers were acquired from the Scopus database. In order to compare it with the existing data, the same publication and citation windows were chosen: [2007–2012]. Therefore, two sequences of values were assigned to the majority³ of *EPL* papers published within 2007 and 2012 years. Since only annual citation data are available, the corresponding downloads were also grouped for years. The annual pairs of cumulative values (downloads vs. citations) per each paper are plotted in Fig. 4.

³Some papers cannot be found in Scopus probably for technical reasons.

The very natural first step is to find the correlation between the final (cumulative) values of downloads d and citations c shown by circles in Fig. 4. Rather low value of Pearson correlation coefficient $PCC \approx 0.49$ shows the absence of evident correlation between the for corresponding values for all papers. The correlation coefficient values for papers published in different years are shown in Fig. 5: higher for papers with longer history and lower for more recent publications. The results are fairly similar for cumulative citations counted in further two years: 2013 and 2014. In spite of the expectations to see better correlation after taking into account 2-years time delay for citations [18], the cumulative downloads d_{2012} are not better correlated with c_{2013} and c_{2014} .

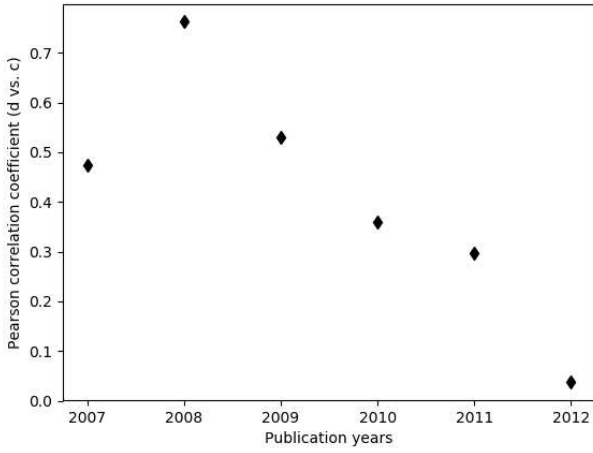


Fig. 5. Pearson correlation coefficients (PCC) characterizing the correlation between the cumulative numbers of downloads d and citations c vs. publication year of papers.

Further speculations are relevant to the shape of d vs. c cumulative curves (lines in Fig. 4). The straight lines are supposed to represent the downloading and citing processes governed by constant motives: the annual increment of ratio c/d remains the same to some extent. The zig-zag-like shape, in the opposite case, demonstrates the change of initial motives providing more attention by readers (faster accumulation of downloads) or more acknowledgement by citing authors (more intensive citing process).

The sequence of value of angles α between the curve and the horizontal axis characterizes the zig-zag-ness of each long enough cumulative curve d vs. c , see an example in Fig. 7.

The larger is difference between $\Delta = \alpha_{\max} - \alpha_{\min}$ for each paper, the steeper direction change can be found in the corresponding cumulative curve. The empirical investigation of the distribution of Δ values for all papers brought us to its possible critical values: $\Delta_{c1} \approx 0.3$ and $\Delta_{c2} \approx 8$. Such a criterium to cluster 2315 *EPL* papers with 3 or more years long publication history was applied. Three groups of papers are shown in Fig. 7: 405 papers characterized by slow zig-zag-ness, $\Delta_i \leq \Delta_{c1}$; 1606 — medium zig-zag-ness, $\Delta_{c2} < \Delta_i \leq \Delta_{c2}$; 304 — high zig-zag-ness, $\Delta_{c2} < \Delta_i$.

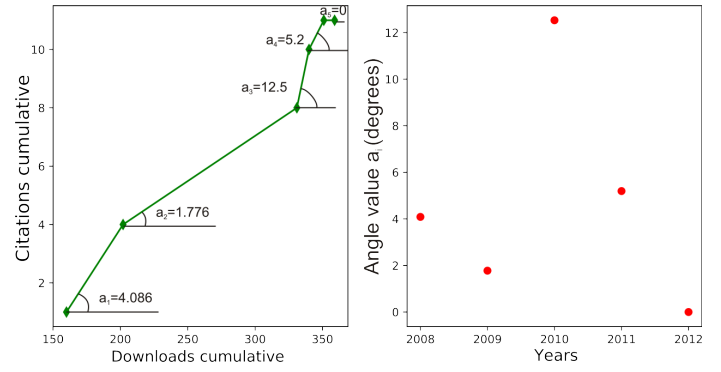


Fig. 6. Right-hand panel: Cumulative annual number of citations c vs. cumulative annual number of downloads d for selected paper published in 2007. Left-hand panel: the corresponding values of angles characterizing the annual changes of cumulative curve direction.

The clustering of papers into several groups makes further conclusions less reliable statistically. However, while the publication years and kinds of publications are presented more or less similarly in all three clusters of papers, the correlations between cumulative values of downloads d and citations c vary. The smallest value of $PCC \approx 0.19$ is found for the first group. $PCC \approx 0.68$ describes much stronger positive correlation for the second. $PCC \approx 0.84$ for the third category. Therefore, one can state that the strongest correlation between the cumulative counts of downloads and citations was observed for the papers characterized by zig-zag-like cumulative evolution of $c(d)$, i.e. by changing of motives governing the downloading and citing.

IV. CONCLUSIONS AND OUTLOOK

A case study of downloading and citing of academic publications is presented in this paper. The data from *EPL* journal were used to analyse the dynamics of download process on a scale of entire journal as well as on a scale of individual publications. The first part of the paper contains our recent results on this topic [14]. The download rate is naturally different for publications of different age — the highest interest of potential readers is attracted by newly published papers. Older papers are downloaded less actively. The corresponding cumulative curve of downloads can be considered as one of the characteristics which describes the entire journal. One of the models designed to describe change of users' motives is applied: the function of two exponents allows one to make the long-term predictions about downloading journal publications. The second part of the papers is devoted to the comparison of downloads and citations — this study is on its initial stage but some very new results are presented here. Not only the correlation between cumulative downloads and citations counts is checked, but a way to cluster publications by their history of downloading and citing is suggested. A further investigation is needed in order to find the reasons for different relationships between these two characteristics.

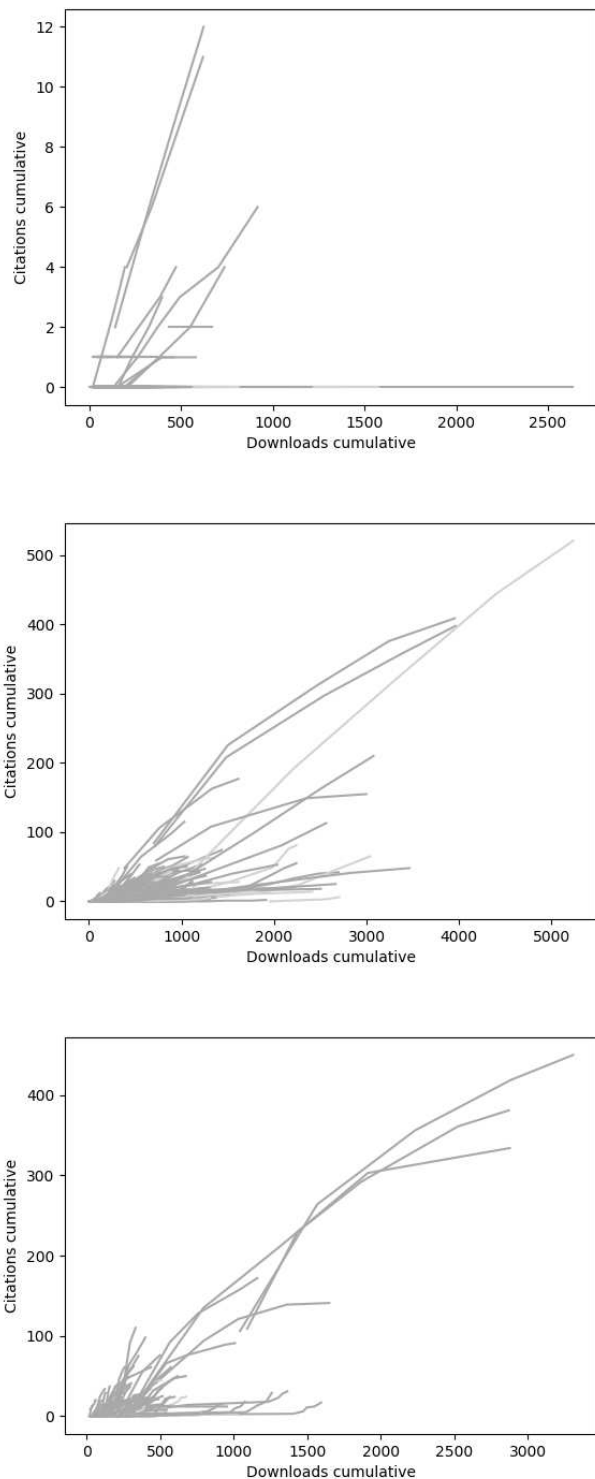


Fig. 7. Cumulative annual number of citations c vs. cumulative annual number of downloads d for papers with 3 or more years long publication history characterized by $\Delta \leq 0.3$ (top panel), $0.3 < \Delta \leq 8$ (middle panel) and $8 < \Delta$ (bottom panel).

ACKNOWLEDGMENT

We thank Daniel Barrett (IOP Publishing) and the staff of *EPL* for providing the data and assistance. The work was

supported by the grant of the National Academy of Sciences of Ukraine, No. 0118U003620 (O.M.) and by the 7th FP, IRSES project No. 612707 “Dynamics of and in Complex Systems” (DIONICOS).

REFERENCES

- [1] Nalimov, V., Mulcjenko, B. (1971). Measurement of Science: Study of the Development of Science as an Information Process. Washington DC: Foreign Technology Division.
- [2] Price D. J. de S. Little Science, Big Science // New York, Columbia U.P. 1963.
- [3] Priem, J. (2014). Altmetrics. In B. Cronin & C. Sugimoto (Eds.), Beyond Bibliometrics: Harnessing Multidimensional Indicators of Scholarly Impact. London: MIT Press.
- [4] Mingers, J. and Leydesdorff, L., A review of theory and practice in scientometrics. European Journal of Operational Research, 2015, 246(1), pp.1-19.
- [5] Garfield, E., 1979. Citation indexing:: its theory and application in science, technology, and humanities. New York: John Wiley & Sons.
- [6] Tahamtan I., Bornmann L., Core elements in the process of citing publications: Conceptual overview of the literature. Journal of Informetrics, 2018, 12(1), pp.203-216.
- [7] Moed H.F. Statistical Relationships Between Downloads and Citations at the Level of Individual Documents Within a Single Journal. J. Am. Soc. Inf. Sci. Tec., **56** (2005) 1088.
- [8] Moed H.F., Halevi G., On full text download and citation distributions in scientific-scholarly journals. Journal of the Association for Information Science and Technology, 2016, 67(2), pp.412-431.
- [9] EPL official web-page. <http://iopscience.iop.org/journal/0295-5075> Accessed 25 March 2018.
- [10] EPL (journal) wikipedia page. [https://en.wikipedia.org/wiki/EPL_\(journal\)](https://en.wikipedia.org/wiki/EPL_(journal)) Accessed 25 March 2018.
- [11] Web of Science, <https://clarivate.com/products/web-of-science/>. Accessed 21 March 2018.
- [12] Scopus, <http://www.scopus.com/>. Accessed 21 March 2018.
- [13] Kurtz, M.J., Eichhorn, G., Accomazzi, A., Grant, C., Demleitner, M., Murray, S.S., Martimbeau, N. and Elwell, B., 2005. The bibliometric properties of article readership information. Journal of the Association for Information Science and Technology, 56(2), pp.111-128.
- [14] Mryglod O., Kenna R., Holovatch Y. Is your EPL attractive? Classification of publications through download statistics. EPL (Europhysics Letters). 2014 Dec 15;108(5):50011.
- [15] Counter project. <https://www.projectcounter.org/>. Accessed 24 March 2018.
- [16] Watson A., Comparing citations and downloads for individual articles at the Journal of Vision, *Journal of Vision*, **9** (2009) i, 1.
- [17] Gorraiz J., Gumpenberger C., Schlogl C., Usage versus citation behaviours in four subject areas. Scientometrics, 2014, 101(2), pp.1077-1095.
- [18] Wan, J.K., Hua, P.H., Rousseau, R. and Sun, X.K., 2010. The journal download immediacy index (DII): experiences using a Chinese full-text database. Scientometrics, 82(3), pp.555-566.

Using Recurrent Procedures to Identify the Parameters of the Large-sized Object Moving Process Model in Real Time

Hanna Rudakova
Engineering Cybernetics Department
Kherson National Technical University
Kherson, Ukraine
RudakovaAnna25@gmail.com

Oksana Polyvoda
Engineering Cybernetics Department
Kherson National Technical University
Kherson, Ukraine
pov81@ukr.net

Anton Omelchuk
Engineering Cybernetics Department
Kherson National Technical University
Kherson, Ukraine
tareon@ukr.net

Abstract—A recurrent identification procedure by the method of least squares is considered to determine the parameters of the large-sized object moving process model by the example of a ship's descent on a slipway. The dynamics of the adjustment of the values of the components of the model main matrix and the state vector obtained as a result of calculations is shown. The conditions for the end of the identification procedure and its re-activation are determined.

Keywords—adaptive control system, the method of least squares, procedure periodic, parameter estimation, large-sized object, a recurrent algorithm, the identification procedure

I. INTRODUCTION

The movement of large-sized objects is carried out through the interconnected operation of electric drives and mechanisms complex. The processes of displacement occur under non-stationary conditions under the influence of external factors, which change significantly throughout the movement, which often leads to the occurrence of emergencies [1]. Reliable functioning of such complexes is possible only with the use of adaptive control systems under the influence of stochastically changing external and internal factors. The tasks of identifying the parameters of the displacement process model arise periodically in the synthesis of the electromechanical complexes components operation rational control. Recently, the trend is to use artificial intelligence techniques to perform the identification process, such as fuzzy logic [2], neural networks [3], etc. However, in order to use the neural network technology, it is necessary to use a high-volume training set for training the system, which is not always possible. When using fuzzy systems, there is a problem creating a base of rules and membership functions, which is especially difficult in the conditions of a sharp change in the factors that influence the process.

II. PROBLEM STATEMENT

The movement of a large object on a surface can be regarded as the sum of translational and rotational motions, which are described by equations with respect to the center of mass [4]:

– for translational motion

$$m \frac{dv}{dt} = \sum F, \quad \frac{dS}{dt} = v; \quad (1)$$

– for rotational motion

$$J \frac{d\omega}{dt} = \sum M, \quad \frac{d\varphi}{dt} = \omega, \quad (2)$$

where m is the mass of the object, v is the speed of the translational motion, S is the displacement, t is the time, $\sum F$ is the sum of the external forces applied to the object, J is the inertia moment of the object relative to the axis of rotation, ω is the angular velocity of the object, φ is the angle of the object rotation, $\sum M$ is the total rotational moment of all forces with respect to axis of the mass center rotation of a large-sized object.

The system of equations (1), (2) can be represented as equations in the state space of the fourth order [5] with state vector of the object being moved $\mathbf{x} = (x_1 \ x_2 \ x_3 \ x_4)^T$ where $x_1 = S$, $x_2 = v = dS/dt$, $x_3 = \varphi$, $x_4 = \omega = d\varphi/dt$; vector of control actions \mathbf{u} and output vector \mathbf{y} characterizing the structure of the monitoring system. This systems model is, as a rule, non-linear dependencies. In the synthesis of optimal control actions, the methods of modern control theory usually use the linearized model of system, represented in the state space [6].

Consider the process of moving a bulky object by example of a ship's descent on a slipway. Forces acting on a distributed, moving "ship-trolleys" object during a controlled descent are shown in Fig. 1.

The model of ship movement on a slipway, developed in the states space, which takes into account all the significant external factors, is nonlinear [7].

As a result of linearization, a mathematical model was obtained in the states space in the vector-matrix form

$$\hat{\mathbf{x}} = \mathbf{A}\hat{\mathbf{x}} + \mathbf{B}\hat{\mathbf{u}}, \quad \hat{\mathbf{y}} = \mathbf{C}(k) \cdot \hat{\mathbf{x}}, \quad (3)$$

where

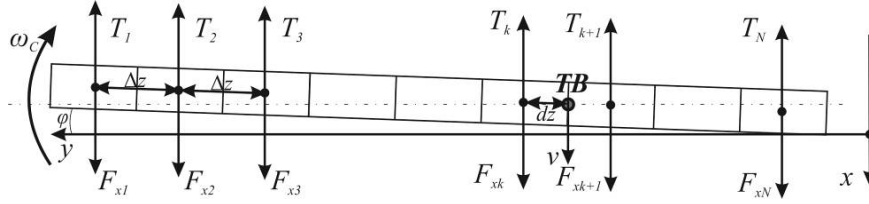


Fig. 1. The structure of forces actions on the "ship-trolleys" object.

$$\mathbf{A} = \begin{pmatrix} 0 & 1 & 0 & 0 \\ a_{21} & 0 & a_{23} & 0 \\ 0 & 0 & 0 & 1 \\ a_{41} & 0 & a_{43} & 0 \end{pmatrix},$$

$$\mathbf{B} = \begin{pmatrix} 0 & 0 & \dots & 0 \\ K_1 & K_1 & \dots & K_1 \\ 0 & 0 & \dots & 0 \\ q_1 K_2 & q_2 K_2 & \dots & q_N K_2 \end{pmatrix}, \quad (4)$$

$$\mathbf{C}(k) = \begin{pmatrix} 1 & 0 & q_1 & 0 \\ 1 & 0 & q_N & 0 \end{pmatrix}.$$

The elements of the matrix \mathbf{A} are defined as $a_{21} = f_{21}(x_{1s}, x_{3s})$, $a_{23} = f_{23}(x_{1s})$, $a_{41} = f_{41}(x_{1s}, x_{3s})$, $a_{43} = f_{43}(x_{1s})$, where x_{1s}, x_{3s} are the elements of the stable state vector of the object $\mathbf{x}_s = (x_{1s}, \dots, x_{4s})^T$ in a bounded neighborhood; $K_1 = -T_m/m$, $K_2 = T_m/J$ are scale coefficients; $q_i = (k-i)\Delta z + dz$, $i = \overline{1, N}$ is remoteness of the i -th force application point from the rotation point (the large-sized object mass center); the elements of the matrix \mathbf{C} correspond to the structure of the measurement system [7]. The values of the components of the matrices \mathbf{A} , \mathbf{B} and \mathbf{C} , can change during the process of moving the object, depending on the conditions of the system functioning, preserving its structure. The values of the parameters of the matrices depending on the external and internal operating conditions of the system may change, which necessitates adjusting the values of the model parameters to maintain its adequacy. Thus, in connection with the changes in the parameters of the external environment on the trajectory of motion, it is necessary to periodically initiate a procedure for identifying model parameters. To identify the parameters of the moving process model in real time, it is advisable to use recurrent procedures that make it possible to obtain an estimate of the model parameters when new dimensions arrive [8]. Recurrent evaluation procedures are determined by the dependence

$$P[k+1] = P[k] + \gamma[k+1] \cdot f(P[k], y[k+1], u[k+1]), \quad (5)$$

where $P[k]$ is the current estimate of the parameter; $\gamma[k]$ – the weight coefficient; f is some function that depends on the current value $P[k]$ and determines the magnitude and direction of the next step; $y[k+1]$ and $u[k+1]$ are the output and input signals that follow after the current value.

The most famous recurrent procedures are the stochastic approximation method and the least squares method [9]. Since the accuracy of the stochastic approximation method can only be considered for $t \rightarrow \infty$, according to [8], therefore, it cannot be used when identifying model parameters for solving problems of controlling the large-sized object movement in real time.

III. AIM OF THE RESEARCH

The aim of the research is to analyze the effectiveness and expediency of applying recurrent procedures for identifying the parameters of a linearized large-sized object moving process model.

IV. MAIN PART

When new measured information about the control object at discrete moments of time is received, it is expedient to use models in the state space of the form

$$\mathbf{x}_m[k] = \mathbf{A}[k-1] \cdot \mathbf{x}_m[k-1] + \mathbf{B} \cdot \mathbf{u}[k-1]. \quad (6)$$

The adequacy of the object model is estimated on the basis of the values deviation analysis of the state variables obtained from the model $\mathbf{x}_m[k]$ and as a result of measurements $\mathbf{x}_o[k]$, that is by the value of the error

$$\mathbf{e}[k] = \mathbf{x}_o[k] - \mathbf{x}_m[k]. \quad (7)$$

The use of the classical least squares method is given in [10], however, for real-time identification, it is more expedient to use the recurrent procedure [8]. In this case the algorithm for setting parameters has the form

$$\mathbf{A}[k] = \mathbf{A}[k-1] + \Gamma[k] \cdot \mathbf{e}[k] \cdot \mathbf{x}_m^T[k], \quad (8)$$

where $\Gamma[k]$ – is the matrix obtained on the basis of real values of the state vector measured over the entire observation interval $t \in [t_0, t_f]$, which can be determined recurrently with the aid of relation

$$\Gamma[k] = \Gamma[k-1] - \Gamma[k-1] \cdot \mathbf{x}_o[k] \cdot \gamma[k-1] \cdot \mathbf{x}_o^T[k] \cdot \Gamma[k-1], \quad (9)$$

where $\gamma[k-1] = [1 + \mathbf{x}_o^T[k] \cdot \Gamma[k-1] \cdot \mathbf{x}_o[k]]^{-1}$.

To implement the least squares method, it is necessary to specify the initial values of the state vector components of the model $\mathbf{x}_m[0] = \mathbf{x}_o[0]$, and also the matrix components initial values $\mathbf{A}[0]$, for the known dynamics of the control

vectors $\mathbf{u}[k]$ and the state of the object $\mathbf{x}_o[k]$ for $k=1,2,3,\dots,N$. The initial values of the matrix $\mathbf{\Gamma}[0]$ are chosen as $\mathbf{\Gamma}[0]=\mathbf{I}\cdot(1/\alpha)$, where α is a numerical coefficient whose value influence the convergence of the identification algorithm. The end of the identification phase occurs with small deviations of the matrix \mathbf{A} components values, that is provided $|a_{ij}[k]-a_{ij}[k-1]|<\delta$, for all i and j .

V. EXPERIMENTS

Simulation of the identification process using the recurrent least squares method was carried out at the initial values of the state vector of the object and the initial values of the matrix \mathbf{A} components in the form

$$\mathbf{x}_o[0]=\begin{bmatrix} 0 \\ 0,05 \\ 0 \\ 0 \end{bmatrix}, \quad \mathbf{A}[0]=\begin{bmatrix} 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 1 & 0 & 1 & 0 \end{bmatrix}.$$

The change in the values of the matrix \mathbf{A} components during the calculations is shown in Figure 2.

As a result of the calculations, a matrix \mathbf{A} of the following form obtained

$$\mathbf{A}[350]=\begin{bmatrix} 0,98 & 1,02 & 0,14 & 5,63\cdot 10^{-5} \\ -0,12 & -0,03 & 0,85 & -6,75\cdot 10^{-5} \\ 0,15 & 2,95\cdot 10^{-3} & 0,02 & 1 \\ -0,13 & -0,03 & 0,85 & -7,22\cdot 10^{-5} \end{bmatrix}.$$

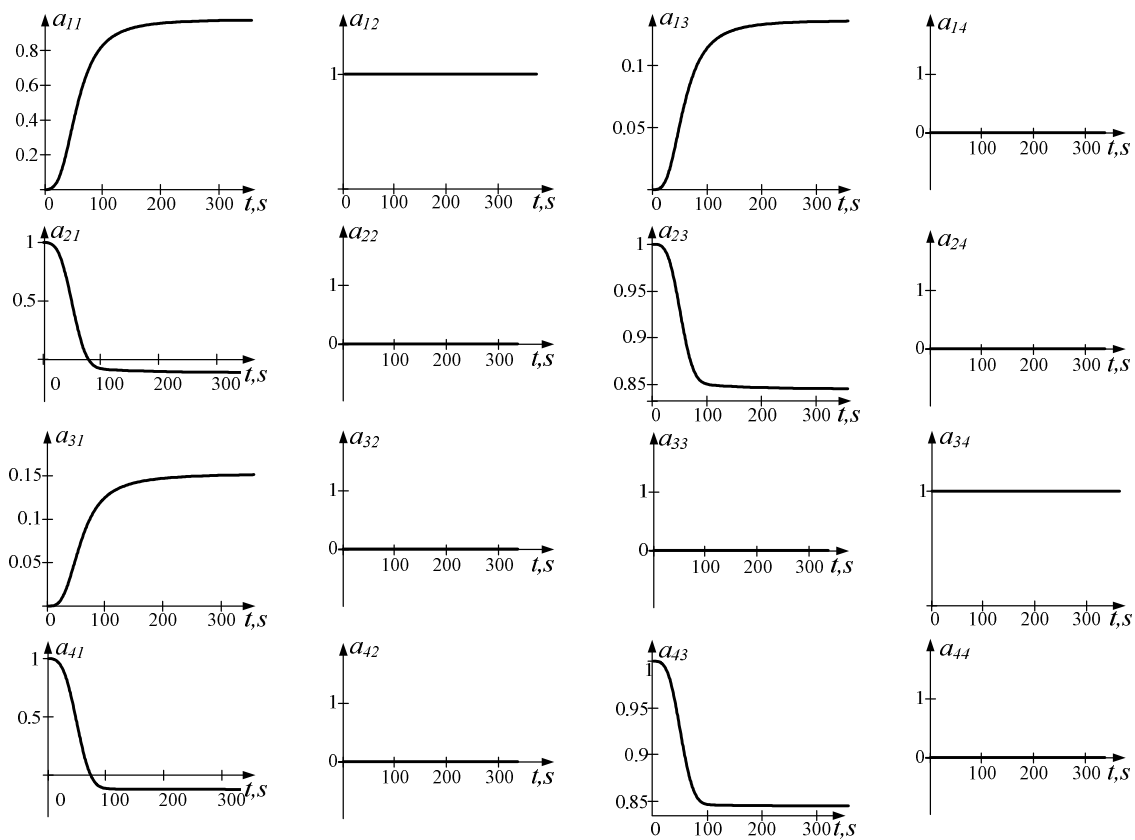


Fig. 2. Dynamics of adjusting the values of matrix \mathbf{A} components

Taking into account the permissible error of 5%, the resulting matrix \mathbf{A} can be written as

$$\mathbf{A}^* = \begin{bmatrix} 0,98 & 1 & 0,14 & 0 \\ -0,12 & 0 & 0,85 & 0 \\ 0,15 & 0 & 0 & 1 \\ -0,13 & 0 & 0,85 & 0 \end{bmatrix}.$$

The best results are obtained if use the algorithm setting factor $\gamma=150$ and $\Delta t=1$ s. The time of the end of the identification stage by the method of recurrent stochastic approximation was $t=100$ s, which is admissible, since only

with $t > 300$ s the external conditions of motion of the large-sized object under consideration change and the determination of new matrix \mathbf{A} values is required. The graphs of the state vector components dynamics of the object (line 1) and the model (line 2) are shown in Fig. 3.

Graphs of the error dependence $\Delta x_i(t) = x_o(t) - x_m(t)$ are shown in Fig. 4. When changing the external conditions for a large-sized object moving, it is necessary to reactivate the identification process again. The activation condition of the identification phase is $|x_o(t) - x_m(t)| > \varepsilon$.

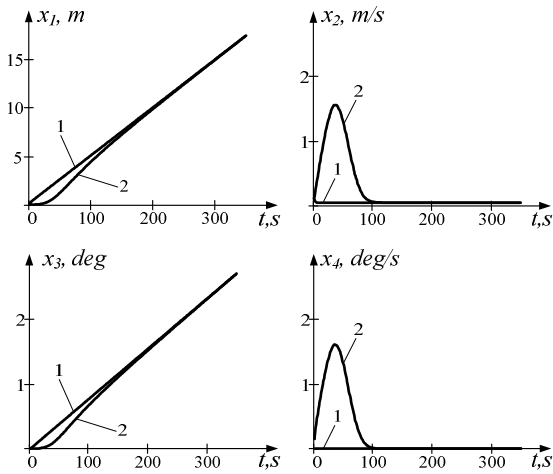


Fig. 3. Graphs of dynamics of the states vector components

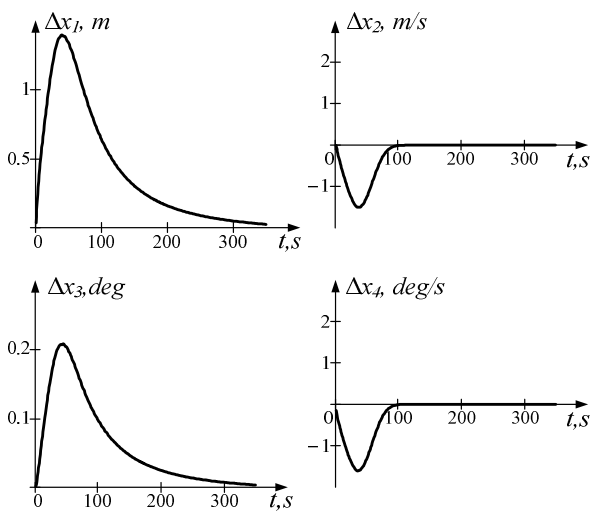


Fig. 4. The graphs of the dynamics error

VI. CONCLUSIONS

Using of least squares method allows getting an adequate estimation of both motion parameters and the model coefficients in real time. The timely identification of the process of moving bulky object is a necessary condition for the effective functioning of the adaptive control system of ship's moving at the slipway. Adaptive control methods allow synthesizing control system whose motion trajectory approaches to a given.

REFERENCES

- [1] A. Omelchuk, U. Lebedenko and G. Rudakova, "Problem of coordinate control of complex electromechanical system," Scientific works national university of food technologies, vol 49, pp. 19 – 23, 2013. (in Ukrainian)
- [2] J. Mendes, R. Araújo and F. Souza, "Adaptive fuzzy identification and predictive control for industrial processes," Expert Systems with Applications, vol 40, pp. 6964-6975, 2013.
- [3] Santosh Ku. Behera and Debaraj Rana, "System Identification Using Recurrent Neural Network," International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering, vol 3, pp. 8111– 8117, 2014.
- [4] A. A. Omelchuk, U. A. Lebedenko and A. V. Rudakova, "Simulation of the slipway cradle motion," Bulletin of KNTU, vol 2(47), pp. 265–270, 2013. (in Russian)
- [5] P. Albertos and A. Sala, Multivariable control systems: an engineering approach. London: Springer, 2004.
- [6] D. P. Kim, Automatic control. Theory Nonlinear and Multivariable System. Seoul: Harnol, 2000.
- [7] A. A. Omelchuk, G. V. Rudakova and O. V. Polivoda, "Optimal control of the ship's motion at the cross slipway," Odes'kyi Politechnichnyi Universytet. Pratsi. no 3(47). pp. 75 – 84. 2015.
- [8] V. N. Kirichkov, Object identification process control systems. Kiev: Vyscha shkola, 1990. (in Russian)
- [9] P. Eykhoff, System identification: parameter and state estimation. Chichester, England: Wiley, 1974.
- [10] G. V. Rudakova, O. V. Polyvoda and A. A. Omelchuk, "Adaptive Control System of the Ship's Motion at the Cross Slipway", IEEE 4th International Conference Methods and Systems of Navigation and Motion Control (MSNMC), pp. 162-165, 2016.

Robust Approach to Estimation of the Intensity of Noisy Signal with Additive Uncorrelated Impulse Interference

Andriy Lozynskyy
Karpenko Phisico-Mechanical Institute
NAS of Ukraine
Lviv, Ukraine
lozynskyy@ah.ipm.lviv.ua

Igor Romanyshyn
Karpenko Phisico-Mechanical Institute
NAS of Ukraine
Lviv, Ukraine
romanyshyn@ipm.lviv.ua

Bohdan Rusyn
Karpenko Phisico-Mechanical Institute
NAS of Ukraine
Lviv, Ukraine
rusyn@ipm.lviv.ua

Volodymyr Minialo
Karpenko Phisico-Mechanical Institute
NAS of Ukraine
Lviv, Ukraine
minialo@ipm.lviv.ua

Abstract – A robust approach to estimation the intensity of a noisy signal with additive uncorrelated impulse interference is proposed. An occurrence of the additive uncorrelated impulse interference leads to increasing of the observed signal dispersion within some sections with impulse interference. Robustness of the intensity estimation is achieved by decreasing the influence of sections with impulse interference. A number of nonlinear filtering methods basing on lower envelope detection are developed: two-parameter recursive filter, dilation filter, clipping derivative filter and filters based on order statistics. Proposed approach was approbated by a numerical simulation. Numerical simulation is validated the efficiency of the proposed approach for estimation the intensity of a noisy signal with additive uncorrelated impulse interference at dynamic data mining and data stream mining.

Keywords – noisy signal, additive uncorrelated impulse interference, random signal parameters estimation, robust method, nonlinear filtering

I. INTRODUCTION

Estimation of quasistationary noisy signal intensity with additive noise is the important problem in dynamic data mining and data stream mining.

If a noisy signal is represented as a set of random variables then a problem of estimation of noisy signal parameters is reduced to a problem of estimation of these random variables parameters [1-15].

A basis for estimation of random variables parameters is probability density function. Mainly, at regular random signal analysis, one is estimated a mean value (amplitude characteristic) and dispersion (energy characteristic).

At analysis of one-dimensional random variables that conforms to a first two moments: mean value and dispersion (location and scale). First two moments are fully characterize a normal distribution. In classical statistics for their estimation are used the arithmetic mean and standard deviation.

The arithmetic mean and standard deviation are very responsive to outliers and, therefore, are unreliable estimations in this case [1, 6]. The most responsive to outliers

is standard deviation. As classical example is the Tukey method, which illustrates the influence of outliers on robustness of standard deviation estimation [2]. It is shown, only two “bad” measurements out of thousand can substantially worse the estimation of standard deviation versus mean absolute deviation [1, 2, 8]. Mean absolute deviation is a robust parameter for estimation of random variable scale [1]. At the same time, the impulse interference is affecting on it also, as on the arithmetic mean.

In presence of outliers, the median is the most insensitive value of location estimation versus mean value [8, 15]. It should be noticed that Kolmogorov paper is one of the first describing the application of median for estimation of random variable location [18].

For estimation a random variables location at outliers occurrence on the base of order statistics, α - trimmed values and α - Winsorized are used [13].

α - trimmed value is obtained, when in variational series α percent of series length are rejected from the beginning and from the end; the arithmetic mean of rested series, in this case, acts as location parameter.

α - Winsorized value is obtained, when in variational series, α percent of series length from the beginning of series are substituted by the least value of rested series, and α percent (also) - from the end of series by the largest value. The arithmetic mean of Winsorized series, in this case, acts as location parameter.

Using methods of ordinal statistics to time series filtering is performing with “window” filtering. In other words, the interval (“window”) is chosen for data filtering and this interval is moving sequentially, sample by sample.

On the other hand, a problem of estimation the intensity of a noisy signal with additive noise is reduced to signal filtering in frequency and time domain [15, 22].

However, modern methods don't provide a good estimation of location and scale with impulse interference.

II. FILTERING METHODS ON THE BASIS OF LOWER ENVELOPE DETECTION

A. Statement of problem

A problem of estimation the intensity of a noisy signal is reduced to estimation the dispersion of random signal with interference.

Dispersion of random signal with additive uncorrelated noise has two components: a desired dispersion of random signal and noise dispersion. Impulse interference leads to increasing a dispersion of the registered signal on interfered interval. Estimation of random signal dispersion is reduced to estimation of random variable location and is achieved by decreasing influence of these intervals.

Let's have a sequence of noisy samples s_{n_i} of signal intensity:

$$s_{n_i} = s_i + \xi_i, \quad \forall i = \overline{1, N}, \quad (1)$$

where $s_i \geq 0$ - utility noisy signal, $\xi_i \geq 0$ - random interference with random occurrence time and random duration.

The problem of estimation the mean value (location) of a random variable is set:

$$\bar{s} = \frac{1}{N} \sum_{i=1}^N s_i \quad (2)$$

on basis of noisy samples (1).

It's clear that averaging of s_{n_i} , leads to:

$$\bar{s}_{n_i} = \bar{s} + \bar{\xi} = \frac{1}{N} \sum_{i=1}^N (s_i + \xi_i), \quad (3)$$

that can't be a good estimator.

In case of normal distribution of random variables s_i , ξ_i , distribution of sum (1), usually is labeled as distributions with «heavy-tailed» [1, 16]. Numerous experiments have shown that application of standard methods (on basis of mean value calculation (3), median filtering, calculation of trimmed- average, Winsorization, etc.) is not a good estimator of location (2).

A new approach is proposed for estimation the intensity of a noisy signal with uncorrelated noise that is based on s_{n_i} lower envelope signal detection. It allows to decrease, substantially, the influence of intervals with impulse interference at estimation of random variable location (2). According to the proposed approach, a number of methods were elaborated: two-parameter recursive filter, dilation filter, clipping derivative filter as well as a series of modified methods, based on order statistics (median filtering, trimmed-average, Winsorization).

B. Two-parameter recursive filter

A method that based on nonlinear filtering for estimation of mean value (2) of noisy realization (1) is proposed.

These filters are based on exponential smoothing [20-22] and operation as follows:

$$\begin{aligned} y_1 &= s_1 + \xi_1, \\ y_i &= \begin{cases} k_1(s_i + \xi_i) + (1 - k_1)y_{i-1}, & s_i + \xi_i > y_{i-1} \\ k_2(s_i + \xi_i) + (1 - k_2)y_{i-1}, & s_i + \xi_i \leq y_{i-1} \end{cases} \quad (4) \\ i &= 2, 3, \dots, N \end{aligned}$$

Here $k_1 \sim 0$ - coefficient, near-zero, $k_2 \sim 1$ - coefficient, near-one. Such selection of coefficients means that if current value of the registered signal exceeds a previous filtered value (usually interfered value) then we receive a value that is near to previous (first row of y_i in (4)). In case of noise absence, filtering is performed according to lower envelope. Such filtering method at which i - value of filtering signal is expressed in terms of $i-1$ (with weight coefficient) leads to displacement of sequence to the right on one step. For elimination of such effect is used an inversely filtration - from N - sample to first:

$$\begin{aligned} \tilde{y}_N &= y_N, \\ \tilde{y}_i &= \begin{cases} k_1 y_i + (1 - k_1)\tilde{y}_{i+1}, & y_i > \tilde{y}_{i+1} \\ k_2 y_i + (1 - k_2)\tilde{y}_{i+1}, & y_i \leq \tilde{y}_{i+1} \end{cases} \quad (5) \\ i &= N-1, N-2, \dots, 1 \end{aligned}$$

Such two-stage filtering allows us to eliminate the effect of sequence displacement that appears during one-stage filtering procedure.

It should be marked that single executing of the proposed filtering procedures, may not be enough to filter the data. Therefore, the filtration continues until the desired result is achieved.

C. Dilation method

Dilation is a filtering method, simpler versus two-parameter and it consists in the following:

$$\begin{aligned} y_1 &= s_1 + \xi_1; id = 0; \\ y_i &= \begin{cases} s_i + \xi_i; id = 0; & s_i + \xi_i \leq y_{i-1} \\ y_{i-1}; id = id + 1; & s_i + \xi_i > y_{i-1} \end{cases} \quad (6) \\ i &= 2, 3, \dots, N \end{aligned}$$

As one can see, dilation method is based on replacement of the samples in increasing interval to previous value. The maximal size of this interval is given by the parameter id_{\max} , which is determined on the estimation of the duration of the interference. The parameter id changes to id_{\max} , then reset, filter output is assigned to an current value and the filtration cycle (6) is repeated. Dilation method, as well as two-parameter method, leads to displacement of sequence to the right. Next stage is used for elimination of this effect:

$$\begin{aligned} \tilde{y}_N &= y_N; id = 0; \\ \tilde{y}_i &= \begin{cases} y_i; id = 0; y_i \leq \tilde{y}_{i+1} \\ \tilde{y}_{i+1}; id = id + 1; y_i > \tilde{y}_{i+1} \end{cases} \quad (7) \\ i &= N-1, N-2, \dots, 1 \end{aligned}$$

D. Method of clipping derivative

The present method is based on comparison of current sample with previous one and when difference between them exceed a defined value, current value is substituted by this value.

$$\begin{aligned} y_1 &= s_1 + \xi_1; \\ y_i &= \begin{cases} s_i + \xi_i; (s_i + \xi_i) - (s_{i-1} + \xi_{i-1}) \leq y_{i-1}p; \\ s_i + \xi_i + p; (s_i + \xi_i) - (s_{i-1} + \xi_{i-1}) > y_{i-1}p \end{cases} \quad (8) \\ i &= 2, 3, \dots, N \end{aligned}$$

The parameter p is chosen on the assumption of allowable increasing of signal intensity (1). Trimming of difference is performed adaptively - proportionally to current filter output.

As in the previous methods, reiteration from end to start sequence may be used for elimination the effect of sequence displacement.

E. Methods on basis of order statistics

Order statistics methods are based on forming of variational series during "window" filtering and selection the value near the beginning of variational series as location parameter.

III. NUMERICAL SIMULATION

A noisy signal with interference was simulated as sum of two random variables with a distribution close to the normal and with given mean values. Probability of interference appearance was 0.03, and probability of disappearance – 0.1.

Figure 1 presents an example of simulated signal and results of filtering by two-parameter recursive filter with $k_1 = 0.05$, $k_2 = 0.5$ parameters.

Figure 2 shows a comparison results of two median filters (standard and modified), each of which is based on 17 samples. An input signal was simulated similarly to the previous case, with the probability of interference appearance rising from experiment to experiment from 0 to 0.5 (dotted line), and probability of "disappearance" equal to 0.5.

Here, a horizontal (black) line corresponds to mean value (etalon) that doesn't change from the experiment to experiment and was equal to 2. A green line - to the result of modified median filtering with the fourth value of each "window" was taken into account. And yellow line - to the result of standard median filtering at which the ninth (central) value of each "window" was taken into account.

As one can see, based on the simulation signal, modified filtering is the better way to estimate of the location (1).

Filtering results (location estimation and its standard deviation) versus level of the interference is presented in Table I. It is shown the results of standard and modified

median filtering (SMF and MMF, accordingly) in the following format: (location \pm standard deviation)/etalon for different interference levels.

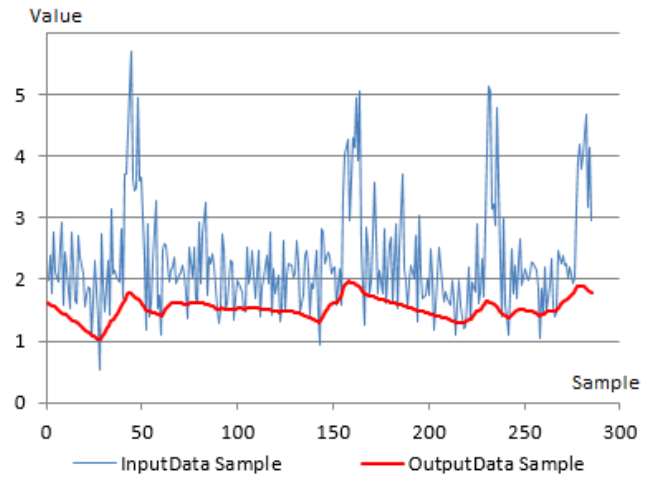


Fig. 1. Simulated signal with interference and filtering result at lower envelope by two-parameter recursive filter

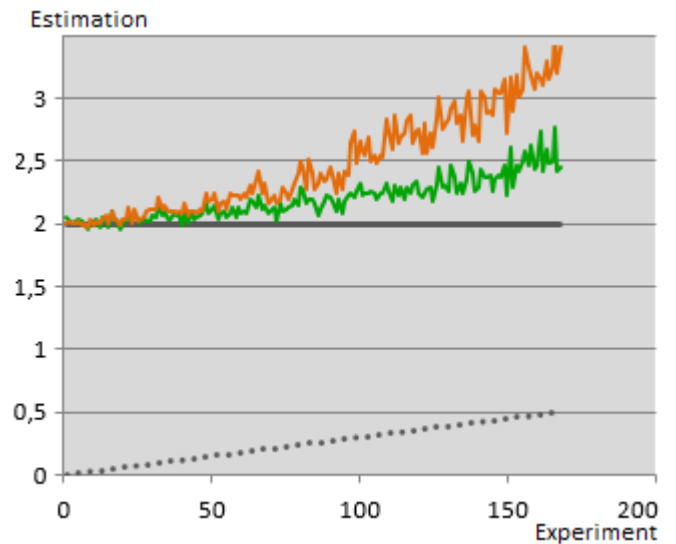


Fig. 2. Comparison of standard and modified median filters

TABLE I. DEPENDENCE OF THE MEDIAN FILTERING RESULTS FOR SOME LEVELS OF INTERFERENCE

Interference level	SMF	MMF
0	1.00 \pm 0.01	1.00 \pm 0.02
1	1.24 \pm 0.03	1.20 \pm 0.03
2	1.46 \pm 0.06	1.25 \pm 0.05
3	1.67 \pm 0.10	1.28 \pm 0.06
4	1.87 \pm 0.11	1.32 \pm 0.07

As one can see, without interference a standard median filtering method gives a less deviation for location estimation. At increasing level of interference, a modified median filtering method, based on the selection of less element, prevails over standard.

Table II shows the results of filtering (location and deviation) using different methods, developed by authors. Input signal was simulated similarly to the previous case, and

probability of the interference appearance and “disappearance” was equal to 0.5 both.

The base of order statistics consists of 17 samples. For the standard median filtering, as usual, a central (ninth) value was selected. For the modified median filtering – the fourth value, trimm–mean was performed basing on first 1-3 values, Winsorization – on the base of 2-4 values. Derivative clipping parameter was 0.07 and dilation parameter value was 11.

TABLE II. FILTERING METHODS COMPARISON

Method	Location/etalon	Deviation/etalon
Median filter	1.688	0.094
Recursive filter	1.331	0.043
Dilation	1.210	0.059
Derivation crop	1.185	0.054
Trimm-filter	1.212	0.045
Winsorization	1.262	0.055
Modified median filtering	1.293	0.059

As one can see, a standard median filter is substantially yield to developed filtering methods, based on lower envelope detection.

IV. CONCLUSIONS

A new approach to estimation of the noisy signal intensity with additive uncorrelated impulse interference in the field of dynamic data mining and data stream mining, based on filtering using lower envelope detection, is proposed. For this, a number of nonlinear filtering methods are developed, namely: two-parameter recursive filter, dilation filter, clipping derivative filter and filters based on order statistics.

Results of numerical experiments are indicated about insufficient applicability of standard filtering methods of noisy signals of considered class and appreciable benefit (on accuracy) of developed methods, based on the proposed approach, versus standard.

REFERENCES

[1] S. A. Ayyvazyan, I. S. Enyukov, L. D. Mehsalkin, Applied statistics: Rudiments of simulation and data preprocessing. M.:Finances and Statistics, 1983.

[2] P. J. Huber, Robust statistics. M.: Mir, 1984. (In Russian)

[3] F. R. Hampel, E. M. Ronchetti, P. J. Rousseeuw, and W. A. Stahel, Robust statistics. M.: Mir, 1989. (In Russian)

[4] J. W. Tukey, A survey of sampling from contaminated distributions. In: Contributions to Prob. and Statist. (Ed. Olkin I. et al.). Stanford Univ. Press. 1960, pp. 448–485.

[5] A. W. F. Edwards, “Three Early Papers on Efficient Parametric Estimation,” Statistical Science, vol. 12, no. 1, pp. 35-47, 1997.

[6] G. Shevlyakov, and P. Smirnov, “Robust Estimation of the Correlation Coefficient: an Attempt of Survey,” Austr. J. of Statistics, vol. 40, no.1&2, pp. 147-156, 2011.

[7] P. O. Smirnov, Robust methods and algorithms of estimation the correlation data characteristics on the basis of new high-performance and rapid robust scale estimations. (Candidate dissertation). St. Petersburg, 2013.

[8] C. Croux, and C. Dehon, Robust estimation of location and scale. Encyclopedia of Environmetrics, A.-H. El-Shaarawi and W. Piegorisch (eds). John Wiley & Sons Ltd: Chichester, UK, Retrieved from 2013. https://feb.kuleuven.be/public/u0017833/PDF-FILES/Croux_Dehton5.pdf.

[9] G. E. P. Box, “Non-Normality and Tests on Variance” Biometrika, vol. 40, pp. 318–335, 1953.

[10] P. J. Bickel, and E. L. Lehmann, “Descriptive Statistics for nonparametric models. I.,” Introduction. The Annals of Statistics, vol. 3, no. 5, pp. 1038-1044, 1975.

[11] P. J. Bickel, and E. L. Lehmann, “Descriptive Statistics for nonparametric models. II.,” Location. The Annals of Statistics, vol. 3, no. 5, pp. 1045-1069, 1975.

[12] P. J. Bickel, and E. L. Lehmann, “Descriptive Statistics for nonparametric models. III.,” Dispersion. The Annals of Statistics, vol. 4, no. 6, pp. 1139-1158, 1976.

[13] D. E. Tyler, A short course on robust statistics. Retrieved from <http://www.rci.rutgers.edu/~dtyler/ShortCourse.pdf>.

[14] P. J. Rousseeuw, and C. Croux, “Alternatives to the Median Absolute Deviation,” Journal of the American Statistical Association, vol. 88, 424, pp. 1273-1283, 1993.

[15] Christophe Leys, Christophe Ley UGent, Olivier Klein, Philippe Bernard and Laurent Licata, “Detecting outliers: Do not use standard deviation around the mean, use absolute deviation around the median,” Journal of Experimental Social Psychology, vol. 49(4), pp.764-766, 2013. Retrieved from <http://dx.doi.org/10.1016/j.jesp.2013.03.013>.

[16] A. Chakrabarty, “Large Deviations for Truncated heavy-tailed random variables: a boundary case,” Indian J. Pure Appl. Math., vol.48 (4), pp. 671-703, 2017.

[17] R. A. Fisher, “On the Mathematical Foundations of Theoretical Statistics,” Phil. Trans. R. Soc. Lond. A., vol. 222, pp. 309-368, 1992. doi: 10.1098/rsta.1922.0009.

[18] A. N. Kolmogorov, “The method of the median in the theory of errors,” Mathematical collect., vol.38, no. 3-4, pp. 47-50, 1931.

[19] A. A. Lyubushin, Analysis of data from geophysical and environmental monitoring systems. M.: Nauka, 2007.

[20] E. S. Gardner, Exponential smoothing: the state of the art. Part II. Houston, 2005.

[21] Yu. S. Dodonov, and Yu. A. Dodonova, “Stable measures of central tendency: weighing as probable alternative of data truncation at the response time analysis,” Psychological researches, vol. 5(19), pp. 1–14, 2011. Retrieved from <http://psystudy.ru>.

[22] Predicting time series using exponential smoothing. Retrieved from <https://www.mql5.com/ru/articles/346>.

Using Stacking Approaches for Machine Learning Models

Bohdan Pavlyshenko
SoftServe, Inc., Ivan Franko National University of Lviv
Lviv, Ukraine
bpavl@softserveinc.com, b.pavlyshenko@gmail.com

Abstract—In this paper, we study the usage of stacking approach for building ensembles of machine learning models. The cases for time series forecasting and logistic regression have been considered. The results show that using stacking technics we can improve performance of predictive models in considered cases.

Index Terms—machine learning, stacking, forecasting, classification, regression

I. INTRODUCTION

One of effective approaches in machine learning classification and regression problems is stacking. The main idea of stacking is using predictions of machine learning models from the previous level as input variables for models on the next level. Using multilevel models with stacking approach is very popular among the participants of Kaggle [1] community. On Kaggle platform, different business companies propose their problems with datasets for data scientists competitions to develop predictive models with the best accuracy. Time series can be analysed by different approaches such as ARIMA, linear models, machine learning models [2].

In this study, we consider the applying of stacking approach to predictive models for time series and for logistic regressions.

II. USING LINEAR REGRESSION FOR MODELS STACKING

We are going to consider several simple cases of approaches in the sales times series forecasting. For our study, we used the data set from Kaggle competition 'Rossmann Store Sales' [3]. Combining different predictive models with different sets of features into one ensemble, one can improve the result accuracy. There are two main approaches for model ensembling - bagging and stacking. Bagging is a simple approach when we use weighted blending of different model predictions. Such models use different types of classifiers with different sets of features and meta parameters. If forecasting errors of these models have weak correlation, then these errors will be compensated by each other under the weighted blending. The less is the error correlation of model results, the more precise forecasting result we will receive. Let us consider the stacking technic [4] for building ensemble of predictive models. In such an approach, the results of predictions on the validation set are treated as input regressors for the next level models. As the next level model, we can consider a linear model or

another type of a classifier, e.g. Random Forest classifier or Neural Network. In our study, linear regression and machine learning models were from scikit-learn python package, neural network was from Keras python package. It is important to mention that in case of time series prediction, we cannot use a conventional cross validation approach, we have to split a historical data set on the training set and validation set by using time splitting, so the training data will lie in the first time period and the validation set - in the next one. Fig. 1 shows the time series forecasting on the validation sets obtained using different models. Predictions on the validation sets are treated as regressors for the linear model with Lasso regularization. Fig. 2 shows the results obtained on the second-level with linear regularized model. Only two models from the first level (gradientBoosting and ExtraTree) have non zero coefficients for their results. For other cases of sales datasets, the results can be different and the other models from the first level can play more essential role in the forecasting.

III. SALES TIME SERIES FORECASTING

The company Grupo Bimbo organized Kaggle competition 'Grupo Bimbo Inventory Demand' [5]. In this competition, Grupo Bimbo invited Kagglers to develop a model to forecast accurately the inventory demand based on historical sales data. I had a pleasure to be a teammate of a great team 'The Slippery Appraisals' which won this competition among nearly two thousand teams. We proposed the best scored solution for sales prediction in more than 800,000 stores for more than 1000 products. Our first place solution can be found at [6]. To built our final multilevel model, we exploited AWS server with 128 cores and 2Tb RAM. For our solution, we used a multilevel model, which consists of three levels (Fig. 3). We built a lot of models on the 1st level. The training method of most 1st level models was XGBoost. On the second level, we used a stacking approach when the results from the first level classifiers were treated as the features for the classifiers on the second level. For the second level, we used ExtraTrees classifier, the linear model from Python scikit-learn and Neural Networks. On the third level, we applied a weighted average to the second level results. The most important features are based on the lags of the target variable grouped by factors and their combinations, aggregated features (min, max, mean, sum) of target variable grouped by factors and their combinations,

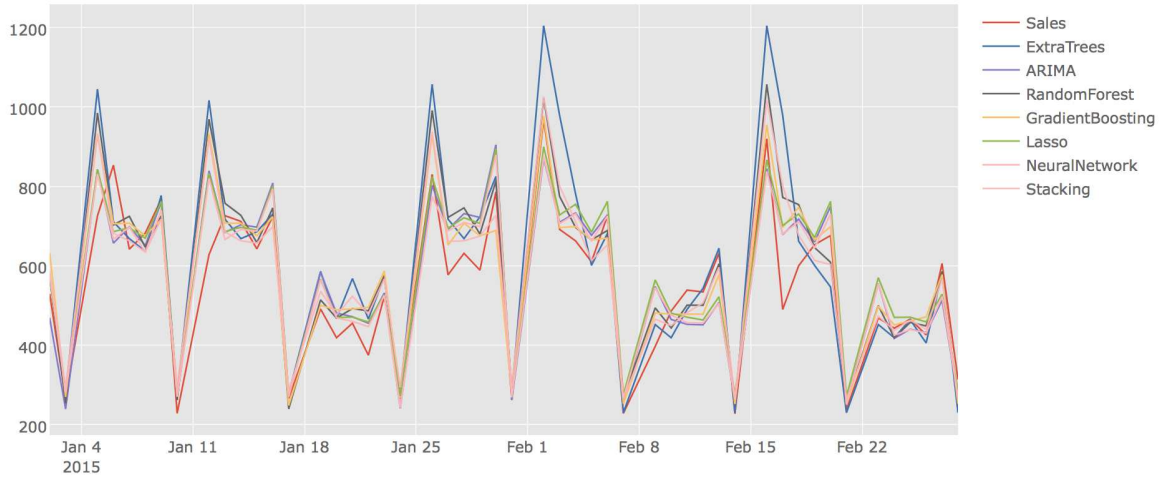


Fig. 1. Different methods for time series forecasting

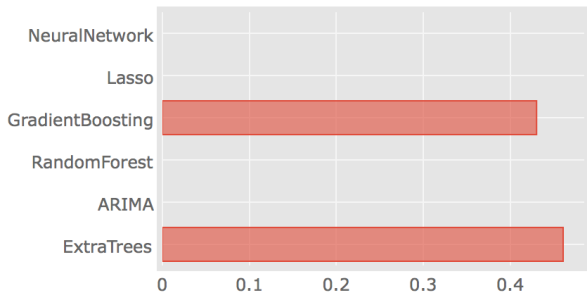


Fig. 2. Coefficients for stacking linear regression

frequency features of factors variables. One of the main ideas in our approach is that it is important to know what were the previous week sales. If during the previous week too many products were supplied and they were not sold, next week this product amount, supplied to the same store, will be decreased. So, it is very important to include lagged values of target variable as a feature to predict next sales. More details about our team's winner solution are at [6]. The simplified version of the R script is at [8]. Our winner solution may seem to be too complicated, but our goal was to win the competition and even a small improvement in forecasting score required essential numbers of machine learning models in the final ensemble. Real business cases with a sufficient accuracy can be simpler.

IV. STACKING APPROACH FOR LOGISTIC REGRESSION

Let us consider using stacking approach for logistic regressions problems. In the Kaggle competition 'Bosch Production Line Performance' [11], the problem of internal failures on assembly lines was considered. The data set consists of measurements for components on assembly lines. This case is a type of logistic regression problem with highly imbalanced classes. The problem lies in predicting which parts will fail a quality control. In the work [12], the logistic regression

approach in manufacturing failure detection was considered. As a data set for the analysis, we used the data from Kaggle competition 'Bosch Production Line Performance' [11]. The data set has a lot of anonymized features. For modeling we used linear, machine learning and Bayesian approaches. To find influence of different factors we exploited the generalized linear model. Using Bayesian approach for logistic regression, we can get the probability distribution function for model parameters. Having statistical distribution we can make risk assessments. To build machine learning models, we used XGBoost classifier from R package 'xgboost' [7], [9], [10]. The data in this set have highly imbalanced classes. To reduce this problem we used undersampling approach. The samples with positive value 1 for target variable were retained without changes. The samples with value 0 for target variable were randomly sampled, so the total number of data items was reduced. For categorical features we used one-hot encoding. The results of the classification were the probability for positive responses. Combining machine learning models and linear or Bayesian models on different levels can give us improved results for logistic regression. Fig. 4 shows a diagram of such possible stacking model. On the first level, there are different XGBoost classifiers with different sets of features and subsets of samples. On the second level, probabilities predicted on the first level can be blended with appropriate weights using linear or Bayesian regression. To evaluate classification performance we used Matthews correlation coefficient (MCC):

$$MCC = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

where TP represents the number of true positives, TN represents the number of true negatives, FP represents the number of false positives, and FN represents the number of false negatives. Let us consider the use of generalized linear model for stacking logistic regression with independent variables which are the probabilities predicted by XGBoost models on the first levels.

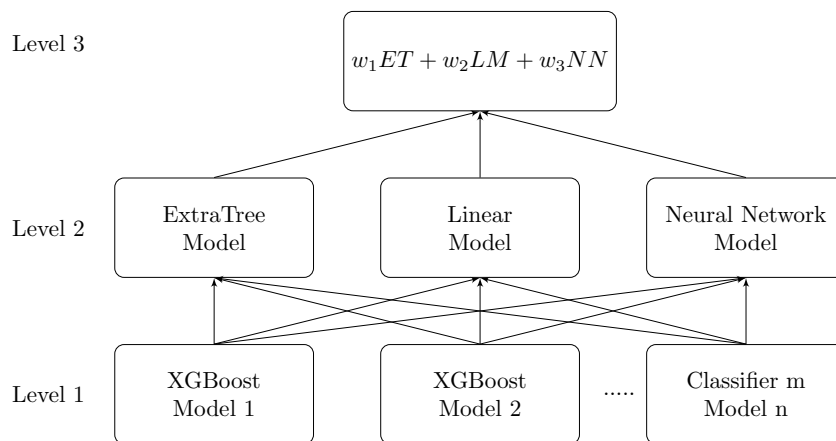


Fig. 3. Multilevel machine learning model for sales time series forecasting

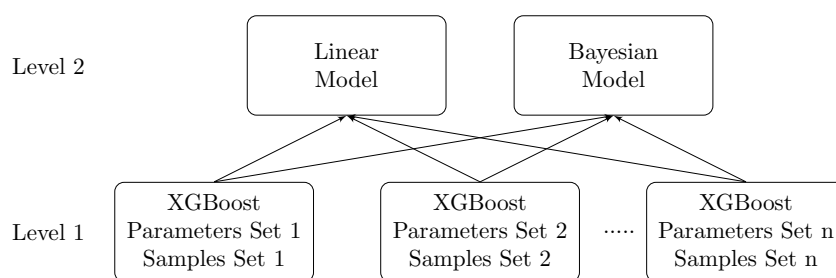


Fig. 4. Stacking model for logistic regression

We used different sets of parameters for 3 XGBoost models, they are - set 1: max.depth = 15, colsample_bytree = 0.7; set 2: max.depth = 5, colsample_bytree = 0.7; set 3: max.depth = 15, colsample_bytree = 0.3. For these 3 models, we used the same subset of samples. Fig. 5, 6 show the dependence of Matthews correlation coefficient from a probability threshold for different subsets of features, where features set 2 is features set 1 with 4 added magic features. So-called magic features which are based on the ID of samples were considered by the participants of the competition at [13]–[15]. For Bayesian models, we used the same 3 subsets of parameters with different subsets of samples. As it was shown above, for different samples subsets, we received slightly different results for Matthews correlation coefficient. These differences can be taken into account using Bayesian model. For Bayesian inference we used JAGS sampling software [16], [17]. We used Bayesian model for logistic regression. As covariates we used the probabilities predicted by three XGBoost models. Fig. 7 shows the boxplots for coefficients of probabilities predicted by different XGBoost models.

V. CONCLUSION

In our study, we considered stacking approaches for time series forecasting and logistic regression with highly imbalanced data. Using multilevel stacking models, one can receive

more precise results in comparison with single models. For stacking machine learning models the linear regression with Lasso regularization, other machine learning model, Bayesian model can be used. Using stacking model on the second level with the covariates that are predicted by machine learning models on the first level, makes it possible to take into account the differences in results for machine learning models received for different sets of parameters and subsets of samples. As obtained results show, using stacking approach for machine learning models we can improve performance of predictive models.

REFERENCES

- [1] Kaggle: Your Home for Data Science. URL: <http://kaggle.com>
- [2] B. M. Pavlyshenko. "Linear, machine learning and probabilistic approaches for time series analysis," in IEEE First International Conference on Data Stream Mining & Processing (DSMP), Lviv, Ukraine, pp. 377-381, August 23-27, 2016.
- [3] "Rossmann Store Sales", Kaggle.Com, URL: <http://www.kaggle.com/c/rossmann-store-sales>.
- [4] D. H. Wolpert. "Stacked generalization." *Neural networks*, 5(2), pp. 241-259, 1992.
- [5] Kaggle competition "Grupo Bimbo Inventory Demand" URL: <https://www.kaggle.com/c/grupo-bimbo-inventory-demand>
- [6] Kaggle competition "Grupo Bimbo Inventory Demand" #1 Place Solution of The Slippery Appraisals team. URL: <https://www.kaggle.com/c/grupo-bimbo-inventory-demand/discussion/23863>

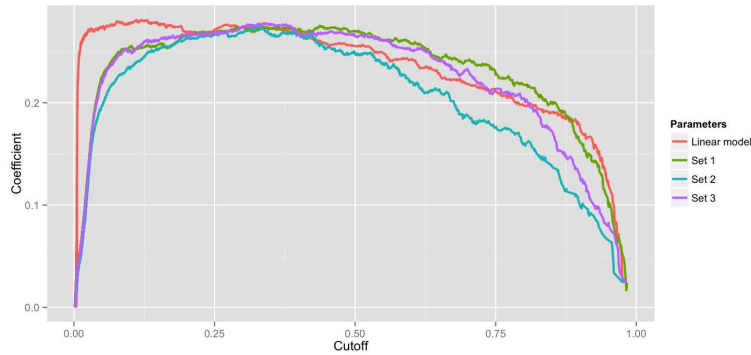


Fig. 5. Matthews coefficient for different XGBoost parameter sets (feature set 1)

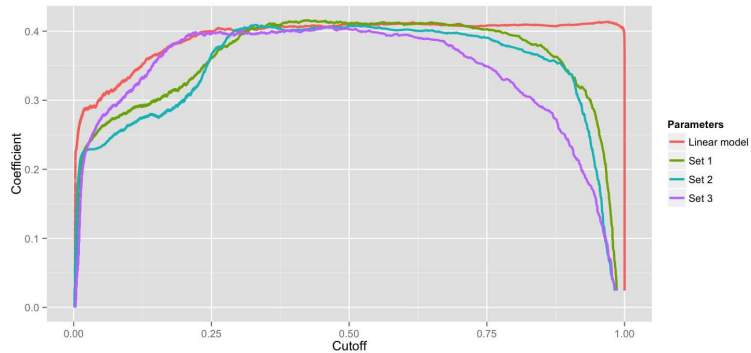


Fig. 6. Matthews coefficient for different XGBoost parameter sets (feature set 2)

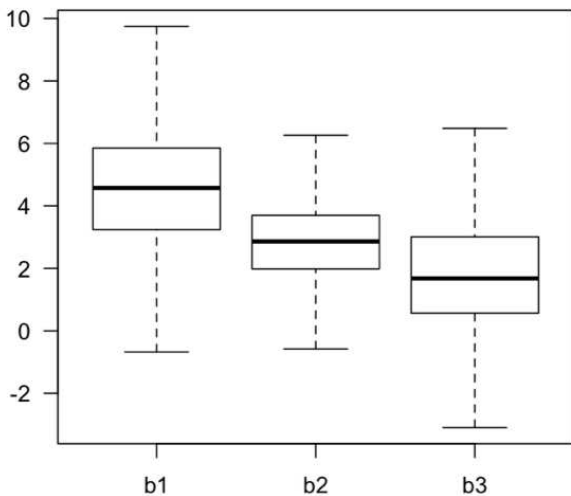


Fig. 7. Boxplots for coefficients of probabilities predicted by different XGBoost models

[7] T. Chen and C. Guestrin. “Xgboost: A scalable tree boosting system.” In Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining. ACM, 2016, pp. 785-794.

[8] Kaggle competition “Grupo Bimbo Inventory Demand” Bimbo XGBoost R script LB:0.457. URL: <https://www.kaggle.com/bpavlyshenko/bimbo-xgboost-r-script-lb-0-457>

[9] J. Friedman. “Greedy function approximation: a gradient boosting machine.”, *Annals of Statistics*, 29(5):1189-1232, 2001.

[10] J. Friedman. “Stochastic gradient boosting.”, *Computational Statistics &*

Data Analysis, 38(4):367-378, 2002.

[11] Kaggle competition “Bosch Production Line Performance”. URL: <https://www.kaggle.com/c/bosch-production-line-performance>

[12] B. Pavlyshenko. “Machine learning, linear and bayesian models for logistic regression in failure detection problems.” in IEEE International Conference on Big Data (Big Data), Washington D.C., USA, pp. 2046-2050, December 5-8, 2016.

[13] Kaggle competition “Bosch Production Line Performance”. The Magical Feature : from LB 0.3- to 0.4+. URL:<https://www.kaggle.com/c/bosch-production-line-performance/forums/t/24065/the-magical-feature-from-lb-0-3-to-0-4>

[14] Kaggle competition “Bosch Production Line Performance”. Road-2-0.4+. URL:<https://www.kaggle.com/mmueller/bosch-production-line-performance/road-2-0-4>

[15] Kaggle competition “Bosch Production Line Performance”. Road-2-0.4+ ->FeatureSet++. URL: <https://www.kaggle.com/alexanderlarko/bosch-production-line-performance/road-2-0-4-feature-set>

[16] John Kruschke. *Doing Bayesian data analysis: A tutorial with R, JAGS, and Stan*. Academic Press, 2014.

[17] Martyn Plummer. *JAGS Version 3.4.0 user manual*. URL:http://sourceforge.net/projects/mcmcjags/files/Manuals/3.x/jags_user_manual.pdf

Modeling of Thermoviscoelasticity Time Harmonic Variational Problem for a Thin Wall Body

Romanna Malets
Department of Programming
Ivan Franko National University of Lviv
Lviv, Ukraine
romannakhmil@yahoo.com

Igor Malets
Department of Project Management, Information
Technologies and Telecommunications
Lviv State University of Life Safety
Lviv, Ukraine
igor.malets@gmail.com

Heorgiy Shynkarenko
Department of Information System
Ivan Franko National University of Lviv
Lviv, Ukraine
kis@lnu.edu.ua
Opole University of Technology
Opole, Poland
h.shynkarenko@gmail.com

Petro Vahin
Department of Information System
Ivan Franko National University of Lviv
Lviv, Ukraine
ppvahin@gmail.com

Abstract—The paper presents the construction and analysis of vibration problem of thermoviscoelastic shells under the influence of non-stationary heat and under forced loads. The studied model was based on application of simplest finite element semidiscretization to mixed variational problem of dynamical thermoviscoelasticity. The problem in addition to the mutual influence of temperature field and stress field is also taken into account the viscoelastic properties of the material thin wall body. For assumptions quite suitable for applications we prove the well-posedness for this model of time harmonic vibrations.

Keywords—initial-boundary value problem, thermo-viscoelasticity, material with short-term memory, variational formulation, semidiscretization, well-posedness of problem, Galerkin discretization.

I. INTRODUCTION

Mathematical modeling methods of thin-walled structures that are under forced, temperature and electromagnetic loads are wide tools base of continuum mechanics and its engineering applications.

Last time, well developed analytical methods for solving this class of problems are actively complemented by methods of computational mathematics and computer simulation, the successful application of which often requires revision and supplementation of classical models, for example, shell theory, developing appropriate software. The filling of mechanics with an intensive influx of engineering problems, for example, with smart materials, makes studies in this field relevant and timely.

In authors' previous articles [2] a development and analysis of the dimension reducing methods for heat conduction problem and thermoelasticity problem for thin flexible bodies have been investigated.

In work [6] theory of thermoviscoelastic thin wall elements for dynamical problems was considered. In this article, similar techniques as in [7] are applied to the problem of forced vibrations of thermoviscoelastic shells [7].

II. PROBLEM STATEMENT

Let be the bounded connected domain $D \in \mathfrak{R}^n$ of points $\mathbf{x} = (x_1, x_2, \dots, x_n)$ with Lipschitz-continuous boundary $\partial D = S$, and $\mathbf{n} = \{n_i\}_{i=1}^n$ is unit outer normal vector $n_i = \cos(\mathbf{n}, x_i)$. Also let us consider time interval $[0, T]$, $0 < T < +\infty$. Notation $\{F_i(\mathbf{x}, t)\}_{i=1}^3$ is a vector of volume mechanical forces, a vector of surface mechanical loads $\hat{\sigma} = \{\hat{\sigma}_i(\mathbf{x}, t)\}_{i=1}^3$ on the boundary $S_\sigma \subset S$, represents volume heat forces $g = g(\mathbf{x}, t)$. Like in classic thermoelasticity problem, our goal is to find vector of elastic displacements $\mathbf{U} = \{U_i(\mathbf{x}, t)\}_{i=1}^3$ and temperature increment $\theta(\mathbf{x}, t)$, which satisfy the following equations in $D \times (0, T]$ (here and everywhere below the ordinary summation by repetitive indices is expected) [2],[3]:

$$\rho U_i'' - \partial_k \sigma_{ki} = \rho F_i, \quad (1)$$

$$c_e \theta' - \partial_i (\lambda_{ij} \partial_j \theta) + \theta_0 \beta_{ij} \partial_i U_j' = g, \quad (2)$$

The above expressions (1)-(2) are equation of motion, heat conduction equation, where $\partial_i := \partial v / \partial x_i$, $v' := \partial v / \partial t$, $v'' := \partial_i (\partial_i v)$. Below we will explain the meaning of each

notation more thoroughly. Here $\sigma = \{\sigma_{ij}\}_{i,j=1}^n$ is a stress tensor, which is defined by the following constitutive equation, namely hypothesis Duhamel-Neumann for material with short-term memory:

$$\begin{aligned} \sigma_{ij}(\mathbf{U}, \theta) &:= \sigma_{ij}^e(\mathbf{U}) + \sigma_{ij}^v(\mathbf{U}') + \sigma_{ij}^t(\theta) \\ &= c_{ijkm} E_{km}(\mathbf{U}) + a_{ijkm} E_{km}(\mathbf{U}') - \beta_{ij} \theta, \end{aligned} \quad (3)$$

$$\begin{aligned}
& \frac{1}{2} \left[\|\mathbf{U}'(t)\|_{\mathbf{H}}^2 + \|\mathbf{U}(t)\|_{\mathbf{Y}}^2 + \|\theta(t)\|_{\mathbf{Z}}^2 \right] + \\
& \int_0^t \left[\|\mathbf{U}'(\tau)\|_{\mathbf{Y}}^2 + \|\theta(\tau)\|_{\mathbf{G}}^2 \right] d\tau = \\
& \frac{1}{2} \left[\|\mathbf{V}_0\|_{\mathbf{H}}^2 + \|\mathbf{U}_0\|_{\mathbf{Y}}^2 + \|\theta_0\|_{\mathbf{Z}}^2 \right] \\
& + \int_0^t \left[\langle l(\tau), \mathbf{V}'(\tau) \rangle + \langle r(\tau), \theta(\tau) \rangle \right] d\tau \quad \forall \tau \in (0, T].
\end{aligned} \tag{11}$$

Here $\frac{1}{2} \left[\|\mathbf{U}'(t)\|_{\mathbf{H}}^2 + \|\mathbf{U}(t)\|_{\mathbf{Y}}^2 + \|\theta(t)\|_{\mathbf{Z}}^2 \right]$ determines the instant total energy value, $\int_0^t \left[\|\mathbf{U}'(\tau)\|_{\mathbf{Y}}^2 + \|\theta(\tau)\|_{\mathbf{G}}^2 \right] d\tau$ determines dissipation of energy was caused by viscosity and temperature field of an elastic body, $\frac{1}{2} \left[\|\mathbf{V}_0\|_{\mathbf{H}}^2 + \|\mathbf{U}_0\|_{\mathbf{Y}}^2 + \|\theta_0\|_{\mathbf{Z}}^2 \right]$ initial energy value, $\int_0^t \left[\langle l(\tau), \mathbf{V}'(\tau) \rangle + \langle r(\tau), \theta(\tau) \rangle \right] d\tau$ an influx of energy.

Formulated in accordance with the problem (1) – (8), the variational task of the dynamic thermoviscoelasticity of an elastic body, taking into account the corresponding linear elastic-viscous properties of the material and the energy balance equation (1.10), will be the basis for investigations of thermoviscoelastic processes in thin-walled bodies.

III. PARTIALLY DISCRETIZED VARIATIONAL PROBLEM OF THERMOVISCOELASTICITY FOR A THIN WALL BODY

Let an elastic body $D \in \mathfrak{R}^3$ referred to fixed curvilinear orthogonal coordinate system $(\alpha_1, \alpha_2, \alpha_3)$ (fig. 1) as follows:

$$\begin{aligned}
D &:= \{\mathbf{r} = (\boldsymbol{\alpha}, \alpha_3) \in \mathbb{R}^3 : \boldsymbol{\alpha} = (\alpha_1, \alpha_2) \in \Omega, \\
&\quad \alpha_3 \in (-\frac{1}{2}h, +\frac{1}{2}h)\} = \Omega \times (-\frac{1}{2}h, +\frac{1}{2}h),
\end{aligned}$$

where thickness $h = const > 0$ is substantially smaller compared to other space dimensions, $h/diam\Omega \ll 1$. The body of such kind we shall name shell, its set $\Omega = \{\mathbf{r} = (\boldsymbol{\alpha}, 0) \in D\}$ will named the middle surface of shell and denote its contour through $\Gamma = \partial\Omega$. In this coordinate system a surface element $d\Omega$ and a volume element dD of the body are defined as:

$$d\Omega = H_1 H_2 d\boldsymbol{\alpha}, \quad dD = H_1 H_2 H_3 d\boldsymbol{\alpha} d\alpha_3 = d\Omega d\alpha_3, \tag{12}$$

$$H_i = A_i(1 + \alpha_3 k_i), \quad H_3 = A_3 \equiv 1, \quad i = 1, 2. \tag{13}$$

Here $A_i = A_i(\boldsymbol{\alpha})$ and $k_i = k_i(\boldsymbol{\alpha})$ – coefficients of the first quadratic form and the principal curvatures of the surface Ω [4]. Notes $\Omega_{\pm} = \Omega \times \{\pm h/2\}$ are facial surfaces and

$\Sigma = \Gamma \times (-h/2, h/2)$ – lateral surface, then $S = \Omega_+ \cup \Omega_- \cup \Sigma$. Assume the surface of the body is divided into parts nonzero measure as follows

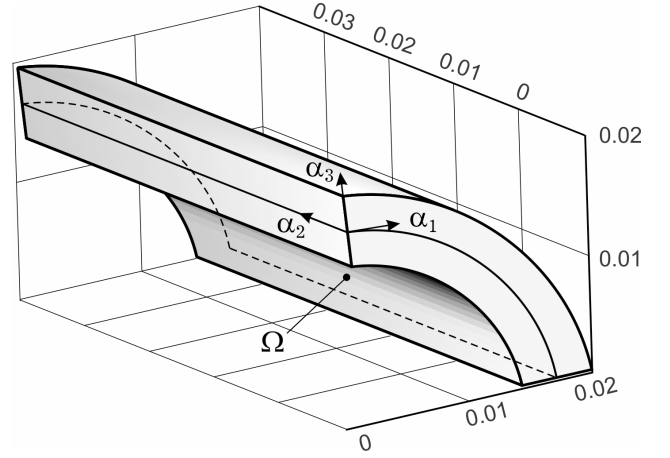


Fig. 1. – Domain D and mid-surface Ω referred to fixed curvilinear coordinate system $(\alpha_1, \alpha_2, \alpha_3)$.

$$S_u = S_\theta = \Sigma := \{\mathbf{r} \in D : \boldsymbol{\alpha} \in \Gamma = \partial\Omega, |\alpha_3| \leq \frac{1}{2}h\},$$

$$S_\sigma = S_q = \Omega_+ \cup \Omega_-, \quad \Omega_{\pm} := \{\mathbf{r} \in \bar{D} : \boldsymbol{\alpha} \in \Omega, \alpha_3 = \pm \frac{1}{2}h\}.$$

By the Timoshenko-Mindlin hypotheses [5] we shall assume that a displacement vector $\mathbf{U} = \{U_i(\mathbf{r}, t)\}_{i=1}^3$ and temperature $\theta = \theta(\mathbf{r}, t)$ can approximated by the linear combinations of a functions $\mathbf{s} = (\mathbf{u}(\boldsymbol{\alpha}, t), \boldsymbol{\gamma}(\boldsymbol{\alpha}, t))$ and $\boldsymbol{\theta} = (\theta_1(\boldsymbol{\alpha}, t), \theta_2(\boldsymbol{\alpha}, t))$ such that

$$\begin{aligned}
\mathbf{U}(\mathbf{r}, t) &\equiv \mathbf{u}(\boldsymbol{\alpha}, t) + \alpha \boldsymbol{\gamma}(\boldsymbol{\alpha}, t), \\
\theta(\mathbf{r}, t) &\equiv \theta_1(\boldsymbol{\alpha}, t) + \alpha \theta_2(\boldsymbol{\alpha}, t) \quad \forall (\boldsymbol{\alpha}, \alpha) \in D.
\end{aligned}$$

Here $\mathbf{u} = \{u_i(\boldsymbol{\alpha}, t)\}_{i=1}^3$ and $\boldsymbol{\theta}_1 = \theta_1(\boldsymbol{\alpha}, t)$ are approximations of the displacement vector and temperature on the middle surface,

$$\begin{aligned}
\boldsymbol{\gamma}(\boldsymbol{\alpha}, t) &\equiv \partial \mathbf{U}(\boldsymbol{\alpha}, 0, t) / \partial \alpha_3, \\
\boldsymbol{\theta}_2(\boldsymbol{\alpha}, t) &\equiv \partial \theta(\boldsymbol{\alpha}, 0, t) / \partial \alpha_3, \quad \forall (\boldsymbol{\alpha}, t) \in \bar{\Omega} \times [0, T].
\end{aligned}$$

As results of partially discretization after the thickness variable of the problem equations (9) we obtained a variation formulation problem for thermoelastic shells in the terms of the displacements vector $\mathbf{s} = (\mathbf{s}_1, \mathbf{s}_2) = (\mathbf{u}(\boldsymbol{\alpha}, t), \boldsymbol{\gamma}(\boldsymbol{\alpha}, t))$ and temperature vector $\boldsymbol{\theta} = (\theta_1(\boldsymbol{\alpha}, t), \theta_2(\boldsymbol{\alpha}, t))$:

$$\left\{ \begin{array}{l} \text{given } \mathbf{s}_0 \in W_h, \mathbf{v}_0 \in \mathbf{H}, \boldsymbol{\theta}_0, \mathbf{g} \in \mathbf{Z}, \mathbf{f} \in \mathbf{H}; \\ \text{find } \boldsymbol{\Psi} = \{\mathbf{s}, \boldsymbol{\theta}\} \in L^2(0, T; W_h \times Q_h) \text{ such, as} \\ m_\Omega(\mathbf{s}''(t), \mathbf{v}) + a_\Omega(\mathbf{s}'(t), \mathbf{v}) + c_\Omega(\mathbf{s}(t), \mathbf{v}) \\ - b_\Omega(\boldsymbol{\theta}(t), \mathbf{v}) = \langle l(t), \mathbf{v} \rangle, \\ \Xi_\Omega(\boldsymbol{\theta}'(t), \boldsymbol{\xi}) + \Lambda_\Omega(\boldsymbol{\theta}(t), \boldsymbol{\xi}) \quad \forall t \in (0, T], \\ + b_\Omega(\boldsymbol{\xi}, \mathbf{s}'(t)) = \langle r(t), \boldsymbol{\xi} \rangle \\ m_\Omega(\mathbf{s}'(0) - \mathbf{v}_0, \mathbf{v}) = 0, \quad c_\Omega(\mathbf{s}(0) - \mathbf{s}_0, \mathbf{v}) = 0 \\ \Xi_\Omega(\boldsymbol{\theta}(0) - \boldsymbol{\theta}_0, \boldsymbol{\xi}) = 0 \quad \forall \mathbf{v} \in W_h, \quad \forall \boldsymbol{\xi} \in Q_h. \end{array} \right. \quad (14)$$

We used the follows introduced spaces:

$$W_h = \{\mathbf{w} \in [H^1(\Omega)]^6 : \mathbf{w} = 0 \text{ na } S_u\},$$

$$Q_h = \{\boldsymbol{\xi} \in [H^1(\Omega)]^2 : \boldsymbol{\xi} = 0 \text{ na } S_\rho\}.$$

The bilinear and linear forms are defined as:

$$m_\Omega(\mathbf{s}, \mathbf{v}) = \rho \sum_{i,j=1}^2 \iint_{\Omega} \phi^{i+j-2} \mathbf{s}_i \cdot \mathbf{v}_j A_1 A_2 d\boldsymbol{\alpha},$$

$$a_\Omega(\mathbf{s}, \mathbf{v}) = \iint_{\Omega} (\mathbf{C}\mathbf{s}) \cdot (\tilde{\mathbf{B}}\mathbf{C}\mathbf{v}) A_1 A_2 d\boldsymbol{\alpha},$$

$$c_\Omega(\mathbf{s}, \mathbf{v}) = \iint_{\Omega} (\mathbf{C}\mathbf{s}) \cdot (\mathbf{B}\mathbf{C}\mathbf{v}) A_1 A_2 d\boldsymbol{\alpha}$$

$$\forall \mathbf{s} = (\mathbf{s}_1, \mathbf{s}_2), \quad \mathbf{v} = (\mathbf{v}_1, \mathbf{v}_2) \in W_h,$$

$$b_\Omega(\boldsymbol{\theta}, \mathbf{v}) = \beta \iint_{\Omega} \Phi(\boldsymbol{\theta}) \cdot (\mathbf{C}\mathbf{v}) A_1 A_2 d\boldsymbol{\alpha},$$

$$\Xi_\Omega(\boldsymbol{\theta}, \boldsymbol{\xi}) = \theta_0^{-1} \sum_{i,j=1}^2 \iint_{\Omega} \phi^{i+j-2} \theta_i \xi_j A_1 A_2 d\boldsymbol{\alpha},$$

$$\Lambda_\Omega(\boldsymbol{\theta}, \boldsymbol{\xi}) = \lambda_\Omega(\boldsymbol{\theta}, \boldsymbol{\xi}) + \kappa_\Omega(\boldsymbol{\theta}, \boldsymbol{\xi}),$$

$$\forall \boldsymbol{\theta} = (\theta_1, \theta_2), \quad \boldsymbol{\xi} = (\xi_1, \xi_2) \in Q_h,$$

$$\langle r, \boldsymbol{\xi} \rangle := -\theta_0^{-1} \iint_{\Omega} \left\{ (q^+ + q^-) \xi_1 \right. \quad (15)$$

$$\left. + \frac{h}{2} (q^+ - q^-) ((k_1 + k_2) \xi_1 + \xi_2) \right\} A_1 A_2 d\boldsymbol{\alpha}$$

$$+ \Xi_\Omega(c_\varepsilon^{-1} \mathbf{g}(t), \boldsymbol{\xi}) \quad \forall \boldsymbol{\xi} = (\xi_1, \xi_2) \in Q_h;$$

$$\langle l, \mathbf{v} \rangle := - \sum_{i,j=1}^2 \iint_{\Omega} (\bar{\boldsymbol{\sigma}}^+ + \bar{\boldsymbol{\sigma}}^-) \left\{ [1 + \frac{1}{2} h(1 + k_1 + k_2)] \mathbf{v}_1 \right.$$

$$\left. - \frac{1}{2} h \mathbf{v}_2 \right\} A_1 A_2 d\boldsymbol{\alpha} + m_\Omega(\mathbf{f}(t), \mathbf{v}) \quad \forall \mathbf{v} = (\mathbf{v}_1, \mathbf{v}_2) \in W_h.$$

Here $\mathbf{C} = \{C_{ij}\}_{i,j=1}^6$, $\mathbf{B} = \{B^{ij}(\boldsymbol{\theta})\}_{i,j=1}^{11}$, $\Phi(\boldsymbol{\theta}) = \{\Phi^i(\boldsymbol{\theta})\}_{i=1}^{11}$, β ar

e data presented in [1], heat flux data q^+ , q^- are given on Ω_+, Ω_- , also surface loads $\bar{\boldsymbol{\sigma}}(\mathbf{r}, t)$ are described such as

$$\bar{\boldsymbol{\sigma}}(\mathbf{r}, t) = \{\bar{\sigma}_i(\boldsymbol{\alpha}, \alpha_3, t)\}_{i=1}^3 =$$

$$= \begin{cases} \boldsymbol{\sigma}^+(\boldsymbol{\alpha}, t) = \{\sigma_i^+(\boldsymbol{\alpha}, t)\}_{i=1}^3, & \text{if } \boldsymbol{\alpha} \in \Omega_+, \\ \boldsymbol{\sigma}^-(\boldsymbol{\alpha}, t) = \{\sigma_i^-(\boldsymbol{\alpha}, t)\}_{i=1}^3, & \text{if } \boldsymbol{\alpha} \in \Omega_-. \end{cases}$$

$$\lambda_\Omega(\boldsymbol{\theta}, \boldsymbol{\xi}) = \theta_0^{-1} \sum_{i,j=1}^2 \iint_{\Omega} \lambda \left[\sum_{k=1}^2 \frac{\lambda_k^{i+j-2}}{A_k^2} \frac{\partial \theta_i}{\partial \alpha_k} \frac{\partial \xi_j}{\partial \alpha_k} \right. \\ \left. + (ij - i - j + 1) \phi^{i+j-4} \theta_i \xi_j \right] A_1 A_2 d\boldsymbol{\alpha},$$

$$\kappa_\Omega(\boldsymbol{\theta}, \boldsymbol{\xi}) = \left\{ (\kappa^+ + \kappa^-) \theta_1 \xi_1 \right.$$

$$\left. + (\kappa^+ - \kappa^-) \frac{h}{2} [(k_1 + k_2) \theta_1 \xi_1 + (\theta_1 \xi_2 + \theta_2 \xi_1)] \right\} \iint_{\Omega} A_1 A_2 d\boldsymbol{\alpha},$$

$$\phi^n := \int_{-h/2}^{h/2} \alpha_3^n (1 + \alpha_3 k_1)(1 + \alpha_3 k_2) d\alpha_3, \quad (16)$$

$$\chi_m^n = \int_{-h/2}^{h/2} (\alpha_3)^n \frac{(1 + \alpha_3 k_1)(1 + \alpha_3 k_2)}{(1 + \alpha_3 k_m)^2} d\alpha_3, \quad m = 1, 2.$$

Here κ^+, κ^- are the heat transfer coefficients on the surfaces Ω_+, Ω_- , respectively.

Details of the construction of the problem (14) see [1].

IV. VIBRATION VARIATIONAL PROBLEM STATEMENT

We suppose that the harmonic loadings with angular frequency ω are applied to the thin shell

$$l(t) = l_c \cos \omega t + l_s \sin \omega t, \quad (17)$$

$$r(t) = r_c \cos \omega t + r_s \sin \omega t, \quad \forall t \in (0, T].$$

Then the approximate solutions of problem (14) can be looked for in the form of the following expansions:

$$\mathbf{s}(\boldsymbol{\alpha}, t) = \mathbf{s}_c(\boldsymbol{\alpha}) \cos \omega t + \mathbf{s}_s(\boldsymbol{\alpha}) \sin \omega t, \quad (18)$$

$$\boldsymbol{\theta}(\boldsymbol{\alpha}, t) = \boldsymbol{\theta}_c(\boldsymbol{\alpha}) \cos \omega t + \boldsymbol{\theta}_s(\boldsymbol{\alpha}) \sin \omega t,$$

where $\mathbf{s}_c(\boldsymbol{\alpha})$, $\mathbf{s}_s(\boldsymbol{\alpha})$, $\boldsymbol{\theta}_c(\boldsymbol{\alpha})$, $\boldsymbol{\theta}_s(\boldsymbol{\alpha})$ are the unknown amplitudes of vector of mechanical displacements and temperature respectively.

Substituting expressions (17) and (18) into variational problem (15) and neglecting its initial conditions, we obtain the variational problem for force harmonic vibrations of thermo-elastic thin shell:

$$\left\{ \begin{array}{l} \text{given } \omega > 0, (l_1, l_2, r_1, r_2) \in \mathbf{W}' = \Phi' \times \Phi', \Phi' = W'_h \times Q'_h; \\ \text{find } \boldsymbol{\psi} = \{\mathbf{s}_c, \boldsymbol{\theta}_c, \mathbf{s}_s, \boldsymbol{\theta}_s\} \in \mathbf{W} = \Phi \times \Phi \\ \text{such that } \forall \{\mathbf{v}_c, \boldsymbol{\xi}_c, \mathbf{v}_s, \boldsymbol{\xi}_s\} \in \mathbf{W} \\ -\omega^2 m_\Omega(\mathbf{s}_c, \mathbf{v}_s) + \omega a_\Omega(\mathbf{s}_c, \mathbf{v}_s) + c_\Omega(\mathbf{s}_c, \mathbf{v}_s) \\ \quad - b_\Omega(\boldsymbol{\theta}_c, \mathbf{v}_s) = \langle l_c, \mathbf{v}_s \rangle, \\ -\omega^2 m_\Omega(\mathbf{s}_s, \mathbf{v}_c) + \omega a_\Omega(\mathbf{s}_s, \mathbf{v}_c) + c_\Omega(\mathbf{s}_s, \mathbf{v}_c) \\ \quad - b_\Omega(\boldsymbol{\theta}_s, \mathbf{v}_c) = \langle l_s, \mathbf{v}_c \rangle, \\ \omega \bar{\Xi}_\Omega(\boldsymbol{\theta}_s, \boldsymbol{\xi}_c) + \Lambda_\Omega(\boldsymbol{\theta}_c, \boldsymbol{\xi}_c) + \omega b_\Omega(\boldsymbol{\xi}_c, \mathbf{s}_s) = \langle r_c, \boldsymbol{\xi}_c \rangle \\ -\omega \bar{\Xi}_\Omega(\boldsymbol{\theta}_c, \boldsymbol{\xi}_s) + \Lambda_\Omega(\boldsymbol{\theta}_s, \boldsymbol{\xi}_s) - \omega b_\Omega(\boldsymbol{\xi}_s, \mathbf{s}_c) = \langle r_s, \boldsymbol{\xi}_s \rangle. \end{array} \right. \quad (19)$$

Having added all the equations of the problem (19) we introduce the linear form $X_\omega: \mathbf{W} \rightarrow \mathfrak{R}$:

$$\begin{aligned} \langle X_\omega, \mathbf{w} \rangle &= \langle l_c, \mathbf{v}_s \rangle - \langle l_s, \mathbf{v}_c \rangle \\ &+ \omega^{-1} (\langle r_c, \boldsymbol{\xi}_c \rangle + \langle r_s, \boldsymbol{\xi}_s \rangle), \\ \forall \mathbf{w} &= (\mathbf{v}_c, \boldsymbol{\xi}_c, \mathbf{v}_s, \boldsymbol{\xi}_s) \in \mathbf{W}, \end{aligned} \quad (20)$$

and the bilinear form $\Pi_\omega: \mathbf{W} \times \mathbf{W} \rightarrow \mathfrak{R}$:

$$\begin{aligned} \Pi_\omega(\boldsymbol{\psi}, \mathbf{w}) &= -\omega^2 [m_\Omega(\mathbf{s}_c, \mathbf{v}_s) - m_\Omega(\mathbf{s}_s, \mathbf{v}_c)] \\ &+ \omega [a_\Omega(\mathbf{s}_c, \mathbf{v}_c) + a_\Omega(\mathbf{s}_s, \mathbf{v}_s)] \\ &+ [c_\Omega(\mathbf{s}_c, \mathbf{v}_s) - c_\Omega(\mathbf{s}_s, \mathbf{v}_c)] \\ &+ [\bar{\Xi}_\Omega(\boldsymbol{\theta}_c, \boldsymbol{\xi}_s) + \bar{\Xi}_\Omega(\boldsymbol{\theta}_s, \boldsymbol{\xi}_c)] \\ &- [b_\Omega(\boldsymbol{\theta}_c, \mathbf{v}_s) - b_\Omega(\boldsymbol{\theta}_s, \mathbf{v}_c)] \\ &+ [b_\Omega(\boldsymbol{\xi}_c, \mathbf{s}_s) - b_\Omega(\boldsymbol{\xi}_s, \mathbf{s}_c)] \\ &+ \omega^{-1} [\Lambda_\Omega(\boldsymbol{\theta}_c, \boldsymbol{\xi}_c) + \Lambda_\Omega(\boldsymbol{\theta}_s, \boldsymbol{\xi}_s)] \\ &\forall \boldsymbol{\psi} = \{\mathbf{s}_c, \boldsymbol{\theta}_c, \mathbf{s}_s, \boldsymbol{\theta}_s\} \in \mathbf{W}, \\ &\forall \mathbf{w} = (\mathbf{v}_c, \boldsymbol{\xi}_c, \mathbf{v}_s, \boldsymbol{\xi}_s) \in \mathbf{W}. \end{aligned} \quad (21)$$

Then variational problem for forced harmonic vibrations of the thermoviscoelastic thin wall body can be rewritten as follows:

$$\left\{ \begin{array}{l} \text{given } \omega > 0, \langle X_\omega, \mathbf{w} \rangle \in \mathbf{W}' = \Phi' \times \Phi'; \\ \text{find } \boldsymbol{\psi} = \{\mathbf{s}_c, \boldsymbol{\theta}_c, \mathbf{s}_s, \boldsymbol{\theta}_s\} \in \mathbf{W} = \Phi \times \Phi \text{ such that} \\ \Pi_\omega(\boldsymbol{\psi}, \mathbf{w}) = \langle X_\omega, \mathbf{w} \rangle \quad \forall \mathbf{w} = (\mathbf{v}_c, \boldsymbol{\xi}_c, \mathbf{v}_s, \boldsymbol{\xi}_s) \in \mathbf{W}. \end{array} \right. \quad (22)$$

V. WELL-POSEDNESS OF THE VARIATIONAL PROBLEM

Let us introduce a scalar product on the space \mathbf{W} in the following way:

$$\begin{aligned} ((\mathbf{y}, \mathbf{w})) &= a_\Omega(\mathbf{s}_c, \mathbf{v}_c) + a_\Omega(\mathbf{s}_s, \mathbf{v}_s) + \\ &\quad \Lambda_\Omega(\boldsymbol{\theta}_c, \boldsymbol{\xi}_c) + \Lambda_\Omega(\boldsymbol{\theta}_s, \boldsymbol{\xi}_s) \\ \forall \mathbf{y} &= \{\mathbf{s}_c, \boldsymbol{\theta}_c, \mathbf{s}_s, \boldsymbol{\theta}_s\} \in \mathbf{W}, \\ \forall \mathbf{w} &= (\mathbf{v}_c, \boldsymbol{\xi}_c, \mathbf{v}_s, \boldsymbol{\xi}_s) \in \mathbf{W}. \end{aligned} \quad (23)$$

And we introduce a norm generated by the scalar product

(23):

$$\| \mathbf{y} \| \|^2 = ((\mathbf{y}, \mathbf{y})) \quad \forall \mathbf{y} \in \mathbf{W}. \quad (24)$$

Then we can easily notice the following estimations:

$$\Pi_\omega(\mathbf{y}, \mathbf{w}) \leq M_c(\omega) \| \mathbf{y} \| \cdot \| \mathbf{w} \| . \quad (25)$$

$$M_c(\omega) = C \max \{ \omega^{-1}, 1, \omega, \omega^2 \} \quad \forall \mathbf{y}, \mathbf{w} \in \mathbf{W},$$

$$\langle X_\omega, \mathbf{w} \rangle \leq M_s(\omega) \| X_\omega \| \cdot \| \mathbf{w} \| . \quad (26)$$

$$M_s(\omega) = C \max \{ \omega^{-1}, 1 \} \quad \forall \mathbf{w} \in \mathbf{W}.$$

Here and everywhere the symbol C – a positive constant value, independent on solutions of variational problem (22).

Now for confirm \mathbf{W} -ellipticity of the bilinear form $\Pi_\omega: \mathbf{W} \times \mathbf{W} \rightarrow \mathfrak{R}$ we consider the expression for $\Pi_\omega(\mathbf{w}, \mathbf{w})$

$$\begin{aligned} \Pi_\omega(\mathbf{w}, \mathbf{w}) &= \omega [a_\Omega(\mathbf{s}_c, \mathbf{s}_c) + a_\Omega(\mathbf{s}_s, \mathbf{s}_s)] \\ &+ \omega^{-1} [\Lambda_\Omega(\boldsymbol{\theta}_c, \boldsymbol{\theta}_c) + \Lambda_\Omega(\boldsymbol{\theta}_s, \boldsymbol{\theta}_s)] \\ &\geq \omega [a_\Omega(\mathbf{s}_c, \mathbf{s}_c) + a_\Omega(\mathbf{s}_s, \mathbf{s}_s)] \\ &+ \omega^{-1} [\Lambda_\Omega(\boldsymbol{\theta}_c, \boldsymbol{\theta}_c) + \Lambda_\Omega(\boldsymbol{\theta}_s, \boldsymbol{\theta}_s)] \\ &\geq \eta(\omega) \| \mathbf{w} \|^2, \quad \eta(\omega) = \min \{ \omega^{-1}, \omega \} \\ &\quad \forall \mathbf{w} = (\mathbf{v}_c, \boldsymbol{\xi}_c, \mathbf{v}_s, \boldsymbol{\xi}_s) \in \mathbf{W}. \end{aligned} \quad (27)$$

Since the statements (25)-(27) are proofed and they are actually the conditions of Lions-Lax-Milgram theorem, the following theorem is then correct:

Theorem 6.1. For each $\omega > 0$ the variational problem (22) has a unique solution $\boldsymbol{\psi} \in \mathbf{W}$, which satisfies the relation:

$$\| \boldsymbol{\psi} \| \leq \eta^{-1}(\omega) M_s(\omega) \| X_\omega \|_* . \quad (28)$$

VI. GALERKIN DISCRETIZATION

Standard Galerkin scheme was used for solving of variational problem (22). We chose some finite-dimensional subspace $\mathbf{W}_h = \Phi_h \times \Phi_h$, $\Phi_h \subset \Phi$, $\dim \mathbf{W}_h = N(h) < +\infty$. Thus, the Galerkin-discretized variational problem (23) looks in the following way:

$$\left\{ \begin{array}{l} \text{given } \omega > 0, X_\omega \in \mathbf{W}', \mathbf{W}_h \subset \mathbf{W}, \dim \mathbf{W} < +\infty,; \\ \text{find } \boldsymbol{\psi}_h = \{\mathbf{s}_{ch}, \boldsymbol{\theta}_{ch}, \mathbf{s}_{sh}, \boldsymbol{\theta}_{sh}\} \in \mathbf{W}_h \text{ such that} \\ \Pi_\omega(\boldsymbol{\psi}_h, \boldsymbol{\varphi}) = \langle X_\omega, \boldsymbol{\varphi} \rangle \quad \forall \boldsymbol{\varphi} \in \mathbf{W}_h. \end{array} \right. \quad (29)$$

We can say the problem (23) is well-posed same as the problem (29). In the space \mathbf{W} we select some basic functions $\{\mathbf{w}\}_{i=1}^\infty$. For each natural number $m \geq 0$, $h = 1/m$ a sequence of approximation spaces \mathbf{W}_h and operators of

orthogonal projection $Pr_h : \mathbf{W} \rightarrow \mathbf{W}_h$ are defined so that a set $\{\mathbf{w}\}_{i=1}^m$ is a basis of \mathbf{W}_h , and $((\boldsymbol{\Psi} - Pr_h \boldsymbol{\Psi}, \mathbf{w})) = 0 \quad \forall \boldsymbol{\Psi} \in \mathbf{W}, \forall \mathbf{w}_h \in \mathbf{W}_h$. Now variational problem (22) is replaced by a sequence of the following problems:

$$\left\{ \begin{array}{l} \text{given } \omega > 0, X_\omega \in \mathbf{W}', h > 0, \\ \mathbf{W}_h \subset \mathbf{W}, \dim \mathbf{W} = m < +\infty; \\ \text{find } \boldsymbol{\Psi}_h = \{\mathbf{s}_{ch}, \boldsymbol{\theta}_{ch}, \mathbf{s}_{sh}, \boldsymbol{\theta}_{sh}\} \in \mathbf{W}_h \text{ such that} \\ \Pi_\omega(\boldsymbol{\Psi}_h, \boldsymbol{\Phi}) = \langle X_\omega, \boldsymbol{\Phi} \rangle \quad \forall \boldsymbol{\Phi} \in \mathbf{W}_h. \end{array} \right. \quad (30)$$

Theorem 5.1. Let be $\forall \boldsymbol{\Psi} \in \mathbf{W}$ a solution of problem (22) with parameter $\omega > 0$. Then a sequence of Galerkin approximations $\forall \boldsymbol{\Psi}_h \in \mathbf{W}$ is unambiguously defined by the solutions of the problems (30) and has the following properties:

$$\| \boldsymbol{\Psi} - \boldsymbol{\Psi}_h \| \leq \eta^{-1} M_c(\omega) \inf_{\forall \mathbf{w} \in \mathbf{W}_h} \| \boldsymbol{\Psi} - \mathbf{w} \| \quad \forall h > 0; \quad (31)$$

$$\lim_{h \rightarrow 0} \| \boldsymbol{\Psi} - \boldsymbol{\Psi}_h \| = 0. \quad (32)$$

Proof. The correctness can be done like in [7]. Since for the inequality (31)

$$\Pi_\omega(\boldsymbol{\Psi} - \boldsymbol{\Psi}_h, \mathbf{w}) = 0 \quad \forall \mathbf{w} \in \mathbf{W}_h$$

and the estimation

$$\begin{aligned} a \| \boldsymbol{\Psi} - \boldsymbol{\Psi}_h \|^2 &\leq \Pi_\omega(\boldsymbol{\Psi} - \boldsymbol{\Psi}_h, \boldsymbol{\Psi} - \boldsymbol{\Psi}_h) = \Pi_\omega(\boldsymbol{\Psi} - \boldsymbol{\Psi}_h, \boldsymbol{\Psi} - \mathbf{w}) \\ &\leq M_c(\omega) \| \boldsymbol{\Psi} - \boldsymbol{\Psi}_h \| \| \boldsymbol{\Psi} - \mathbf{w} \| \quad \forall \mathbf{w} \in \mathbf{W}_h. \end{aligned}$$

Taking into account the density of sequence of spaces $\{\mathbf{W}_h\}$ in the separable space \mathbf{W} :

$$\lim_{h \rightarrow 0} \| \boldsymbol{\Psi} - Pr_h \boldsymbol{\Psi} \| = 0 \quad \forall \boldsymbol{\Psi} \in \mathbf{W}. \quad (34)$$

Therefore, basing on the equality

$$\inf_{\forall \mathbf{w} \in \mathbf{W}_h} \| \boldsymbol{\Psi} - \mathbf{w} \| = \| \boldsymbol{\Psi} - Pr_h \boldsymbol{\Psi} \| \quad (35)$$

and inequality (31) we can conclude the correctness of (32), when $\omega > 0$.

VII. NUMERICAL EXPERIMENTS

Below we present some results of our numerical experiments on computations of eigenvalue problem for our semidiscreted model. We consider a circular cylindrical shell made of homogenous material with radius $R=10$ m and length $L=10$ m and which is under constant temperature.

Young's modulus of shell material is equal to 1 Pa, Poisson's coefficient is 0.3, and mass density is 1 kg/m³.

Boundary conditions are following type:

$$\begin{aligned} u_2 = \gamma_2 = u_3 = \gamma_3 = 0, \quad \text{on } \alpha_1 = 0, \alpha_1 = L; \\ u_1 = \gamma_1 = u_3 = \gamma_3 = 0 \quad \text{on } \alpha_2 = 0, \alpha_2 = \pi/8. \end{aligned}$$

The first column of the Table includes the number of quadratic finite element mesh, the second and third columns include the computed eigenvalues $\omega^2 \cdot 10^3$ and their relative errors δ taking from [1]. Same our results are in the two last columns of the Table.

Mesh	$\omega^2 \cdot 10^3$ [1]	δ , %	$\omega^2 \cdot 10^3$	δ , %
3×3	0,3305068	11,7	0,3583986	11,9
4×4	0,3024595	2,23	0,3401321	6,2
5×5	0,2974929	0,55	0,3288990	2,7

VIII. CONCLUSION

The partially variational problem for a thin wall body was constructed on base the dynamic coupled three-dimensional problem. Under the assumptions about harmonic vibration with known angular frequency we have formulated the corresponding variational problem and then we proved its well posedness. These results shows that we can use well known finite element approximations for Sobolev spaces and obtain the convergence rate its errors.

REFERENCES

- [1] R. Malets, and H. Shynkarenko, "Modeling and solvability of the variational problem of thermo-elastic thin shells, compliant to shear and compression," Manufacturing Processes. Actual Problems, Basic Science Applications. Opole: Politechnika Opolska, vol.1, pp. 103-121, 2014.
- [2] R. B. Malets, "Modeling of heat conduction processes in a thin three-dimensional layer," Visnyk of the Lviv University. Series Appl. Math. and Informatics, iss. 1, pp. 240-250, 2009.
- [3] P. M. Naghdi, "The Theory of Shells and Plates," Handbuch der Physik Berlin-Heidelberg-New York: Springer, vol. VIa2, pp. 425-640, 1972.
- [4] Ya. S. Podstrigach, and R. N. Shvets, Thermoelasticity of thin shells. K.: Naukova dumka, 1978.
- [5] P. P. Vahin, R. B. Malets, and H. A. Shynkarenko, "Variational formulation of the problem of nonstationary thermo-elasticity for thin shells compliant to shears and compression," J. Math. Sci., no. 3, pp. 345-364, 2016.
- [6] R. Malets, and H. Shynkarenko, "Construction and analyse one-step integration time scheme for problem of thermoelastic shells compliant to shear and compression," Applied radioelectronics, vol. 14, no 2, pp. 176-184, 2015.
- [7] J. Necas, and I. Hlavacek, Mathematical Theory of Elasticity and Elastic-Plastic Bodies: An Introduction. Amsterdam:Elsevier, 1981.
- [8] V. Stelmashchuk, and H. Shynkarenko, "Finite Element Analysis of Green-Lindsay Thermo-piezoelectricity Time Harmonic Problem," Visnyk of the Lviv University. Series Appl. Math. and Informatics, iss. 25, pp. 136-147, 2017.
- [9] Ya. G. Savula, and N. P. Fleishman, Calculation and optimization of shells with curved median surfaces. Lviv: Vishcha School, 1989..

Educational Schedule Development Using Evolution Technologies

Oleh Suprun
*Intellectual and Information Systems
Department
Taras Shevchenko National University
of Kyiv
Kyiv, Ukraine
oleh.o.suprun@gmail.com*

Olena Sipko
*Information Technologies Engineering
Department
Cherkasy State Technological
University
Cherkasy, Ukraine
barchat_08@mail.ru*

Vitaliy Snytyuk
*Intellectual and Information Systems
Department
Taras Shevchenko National University
of Kyiv
Kyiv, Ukraine
snytyuk@gmail.com*

Abstract — An effective and optimal schedule is one of the keys to obtaining the necessary knowledge and skills by students, as well as creating comfortable conditions for teachers to work. However, in most universities, schedules are still being developed manually, which takes a long time, and always has a certain percentage of subjectivity. This is due to the relevant software inability to take into account the preferences and requirements of students and teachers and to give them the priorities as a human expert can do. The proposed method makes it possible, by conducting surveys among students and faculty members, to determine and take into account their requirements, to evaluate the possibility of fulfilling their desires and to prioritize them, depending on various factors such as material provision, the number of students with certain proposals for the schedule, teacher's position and status, and others. Using evolutionary technologies allows to quickly conduct an analysis that makes possible to perform multiply experiments, changing certain parameters, and choose the best option. The automation of this process guarantees taking into account all the restrictions and desires that a human expert can not handle with when dealing with a large number of students and teachers. This removes the influence of subjectivity. The proposed method was tested on real data, its efficiency and advantages are shown in the paper.

Keywords—education process, schedule development, genetic algorithm, penalty functions.

I. INTRODUCTION

The schedule development problem is being solved by every person almost daily, regardless of whether it happens deliberately or not. Correct and optimal planning of future actions is one of the key factors affecting the final result of any process, it determines the efficiency and profitability of the conducted economic and financial transactions.

Education process at universities is no exception. Creating a valid curriculum is an essential prerequisite for necessary skills acquisition by future specialists. This is especially important recently, due to the information society rapid development. As new in-depth disciplines and areas to study are emerging constantly, students have to learn more and more information. Also, such aspects as providing sufficient amount of time for learning, preparing for exams and rest, must be taken into account. The information understanding and overall training level depends on the timely submission of information, laboratory workshops organization, practical trainings and seminars. Correct schedule preparation is also equally important for teachers since it regulates the labour rhythm and directly affects their productivity and efficiency. Not to mention the universities

technical provision, that is often far from ideal. The lack of audiences for the simultaneous placement of all students, computer equipment, or the necessary training material appears rather often.

Thus, the curriculum and schedule directly influences the level of specialist's final training, and some little at first glance errors or inaccuracies may have significant outcomes in the future.

Despite this, mathematicians began the necessary methodology development relatively recently. In 1967, in United States the world's first book on the theory of schedules was published [1]. One of the earliest conferences devoted exclusively to the problem of scheduling was the International Conference on Theory and Practice of Automated Timetables, held in Edinburgh in 1995 [2].

Nowadays, many literature on the curriculum development problem in higher education institutions can be found, but usually they are of a purely theoretical nature and are interesting only as a research of a non-trivial mathematical problem, or on the contrary - it has a narrow specialization, and the proposed method can be used only for a particular situation, such as the development of a curriculum for distance education [3], for software developers training [4,5], a schedule optimization for providing students with the necessary practical skills [6], or even to encourage the study of certain disciplines [7], and others.

Also many programs that allow to create schedule according to the specified rigid, that is technical, conditions, were developed. However, such programs do not allow to take into account the priorities and wishes of teachers and students, that is a negative psychological aspect. Most of the proposed programs often do not meet the requirements and are quite uncomfortable, therefore rarely used.

It is the subjectivism that is presented in the schedule creation process, which leads to numerous conflicts, significant time costs and appearance of suboptimal by different criteria decisions. The automation of the scheduling problem process is rather complicated problem, its algorithmization encounters aspects of NP (nondeterministically polynomial) complication. The search for precise algorithms for solving these problems, the time of which is limited by the polynomial of the input data size, can't give any proper results. Exhibitory selection algorithms require significant computational cost even when solving average dimensional examples. Therefore, one of the important areas of research is the construction and analysis

of approximate algorithms with guaranteed accuracy estimation for NP-complete problems.

II. PROBLEM ANALYSIS

Solving the problem of schedule development in general is the process of executing some fixed tasks system using a certain set of resources or service devices. When transferring the general schedules theory to the educational and training schedule, the formulation of this problem class is as follows: "For a given set of training classrooms and a given set of time intervals (lessons), build such a distribution of training sessions for all objects (teachers and training groups) for which the chosen criterion of optimality is the best".

The educational schedule must fulfil the following basic requirements:

- precise and full schedule compliance with the curriculum by volume, content, type and time of classes, provision of materials for the curriculum and programs, systematic and continuous learning process throughout the day, and the evenly distributed students' work during the week, month and semester;
- providing on the classes of interdisciplinary and internal logical connections for each discipline, which are determined by its structural-logical scheme;
- providing of necessary time intervals for students to work independently between lectures and practical classes for each discipline, alternating between disciplines with different complexity levels and classes types during the day and the week;
- implementation of principles for teacher's and student's scientific work organization, achievement of an equal teachers employment in order to ensure their preparation before classes, systematic conducting of methodological, publishing and research activity (here the individual needs department head recommendations may be taken into account);
- providing the effective use of the auditor's and the training laboratory base.

Also there are other requirements of ergonomic and organizational-methodical nature. For example, they include the reasonable execution of the individual teachers preferences, the implementation of the individual training principle, conducting classes with small students groups, and others.

Developing a schedule, the problem of optimal resource management, such as the teaching staff and the auditorium, appears. Solving this problem, it is necessary to take into account the strict restrictions, as well as additional requirements that may be violated in some cases.

Strict restrictions are limitations that must surely be fulfilled; those that physically can not be violated. As example of these: "At the same time, the same audience should have only one teacher and one subject." As a result of solving this problem, it is necessary to get a schedule that simultaneously satisfies all the strict restrictions. If this is not possible, then the list of such restrictions should be changed or some measures should be taken to allow for an acceptable schedule.

Weak restrictions are limitations that can be violated, but this violation must be kept to a minimum level. Their performance is not as obligatory as the strict ones. In contrast to the strict restrictions that have an objective nature, soft restrictions are subjective. Thus, the restriction "The lesson must not be conducted in the laboratory" is objective-subjective, and the restriction "Lecturer Goroshko's lectures should be conducted on Monday and Tuesday" - subjective. It is obvious that violation of weak restrictions leads to the schedule deterioration, but does not exclude its admissibility. Since such violations can be many and they are of a versatile nature, the relevance of obtaining an acceptable optimal (acceptable) timetable is indisputable.

III. PROBLEM FORMALIZATION AND OBJECTIVE FUNCTION

Let the set $R = \{r_1, r_2, \dots, r_n\}$ be a finite set of all possible schedules for a certain educational institution. Its finiteness is guaranteed by the finiteness of educational disciplines set $P = \{p_1, p_2, \dots, p_m\}$, teachers set finiteness $L = \{l_1, l_2, \dots, l_k\}$, students set finiteness $S = \{s_1, s_2, \dots, s_k\}$ and auditoria's set $A = \{a_1, a_2, \dots, a_v\}$.

The schedule development problem, without generality limitation, can be presented as follows [9]:

$$\max_{r \in R} F(r), r \in \Omega(P, S, L, A) \in R,$$

where Ω – a set of restrictions that are determined by the auditoria's and classrooms presence and specialization, the teachers distribution according to the disciplines, discipline according to the classrooms, etc.

Taking into account the requirements and priorities of teachers and students separately, the following formula can be obtained:

$$\alpha_s F_s(r) + \alpha_L F_L(r) \rightarrow \max, r \in \Omega(P, S, L, A),$$

where F_s – students objective function, F_L – teachers objective function, α_s i α_L are weighting coefficients indicating the priorities of teachers and students as the educational process subjects.

Considering the student as the dominant subject in a higher educational institution, it is rational to establish a priori $\alpha_s = 0,6$, $\alpha_L = 0,4$.

For further correction, the following rule is used: if the ratio of the students' number to the teachers' number corresponds to the normative value, the values of the coefficients do not change, if the real ratio is different from the normative, then α_s and α_L must be corrected.

Let N_s be the students number, N_L – teachers number, Nom – the nominal value of students number ratio to the teachers number determined by the managing authority. If

$$\text{inequality } \frac{1}{2} Nom \leq \frac{N_s}{N_L} \leq Nom, \text{ is true, then}$$

$$\alpha_s = 0,6 - \frac{1}{2} \text{Nom} \left(\frac{N_s}{N_L} - \frac{1}{2} \text{Nom} \right) \cdot 0,4.$$

The features of teachers' and students' preferences and requirements set forming are presented below.

Obviously, students can be considered as a certain set, divided into classes (groups according to courses and specialties). Students in each group independently formulate their schedule requirements, which form a single list. If it meets similar requirements - they are united. The authors of the opposing requirements are offered to reach the agreement or to withdraw their claims altogether. In the event of their disagreement among the students, the vote is conducted and the requirement is chosen by the majority.

Thus, a set of requirements is obtained Z^v , it has different priority for each student. In order to reconcile individual preferences, each student is given the opportunity to determine the advantage of each requirement. For this purpose, the hierarchies analysis method is used [8], the matrices of pairwise comparisons are constructed, for which the maximum eigenvalues and corresponding vectors are found. Let the λ_{\max} be the eigenvalues, for the i -th student, $i = \overline{1, m}$, $x^i = (x_1^i, x_2^i, \dots, x_l^i)$ - corresponding eigenvector. Performing the normalization of this vector elements by the formula

$$x_j^{in} = \frac{x_j^i}{\sum_{j=1}^l x_j^i}$$

is obtained, that $x_j^{in} \in (0,1)$ and $\sum_{j=1}^l x_j^{in} = 1$. It can be said that

the value x_j^{in} shows a priority of j -th criteria for i -th student. Since all students are equivalent (equally competent), the requirements priority for them (most often, for the group) is defined as the average value of the criteria priorities for each student, that is:

$$x_j = \frac{1}{m} \sum_{i=1}^m x_j^{in}, j = \overline{1, l}.$$

Thus, for the students group, a requirements priority vector is obtained: $X = (x_1, x_2, \dots, x_l)$. Objective formula can be modified:

$$\alpha_s \cdot F_s = \alpha_s \cdot \sum_{j=1}^l x_j \cdot \chi\{Z_j^v\},$$

$$\chi\{A\} = \begin{cases} 1, & \text{if } A \text{ is true,} \\ 0, & \text{otherwise.} \end{cases}$$

Teachers do not have groups and their requirements need to be individually implemented. Let M be the number of

teachers, divided into sets $T = \{T_1, T_2, \dots, T_K\}$ (department head, doctors of sciences, assistant professors, assistants, etc.).

The specialist, who makes the schedule, forms a matrix of pairwise comparisons: defines the priorities of teachers, representatives of groups:

$$y = \{y_1, y_2, \dots, y_K\}, \quad y_i \in (0,1), \quad \sum_{i=1}^K y_i = 1,$$

Each teacher has his own advantages in forming a schedule, and the number of such advantages from different teachers will be different. Let $Z_i^T = \{Z_{i_1}^{T_j}, Z_{i_2}^{T_j}, \dots, Z_{i_{n_i}}^{T_j}\}$ be the advantages vector for i -th teacher from j -th set, n_i - number if its elements, $j = \overline{1, M}$, $i = \overline{1, K}$. Vectors Z_i^T will correspond the values of the priority vector calculated using the method given above $D_i^j = \{d_{i_1}^j, d_{i_2}^j, \dots, d_{i_{n_i}}^j\}$, $i = \overline{1, M}$, $j = \overline{1, K}$. Thus:

$$\alpha_L F_L = \sum_{j=1}^K y_j \cdot \sum_{i=1}^M \chi\{L_i \in T_j\} \cdot \sum_{i=1}^{n_i} d_{i_1}^j \cdot \chi\{Z_{i_1}^{T_j}\},$$

$$\chi\{L_i \in T_j\} = \begin{cases} 1, & \text{if } L_i \text{ belongs to } T_j, \\ 0, & \text{otherwise.} \end{cases}$$

The objective function can be rewritten as follows:

$$F(r) = \alpha_s \sum_{j=1}^l x_j \chi\{Z_j^v\} +$$

$$+ \alpha_L \sum_{j=1}^K y_j \sum_{i=1}^M \chi\{L_i \in T_j\} \sum_{i=1}^{n_i} d_{i_1}^j \cdot \chi\{Z_{i_1}^{T_j}\} \rightarrow \max,$$

$$r \in \Omega(P, S, L, A).$$

where x_j and y_j are students and teachers preferences' priorities, Z_j^v students preferences, L_i teachers, T_j teachers' groups, $Z_{i_1}^{T_j}$ teachers' advantages, $d_{i_1}^j$ - priority of these advantages, l number of students' preferences, K number of teachers' groups, determined by their positions, scientific degrees and academic rank, M number of teachers' groups, n_i number of teachers' in i -th group.

IV. MATRIX-EVOLUTIONARY METHOD

In order for better visualization, the schedule representation can be shown as a rectangular parallelepiped (Fig. 1).

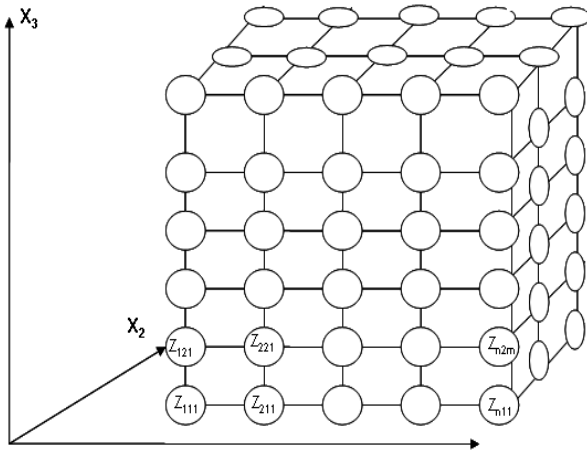


Fig. 1. Schedule representation.

where X_1 is day-lesson, X_2 - course-group, X_3 - auditory, Z - teacher-subject

Since the problem of automated schedule development is an NP-complete task, it is expedient to use the evolutionary technologies algorithm, namely, a modified genetic algorithm. In this case, the method of penalty functions is used, which makes it possible to transform a problem with constraints into a sequence of unconditional optimization problems of some additional functions [10]. They are obtained by modifying the target function with the help of restriction functions in such a way that limitations are not presented in the optimization problem in explicit form.

The modified method of solving the schedule development problem, using the penalty function, has the following steps:

- Step 1. Define structure S of potential schedule r .
- Step 2. Define the criteria E for the search stop.
- Step 3. Perform a potential solution encoding.
- Step 4. Until criteria E is done:
 - Step 4.1. Until a sample of potential solutions Z incomplete:
 - Step 4.1.1. Generate potential solution r .
 - Step 4.1.2. If it is unacceptable ($r \notin \Omega_1(P, S, L, A)$) go to Step 4.1.1.
 - Step 4.1.3. If it is acceptable ($r \in \Omega_2(P, S, L, A)$), add it to Z and go to Step 4.1.
 - Step 4.1.4. If the solution r is unacceptable according to at least one of $\Omega_2(P, S, L, A)$, one of three variants is done:
 - A: If to Step 4.1.1.
 - B: If variant A was made more than A_{\max} times, go to variant C.
 - C: Set

$$F(r) = \alpha_S F_1(r) + \alpha_L F_2(r) - \beta_{it} \phi(F_1(r) \vee F_2(r)),$$

where β - weight coefficient, it - iteration number, $\phi(*)$ - penalty function. Consider r as potential solution and go to Step 4.1.1.

Step 4.2. For all potential solutions calculate $F(*)$, taking into account, that if the solution is acceptable, $\phi(*)$.

Step 4.3. Generate new potential solutions based on the values of the target function, using crossover operations and mutations (if the optimization method is a genetic algorithm) or using normally distributed numbers if this is an evolutionary strategy.

Step 5. Calculate the criteria E .

Penalty function construction is presented below.

$$\phi(F_1(r) \vee F_2(r)) = \begin{cases} 1, & \text{if } r \in \Omega_2(P, R, L, A) \\ f(\bigvee_{j=1}^I x_j, \bigvee_{j=1}^K y_j, D, \gamma), & \text{otherwise} \end{cases}$$

The zero value of a penalty function presents the situation where at least one student's requirement or at least one teacher's requirement is not fulfilled, D - value that integrates the teachers' requirements priorities, γ - penalty parameter.

Since "fined" may be solutions that do not fulfil the requirements of student groups, and solutions that do not fulfil the individual teachers' requirements, it is expedient to consider the penalty function additive and to write in this form:

$$\begin{aligned} \phi(F_1(r) \vee F_2(r)) &= \beta_1 \phi_1(F_1(r)) + \beta_2 \phi_2(F_2(r)) = \\ &= \beta_1 f_1(\bigvee_{j=1}^I x_j) + \beta_2 f_2(\bigvee_{j=1}^K y_j, D). \end{aligned}$$

Obviously, the more restrictions are violated, the greater the penalty function value is. According to the construction, penalty function is an integral function, and based on its purpose, the inequality is valid:

$$0 \leq \phi(*) \leq F_{\max}.$$

Similarly, for its component: $0 \leq \phi_i(F_i(r)) \leq F_{i\max}$, $0 \leq f_1(\bigvee_{j=1}^I x_j) \leq F_{1\max}$ and $0 \leq f_2(\bigvee_{j=1}^K y_j) \leq F_{2\max}$.

V. EXPERIMENTAL RESULTS

To check the proposed algorithm, an automated system was created that develops schedule in accordance with the requirements of the educational process, the disciplines sequence correctness, and the desired of teachers and students. The system has been tested both on specially created and real data.

The system kernel and the interface part were written in the programming language Delphi 7.0. The presented solution is performed using object-oriented technologies, which will allow to easily encapsulate them in future in new system modifications, without violating the algorithms

integrity. The database was implemented on the InterBase 6.0.

As real testing data were used information about students groups, teachers and subjects of full-time studying at the Faculty of Information Technologies of the Cherkasy State Technological University, as well as randomly generated initial data (disciplines were randomly assigned to the classes).

To determine the developed models and target functions effectiveness and expediency, as well as the created automated system relevance, calculations were made on the schedules created automatically and manually. The research was conducted on schedules for the Faculty of Information Technologies and Systems of Cherkasy State Technological University, created on various samples of the initial data: for different semesters (autumn, spring), for different academic years (2014-2015, 2015-2016) The target function results for different initial data are shown in Table I.

TABLE I. OBJECTIVE FUNCTION VALUES

Creation method	Autumn semester 2014-2015	Spring semester 2014-2015	Autumn semester 2015-2016	Spring semester 2015-2016
Manually	643	488	604	543
Automated	682	521	663	571
Difference in percents	≈5,7%	≈6,3%	≈8,9%	≈4,9%

The calculation of the target functions values obtained with automatic and manually creating a schedule methods showed that the effectiveness of the developed models and methods is about 5-9% comparing to the manually schedule creation for classes in higher educational institutions.

In addition to testing the automatic system optimality, a test was made on the dependence of the time, used for creating the optimal schedule, to the problem dimensionality. On average, from 5 to 10 tests for each dimensionality of the initial data was made. The result data are presented in the Table II.

TABLE II. CALCULATION TIME DEPENDING ON PROBLEM DIMENSIONALITY

Problem dimensionality (lessons number)*(groups number)*(days of week number)	Calculation time		
	max	min	average
5	0,8	0,05	0,24
25	3,2	0,9	1,96
45	5,4	2,1	3,5
65	12	3,1	5,8
85	14	5,2	7,6
105	25	10	14,05
125	39	14,5	18,1
145	46	19	26,5
165	51	25	32

In Figure 2 the dependence of the time used for solving the problem on the problem dimensionality is shown (the number of pairs per week and the number of groups).

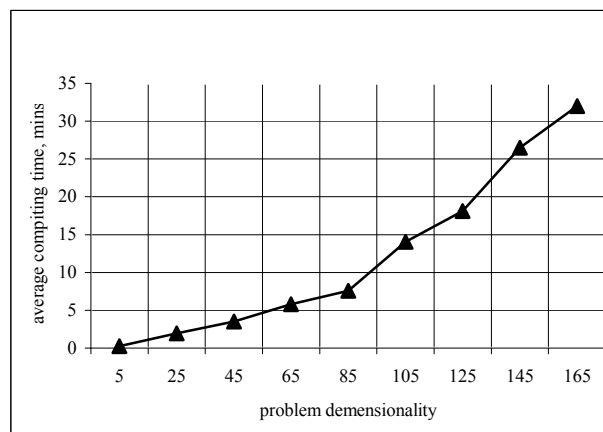


Fig 2. Calculation time.

As can be seen, the problem solving time increases with the increase of input data amount. This is due to a fast increase in the number of restrictions in the model, which increases the size of the arrays and, accordingly, the time used to solve the problem.

VI. CONCLUSIONS

Despite the existence of many programs for automated schedule development in higher education institutions, in most cases the schedule is formed manually, which requires considerable effort, it is time consuming, and such a schedule is far from always optimal and effective. This is due to the inability of such software to take into account the large number of preferences of students and teachers, to choose acceptable and unfulfilled preferences, and to take into account their priority.

The proposed schedule development problem formalization, and the algorithm for its direct solving, allows using the hierarchies' analysis method, conducting surveys among students and teachers, formalizing their preferences, and appropriately taking them into account, when developing a schedule. On the one hand, it can not completely replace the expert's analytical ability, based on his own experience and knowledge of the task, but at the same time, it will avoid subjectivity. In addition, an automated schedule creating system can take into account absolutely all preferences, even if some will be inappropriate, and optimise the schedule, while the specialist will not be able to effectively evaluate all the requirements and wishes, which is especially important for institutions with a large number of students.

The search for the optimal schedule is performed using a modified genetic algorithm, which has demonstrated itself well solving the optimization problems, which can not be solved by classical methods. It allows to find the optimal, or at least acceptable, solution using a small period of time.

Taking into account the preferences of students and teachers is made using the penalty functions, which allows to take into account the priority of these preferences, to evaluate their importance in general, and in comparison with each other. Due to this, the preferences and requirements are perceived more effectively than when manually developing the schedule, the optimal schedule contains the least amount

of violated desires. Also, using the penalty method allows to simplify the target function, which generally accelerates the algorithm.

Experimental researches were done using the developed system for automated schedule creating, while both theoretical and real data were used. A comparison of the obtained schedule and the time used for its creation, with the schedules obtained by classical methods, in particular, generated manually by an expert, is performed, and the results are presented.

REFERENCES

- [1] E. Balas "A Note on the Branch-and-Bound Principle," *Operations Research*. New York, vol. 16, iss. 2, pp. 442-445, April 1968.
- [2] E. Burk, and P. Ross, *Practice and Theory of Automated Timetabling: First International Conference*. Edinburgh: UK, 1995.
- [3] Y. Yang, and D. Tian, "The research of multimedia curriculum design and curriculum development in modern distance education", *2016 International Conference on Educational Innovation through Technology (EITT)*, September 2016. Published in: 2017 3rd IEEE International Conference on Computer and Communications (ICCC), 13-16 Dec. 2017. DOI: 10.1109/CompComm.2017.8322799
- [4] E. S. Grant, "A proposal for technology-based software engineering curriculum development," *2016 IEEE 8th International Conference on Engineering Education (ICEED)*, Kuala Lumpur, Malaysia, December 2016.
- [5] E. S. Grant and V. Shankaraman, "Technology-driven software engineering curriculum development," *2014 IEEE 27th Conference on Software Engineering Education and Training (CSEE&T)*, Klagenfurt, Austria, pp. 168-170, April 2014.
- [6] H. Fugang, C. Le, and Z. Qinhu. "Curriculum Development for Practical Training: A Design-Based Research of Network Detection Combat," *2016 International Conference on Educational Innovation through Technology (EITT)*, Tainan, Taiwan, September 2016.
- [7] W. Yoo, S. Pattaparla, and S. Shaik, "Curriculum development for computing education academy to enhance high school students' interest in computing", *Integrated STEM Education Conference (ISEC)*, 2016 IEEE, Princeton, NJ, USA, pp. 282 - 284 April 2016.
- [8] T. L. Saaty, and Kirti Peniwati, *Group Decision Making: Drawing out and Reconciling Differences*. Pittsburgh, Pennsylvania, 2008.
- [9] O. Sipko, and V. Y. Snytyuk, "Aspects of formulation of the objective function in the problem of scheduling in higher educational institutions based on subjective preferences," *Nauka i Studia*. Polska, Przemysł: Sp. zoo «Nauka i studia», no. 15 (125), pp. 39-51, 2014.
- [10] V. Y. Snytyuk, and O. M. Sipko "Penalty Functions Using in the Schedule Development Problem," *Mathematical Machines and Systems*, Kyiv, no. 3, pp. 158-164, 2015.

Automated Labeling of Bugs and Tickets Using Attention-Based Mechanisms in Recurrent Neural Networks

Volodymyr Lyubinetz
Forethought Technologies
San Francisco, USA
vlyubin@gmail.com

Taras Boiko
Lviv National University
Lviv, Ukraine
me@tboiko.com

Deon Nicholas
Forethought Technologies
San Francisco, USA
deon@forethought.ai

Abstract—We explore solutions for automated labeling of content in bug trackers and customer support systems. In order to do that, we classify content in terms of several criteria, such as priority or product area.

In the first part of the paper, we provide an overview of existing methods used for text classification. These methods fall into two categories - the ones that rely on neural networks and the ones that don't. We evaluate results of several solutions of both kinds.

In the second part of the paper we present our own recurrent neural network solution based on hierarchical attention paradigm. It consists of several Hierarchical Attention network blocks with varying Gated Recurrent Unit cell sizes and a complementary shallow network that goes alongside.

Lastly, we evaluate above-mentioned methods when predicting fields from two datasets - Arch Linux bug tracker and Chromium bug tracker.

Our contributions include a comprehensive benchmark between a variety of methods on relevant datasets; a novel solution that outperforms previous generation methods; and two new datasets that are made public for further research.

Index Terms—text classification, recurrent neural network, hierarchical attention, machine learning, natural language processing

I. INTRODUCTION

When dealing with a customer support ticket, one of the first things a customer service agent has to do is to label the ticket in terms of multiple criteria. These could be priority, product area, or whether action is required from an engineering team. Such labels are used for effective handling of the ticket - for example, tickets with high priority will be dealt with before low priority tickets, or engineering team will intervene only if the ticket was marked for intervention. Figure 1 shows an example of a labeling panel in one of these customer service platforms.

A similar scenario is also true for project tracking systems such as JIRA or Bugzilla. Employees often label tasks in terms of area (e.g. kernel vs front-end) so that the appropriate team takes a look at it, or in terms of type - bugs typically have higher priority than new feature requests.

Considering that before such labeling takes place, hours or even days can pass, an ability to perform it automatically would increase the speed at which businesses operate and

The image shows a web interface for labeling a customer service ticket. At the top, there is a 'Tags' section with a text input field containing 'needs_manual_action' and a close button. Below this are several form elements: a 'Type' dropdown menu, a 'Priority' dropdown menu (which is open, showing options: '-', 'Low', 'Normal', 'High', 'Urgent'), a 'Product Area*' text input field, an 'Action to perform*' text input field, a checkbox labeled 'Send Stella Connect', and a 'Ticket Level' dropdown menu.

Fig. 1. Setting labels for a customer service ticket

dramatically reduce the costs. Therefore, this is an extremely important problem, which got more traction recently due to evolution of deep neural networks and results achieved by leveraging word embeddings.

The general problem we are solving is that of text classification. Given a body of text we have to derive its class from a known fixed set of classes. However, using text classification on data from customer service systems and task trackers has its caveats. For example, such data often includes unique fragments that are hard for automatic systems to reason about, such as stack traces or HTML snippets. On the other hand, they often are well-structured and this structure can be leveraged by some of the novel methods, including the one we are proposing.

We are primarily interested in multi-class text classification, where number of classes we're predicting is larger than two. Historically, great results have been reached on binary classi-

fication tasks such as sentiment analysis (e.g. Twitter dataset, where you have to tell whether a tweet is positive or not), or spam filtering, where you have to tell whether email is a spam or no. Depending on the task, 95%+ accuracies can often be achieved. This is due to fact that binary categories often have a lot of clue words present, which simplify classification task (e.g. "good" or "great" in positive reviews, or "spam" in spam email). However, if you look at data from Table 1, which shows state of the art results on multi-class classification dataset, it becomes clear that this is not a solved problem. Accuracies on Amazon and Yelp reviews datasets, where you have to predict the ranking of a review on 1 to 5 scale, hover around 60% to 70%, with nobody beating the 50% threshold on IMDB dataset for movie genre prediction with 15 categories.

TABLE I: State of the art accuracies on multi-class classification, results taken from [1]

Paper	Yelp'15	IMDB	Amazon
Zhang et al., 2015	59.9%	-	55.3%
Tang et al., 2015	67.6%	45.3%	-
Yang et al., 2016	71.0%	49.4%	63.6%

We begin with an overview of existing methods, then we will present our own solution, and lastly we will provide a comprehensive benchmark of all methods in question.

II. EXISTING METHODS

Text classification is one of the most important problems of Natural Language Programming (NLP) research and variety of methods have been proposed for it.

These methods can be split into two large families - classic solutions that don't leverage neural networks, and novel solutions that leverage recurrent neural networks, especially with the use of word embeddings. Among the former methods are Naive Bayes and algorithms that use term count data, such as Term Frequency Inverse Document Frequency (TF-IDF) fed into the Support Vector Machine (SVM) classifier. Among the latter are methods that use recurrent neural networks on word embeddings data, which differ in network structure, loss functions, algorithms used to derive embeddings and preprocessing routines.

In this section we provide a brief overview of these methods, as they will be part of the benchmark in the section 4.

A. Naive Bayes

Naive Bayes used to be one of the most popular algorithms for text classification, coming into NLP scene in 1960's. It was widely used for early spam filters, where it still performs fairly well [2]. But as we will see later, Naive Bayes shows poor results on multi-class classification. The key idea behind Naive Bayes classifier is using the Bayes theorem - for a document d and class c , we can say that probability of that document having class c is:

$$P(c|d) = \frac{P(d|c)P(c)}{P(d)}$$

After making the "naive" assumption about independence of conditional probabilities for individual terms, we get

$$P(d|c) = P(x_1, x_2, \dots, x_k|c) = P(x_1|c)P(x_2|c)\dots P(x_k|c)$$

, where x_i are terms contained in document d . Then we simply choose the class that maximizes $P(x_1, x_2, \dots, x_k|c)P(c)$, with each of these probabilities computed on the training set.

B. TF-IDF with SVM

Since 1990's algorithms using term count statistics such as TF-IDF took prominence in NLP community. The idea behind TF-IDF is to represent each sentence as a vector of scores determined by term frequencies. The score consists of two parts - one determined by counts of that term inside a document and the other by presence of the term across the body of documents, with the score being the multiplication of the two .

Once we have the TF-IDF data, we can use it with any supervised classification method, such as Softmax classifier or Support Vector Machine. The latter is a popular choice and used to be state of the art method before emergence of neural networks [3]. We evaluate SVM with linear kernel on our data in section 4, where it shows good results.

We have also tried TF-IDF data with several other classifiers, such as neural networks, but found the results to be worse than that with Support Vector Machines. This is due to the fact that neural networks overfit quite easily on sparse data like TF-IDF, while SVMs are unable to achieve perfect fit on it and thus act as a "natural" regularizer.

C. Word embeddings and Recurrent Neural Networks

With the introduction of Mikolov et al. [4] paper in 2013, the vector of NLP research turned towards word embeddings. The idea behind word embeddings is to represent each word with a vector, rather than the entire document as TF-IDF does. The general idea behind how these embeddings are computed is that words that occur together a lot should have similar values (as determined by an appropriate loss function), while those that rarely occur together should be different. Refer to [5] and [4] for more details about the training process. Word embeddings are an excellent candidate to be used with recurrent neural networks (RNN), with each embedding vector typically used as one of the inputs to the first RNN in the stack.

One of the libraries that provides an efficient way to compute word embeddings is fastText [6]. In addition to that, it provides an out of the box classification solution, which we are going to evaluate on our datasets.

The architecture of classification fastText is a vanilla many-to-one RNN that takes word embeddings as inputs, and the resulting output fed into a Softmax classifier.

D. Solution by DeepTriage

Another solution using recurrent networks and word embeddings, that was built to perform bug triaging is DeepTriage from [7]. Considering a similarity of their use case (trigaging can be considered as an extreme multi-class classification, with

number of classes being in the hundreds), it is an excellent candidate to benchmark against.

The architecture of DeepTriage consists of a bidirectional RNN, followed by two fully-connected layers. While in the paper they mention using soft attention modules, the provided code does not use them by default. DeepTriage is going to be used as another candidate for our benchmark.

III. PROPOSED APPROACH

The solution that we propose is based on using hierarchical attention paradigm with varying Gated Recurrent Unit (GRU) [8] cell sizes, and a shallow network that goes alongside. This allows network to outperform regular hierarchical attention on datasets where simpler term-based approaches work well.

A. Preprocessing pipeline

Before we begin the overview of our solution, it's worth mentioning the preprocessing that we have done with the data. This preprocessing routine was shared across all approaches in the benchmark, as each of them has shown better results on the preprocessed data. In general, data cleaning is an extremely important step when developing a machine learning solution, and this is especially true for data in customer service systems and task trackers. For example, in the two datasets that we will be using for benchmarking, people often include stacktraces and error messages. And while error messages carry some weight, stacktraces in most cases are meaningless numbers that only add noise. At Forethought Technologies, we have seen similar issues with customer support tickets, which often contain HTML snippets.

Our final pipeline includes the following steps:

- Casting everything to lowercase.
- Removing stopwords.
- Filtering dataset-specific garbage. This was done by custom regex expressions created upon inspection of the datasets.

We have tried several other common preprocessing routines such as stemming, but they have not led to improved results.

B. Hierarchical Attention

Hierarchical Attention is an approach proposed by Yang et al in [1] and it consists of two key ideas - use of sentence hierarchy, and use of attention vector.

The idea of using sentence hierarchy means that we are going to use one RNN that takes in word embeddings from a particular sentence as inputs and compute another vector that acts as a representation of that sentence. Afterwards, a second RNN will take those sentence vectors, and compute a final vector for the document, that will be passed into the Softmax layer to derive final probabilities. Considering that language is structured in sentences, this paradigm works quite nicely in practice, with authors able to beat the best result on Yahoo Q&A dataset with using this approach alone (and without the use of attention vectors). Figure 2 contains an architecture diagram from the paper.

Considering that documents in a dataset can often follow a particular structure (e.g. the most important information is located in the end), it would be good to have appropriate coefficients for outputs of both word and sentence encoders. This task is done by introducing attention vectors, which are marked as u_s and u_w in the diagram. They are shared across all outputs at their level and are trained alongside other parts of the model. When it comes to combining sentence or word vectors into one, the coefficient that we are going to use will be a dot product of an appropriate attention vector with sentence or word. So attention vector serves as "the ideal vector", which if present would achieve perfect score. Using attention vectors in our networks makes a lot of sense, as data is often well structured, with a significant portion of items in the Linux Bugs Dataset (see section 4) filling out a predefined template for their bugreport. Such scenario is perfect to be used with attention-based mechanisms.

It is worth mentioning that idea of using hierarchy for detection and classification is not new and has been successfully applied in other fields, such as visual object recognition [9].

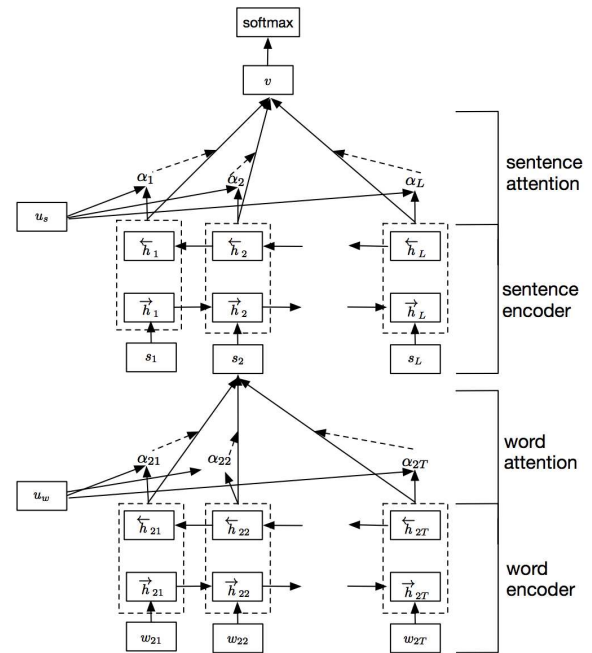


Fig. 2. Architecture of network in Hierarchical Attention paper

C. Network Architecture

As you will see in the benchmark, hierarchical attention does well on a dataset that contains well structured data (Linux bugs), but performs poorly on a dataset where such structure doesn't exist. To combat these problem, we introduce two changes.

First, we are going to use several hierarchical attention blocks like the ones you see on Figure 2, with each of them having a different GRU cell size. Hierarchical attention uses GRU cells rather than more traditional Long Short Term

Memory cells (LSTM) [10], claiming higher performance, albeit with a small margin. The architecture of one such block is depicted on Figure 3. These Deep Attention Blocks will be used for both word-level and sentence-level processing afterwards.

Second, we are going to introduce an additional "shallow" network that is just a simple RNN that takes in word embeddings and produces one vector. It is using GRU cells as well.

Afterwards, the outputs from the shallow network and deep attention blocks are stacked together and go into the fully connected layer, and then into the Softmax layer that produces the final result.

We use cross-entropy loss function for training:

$$Ly'(y) := - \sum_i y'_i \log(y_i)$$

where y_i is the predicted probability value for class i and y'_i is the true probability for that class.

Our network structure is presented on Fig. 4.

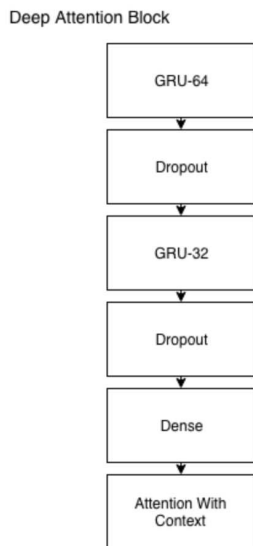


Fig. 3. Deep Attention Block

D. Training Details

Training deep neural networks can often be a finicky task, so we would like to mention several details from training our solution. First, we extensively use dropout [11] to avoid overfitting, which could easily happen considering datasets sizes. What is even more interesting is that we found dropout probability to work best when set at around 1/2, which is higher than typical values. Dropout layers are present in between any two RNN or affine layers in our solution (e.g. see Figure 3). Second, we use RMSprop [12] method for optimization. Lastly, word embeddings that we use for our RNNs are computed by Word2vec.

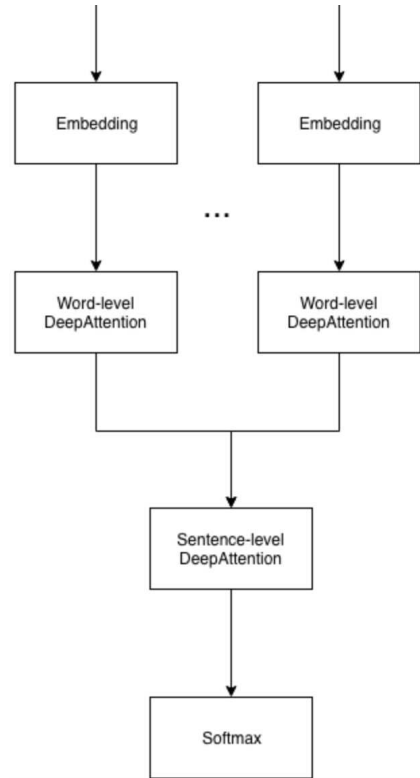


Fig. 4. Our Network Architecture (without an auxiliary shallow network)

IV. COMPARATIVE BENCHMARK

Next, we would like to explore how the aforementioned solutions, including our own, perform on the real data. For this purpose, we have collected two datasets, that are made public for further research. They and the code behind solutions in the benchmark is available at <https://github.com/Forethought-Technologies/ieee-dsmp-2018-paper>.

A. Arch Linux bugtracker dataset

The first dataset that we are going to use contains bugs from open Arch Linux bugtracker at <https://bugs.archlinux.org>. We wrote a simple web scraper to acquire this data. Just like with other bug tracking systems, authors label bugs in terms of various criteria, and two such criteria that we are going to predict are priority and product area. It is easy to see a practical use case for a system that can predict such fields, as the former allows to establish priority of which bugs to fix first and the latter allows to pinpoint the team best suited for the task. Priority field has 9 classes (from P1 high to P3 low), the product field has 16 classes (Network, Drivers, etc.). Below is an example of a bug from this dataset:

Title: *i2o_scsi does not handle reset properly*

Content: *The i2o scsi driver should sleep in the reset handler until the i2o reset message is replied to by the firmware. James has discussed infrastructure to make this generic*

Priority: P2_low (9 classes)

Product: Drivers (16 classes)

The complete dataset contains 16,456 entries.

B. Chromium bugtracker dataset

The second dataset was adapted from the Chromium dataset used by DeepTriage paper. The field that we are going to predict is called "Type" and can be one of Bug / Feature / Compatibility issue.

Title: *Scrolling with middle-mouse button does not work (autoscroll)*

Content: *Product Version: chrome beta 1 URLs (if applicable) : Other browsers tested: Add OK or FAIL after other browsers where you have tested this issue: Safari 3: OK Firefox 3: OK IE 7: OK What steps will reproduce the problem? What is the expected result? Clicking the middle-button on the mouse should show a ""fast scroll"" feature. What happens instead? Nothing. Please provide any additional information below. Attach a screenshot if possible.*

Type: Feature

The complete dataset size contains 58,871 entry.

C. Evaluation methodology

For each solution we measure two results - accuracy and weighted F1 score. The results are computed on the test set, with test set size being 15% of the original data.

Each classifier had optimal hyperparameters picked via a usual grid search.

TABLE II: Benchmark of accuracies

Method	Linux bugs: Importance (9 classes)	Linux bugs: Product (16 classes)	Chromium bugs: Type (3 classes)
Naive Bayes ¹	51.6%	45.6%	80.5%
TF-IDF with SVM	65.0%	61.6%	80.5%
fastText	64.2%	58.7%	82.2%
DeepTriage ²	61.4%	63.8%	81.6%
Hierarchical Attention (regular)	66.4%	58.9%	75.9%
Our Solution	69.1%	58.7%	88.2%

TABLE III: Benchmark of F1 scores

Method	Linux bugs: Importance (9 classes)	Linux bugs: Product (16 classes)	Chromium bugs: Type (3 classes)
Naive Bayes	0.479	0.411	0.787
TF-IDF with SVM	0.568	0.590	0.804
fastText	0.542	0.579	0.821
DeepTriage	0.516	0.604	0.816
Hierarchical Attention (regular)	0.573	0.574	0.758
Our Solution	0.579	0.567	0.879

D. Results

Results are presented in tables 2 and 3. We can make several conclusions from these results:

- Naive Bayes is not a good solution for multi-class classification.

- Decade old solutions like TF-IDF with SVM still show pretty good results and can be a great solution for cases where resources are limited.
- Novel solutions can outperform the classic ones on each dataset.
- Our solution shows superior performance, especially on the last task. However, it did poorly when predicting the product field from the first dataset.

V. CONCLUSION

In this paper we went over a variety of methods used for text classification, presented our own solution based on hierarchical attention paradigm and benchmarked these solutions on two real datasets. We see that novel approaches that use RNNs on word embeddings data outperform the classic solutions, which opens doors to many practical use cases. Nevertheless, the resulting accuracies are still far from perfect and multi-class text classification remains an open problem.

REFERENCES

- [1] Z. Yang, D. Yang, C. Dyer, X. He, A. J. Smola, and E. H. Hovy, "Hierarchical attention networks for document classification," in *NAACL HLT 2016, The 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, San Diego California, USA, June 12-17, 2016*, 2016, pp. 1480–1489. [Online]. Available: <http://aclweb.org/anthology/N/N16/N16-1174.pdf>
- [2] V. Metsis, I. Androutsopoulos, and G. Paliouras, "Spam filtering with naive bayes - which naive bayes?" in *CEAS*, 2006.
- [3] I. Pilászy, "Text categorization and support vector machines," 2005.
- [4] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013. Proceedings of a meeting held December 5-8, 2013, Lake Tahoe, Nevada, United States.*, 2013, pp. 3111–3119. [Online]. Available: <http://papers.nips.cc/paper/5021-distributed-representations-of-words-and-phrases-and-their-compositionality>
- [5] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," *arXiv preprint arXiv:1301.3781*, 2013.
- [6] A. Joulin, E. Grave, P. Bojanowski, and T. Mikolov, "Bag of tricks for efficient text classification," *arXiv preprint arXiv:1607.01759*, 2016.
- [7] S. Mani, A. Sankaran, and R. Aralikatte, "Deeptrriage: Exploring the effectiveness of deep learning for bug triaging," *CoRR*, vol. abs/1801.01275, 2018. [Online]. Available: <http://arxiv.org/abs/1801.01275>
- [8] K. Cho, B. Van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using rnn encoder-decoder for statistical machine translation," *arXiv preprint arXiv:1406.1078*, 2014.
- [9] D. B. Walther and C. Koch, "Attention in hierarchical models of object recognition," *Progress in brain research*, vol. 165, pp. 57–78, 2007.
- [10] S. Hochreiter and J. Schmidhuber, "Long short-term memory," vol. 9, pp. 1735–80, 12 1997.
- [11] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *The Journal of Machine Learning Research*, vol. 15, no. 1, pp. 1929–1958, 2014.
- [12] T. Tieleman and G. Hinton, "Lecture 6.5—RmsProp: Divide the gradient by a running average of its recent magnitude," COURSERA: Neural Networks for Machine Learning, 2012.

#Euromaidan: Quantitative Analysis of Multilingual Framing 2013-2014 Ukrainian Protests on Twitter

Yehor Lyebyedyev
Perfectial
Lviv, Ukraine
gor.qkop@gmail.com

Mykola Makhortykh
Department of Slavonic Studies
University of Amsterdam
Amsterdam, Netherlands
makhortykh@yahoo.com

Abstract—In this paper we investigate the use of social media for framing the Euromaidan protests in Ukraine. Using automatic classification of a large set of Twitter data, we investigate how the online representation of protests changed between different stages of the *protest* campaign; furthermore, we question how the framing of Euromaidan varied between different language streams. Our findings suggest that framing of Euromaidan on social media evolved from a peaceful movement to revolutionary force and then to existential danger to the Russophone population, prompting the continuation of political crisis in Ukraine and the annexation of Crimea.

Keywords—*automatic classification; Twitter; Ukraine; protests; framing*

I. INTRODUCTION

In recent years social media became increasingly used by protest movements, varying from the Green Movement in Iran [1] and Los Indignados in Spain [2] to Occupy Wall Street in the United States [3] and For Fair Elections in Russia [4]. The accessibility and wide reach of social media outlets not only provide protesters with logistical support, but also facilitate public mobilization by allowing activists to disseminate information about their agendas and broadcast their identity to the world. In our article we examine how the polyphony of social media affected the representation of the Euromaidan protest movement in Ukraine between January and March 2014. As Onuch and Sasse [5] argue, the Euromaidan movement was significantly influenced by the use of social media, which were employed by pro-Western activists for public mobilization since the beginning of protests in November 2013. A number of scholars [6, 7, 8] point that on the later stages the Euromaidan protest campaign was also accompanied by the extensive use of social media by activists and bystanders; yet, the majority of existing studies either omit the question of online framing of the protests or focus on short time periods, such as the peak of violence in Kyiv between February 18 and 22 [7]. By contrast, our study employs a large set of data to examine how Euromaidan's image developed over time and discuss its changes as the degree of confrontation between pro- and anti-government forces fluctuated in the course of protests. In addition to examining how the image of Euromaidan evolved from January to March, our paper investigates how the representation of protests changed depending on the audience's language. Using automatic classification of protest-related content on Twitter, we compare how the

Euromaidan protest campaign was represented in English, Russian, and Ukrainian languages and examine how linguistic differences affected the understanding of the political crisis in Ukraine.

II. METHODOLOGY

A. Data collection

In order to examine the use of social media for framing Euromaidan, we employed data from Twitter. While Twitter was not the most popular social media platform in Ukraine before 2013, a number of studies point to its particular importance in the context of protests, both in terms of amount of content generated [9] and the involvement of users, including the ones who never employed the platform before [7]. These observations together with the relative ease of obtaining data from Twitter, when compared with other popular social media platforms, explain our decision to use Twitter for implementing our study. We used Twitter Streaming API for collecting data in real-time mode from January 11 to March 14 2014. The beginning of data collection relates to the increase of protest-related online activity, following a series of clashes between protesters and police in Kyiv in the beginning of January. The end date corresponds to the significant decrease of Euromaidan-related activity on Twitter, caused by the switch of audience's interest to the events in Crimea and Eastern Ukraine. As criteria for data selection we used the presence of term "Euromaidan" in English, Russian, and Ukrainian in the body of a tweet; based on this criteria we collected 1.200.000 messages, which became our data set for this study.

B. Data analysis

In order to examine how Euromaidan was framed in each of major language streams, we divided our data set into English, Russian, and Ukrainian subsets which included all messages written in a respective language. The division was made on the basis of Twitter API language identifier, which automatically determined the language of a particular tweet. Then, we divided each language subset into chronological subsets, using fluctuations of online activity as indicators of framing cycles on Twitter. As Fig. 1 shows, we distinguished six distinct periods in the protests' framing; in some cases the transition was marked by the dramatic change in the scope of Euromaidan-related activity (e.g. the transition between the

first and the second periods), whereas in other cases it depended on the change in proportions between different language streams (e.g. the transition between the fifth and the sixth periods). Because of the large volume of available data, we employed automatic classification for exploring how Euromaidan was framed on Twitter. For this purpose, we used a naive Bayes classifier which was applied to all unique tweets and retweets in each of three language streams. We assumed that tweeted content will be closer to natural

communication, whereas retweeted content will include the large number of messages produced through organized Twitter campaigns; by comparing them, we wanted to check possible differences in framing Euromaidan between these two streams. N-grams were generated for tweets and retweets in each language; then, we compared the probability of n-gram occurrence to identify subjects, which were the most and the least common for tweets and retweets in each particular language.

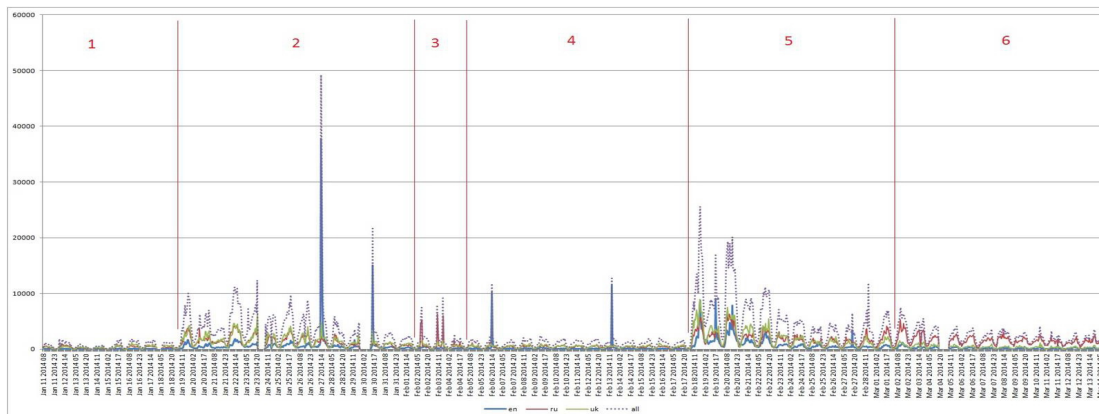


Fig. 1. Number of tweets per day by language

III. FINDINGS

1) *First period.* The first period (11-18 January) was characterized by low activity on Twitter that contrasts with a number of protest-related events which occurred during this time, including the revival of clashes between protesters and police in Kyiv and the adoption of the anti-protest legislation. Yet, neither of these events spawned any significant activity on Twitter; instead, the communication in all language streams was dominated by retweeted content. Based on the ratio of retweets to tweets (the lowest in the Russian stream, the highest in the Ukrainian one) we suggest that natural communication between in relation to Euromaidan mostly occurred in the Russian stream, whereas messages in the English and Ukrainian ones were more actively used for dissemination of selected protest frames through retweets.

In the Russian stream the most common subjects for tweets and retweets included terms referring to the peaceful nature of protests (мирн протест), which contrasted with the use of force by police (беркут луценк) and the repressive legislation adopted by the government (закон колесничен). Compared with retweets, tweets in Russian often included negative assessments of the protests; often, these assessments compared protesters with Nazi (евромайда нацизм) or wild animals (евромайда еврозвер). Conspiracy theories, including references to the psychological warfare (информационнопсихологическаявойн евромайда), also were more present in tweets; together, these observations can be viewed as an indicator of emerging anti-Maidan activity. Compared with the Russian stream, the Ukrainian one referred more often to the Ukrainian political landscape, including specific politicians (олег тягнибок) and issues (тимошенко евромайдан); such references were common both for tweets and retweets. Contraposition of protesters' peaceful actions with authorities' brutality was another common subject; while, the subject of repressive legislation was mostly ignored, the clashes between protesters and police were covered in more details (києво святошинським, побиття луценка). These articulations of police violence

were supplemented with calls for more active participation in the protests (київ вставай) and public mobilisation (потрібна допомога). The only difference between retweets and tweets in the Ukrainian stream related to more explicit framing of the ongoing protests as the patriotic revolution (єврореволюція еврореволюція) among tweeted messages.

Similar to the Ukrainian stream, the English one included little variation between tweeted and retweeted content. Unlike other two streams, the English stream rarely referred to specific political issues; instead, it was used for reporting the ongoing protests in different parts of Ukraine (ralli kyiv, kharkiv euromaidan), thus framing Euromaidan as an all-Ukrainian movement. Messages which called for help (help pleas) and mobilization (spread word) were also common for tweets and retweets in English well as references to the importance of the protest for Ukrainians (ukrain futur, make histori) and brutal actions of authorities (sweep crack, human right). The major difference between retweets and tweets involved a few episodes of the protests, references to which were actively promoted through individual tweets; one of such episodes was a conversation between policemen and an old woman with a mirror in her hands (hold mirror, mirror polic).

2) *Second period.* The second period (19 January-1 February) was marked by the rapid increase in the number of messages in all language streams, in particular the Russian and Ukrainian ones. This increase can be contributed to the growing number of retweets, which constituted the majority of messages in all three streams. Such a dramatic change in Twitter activity corresponded to the beginning of Hrushevskogo street riots in Kyiv; while after the end of riots on January 22 the scope of online activity decreased, it still remained higher than before. Two single peaks of activity in the English language stream (January 27 and 30) corresponded to two "Twitter storms" (Watts 2012), organized by the #DigitalMaidan activist group for attracting the Anglophone Twittersphere's attention to the events in Ukraine.

In the Russian stream the most frequent subjects included violence on the Hryshchivskogo Street (горя шин, територія динам) and brutality of the police (беркут убиває, погибш майдан). Another common topic for both tweets and retweets was the role of Ukrainian and Russian political figures in the crisis: Klichko, Yanukovich, and Putin were referenced most often, yet only the former was presented in a positive light (кличк пита), whereas the latter two were referenced more negatively (путенск деньг, янукович поїдет). Similar to the first period, a number of tweets involved negative assessments of the protests, calling to boycott the pro-Euromaidan artists (океа бойкот) and pointing to the presence of Nazi (нацист вромайда) and snipers on the Maidan (чита снайпер). These negative frames were counteracted by another distinct group of tweets, which expressed mourning and called to pray for the victims of the Hrushevskogo Street (слав иисус, неб сто).

Unlike the Russian stream, where the main emphasis was made on the street riots, the Ukrainian stream paid meager attention to the clashes on Hryshchivskogo; instead, the stream focused on the government's actions (від'їжджають автобуси, титишки вромайдан), in particular, repressions against activists (козак михайло, увага зник). The contraposition of peaceful protests and the authorities' brutality (мирну акцію, злочинної влади) was a common topic in the Ukrainian stream; however, a number of messages calling for violent responses to the government's reprisals (народний гнів, захистити всіх) also appeared during this period. Unlike retweets, tweets mostly referred to the oppositional political figures – both the Ukrainian (удар кличко) and the Russian (кличко навалын) one – and emphasized the revolutionary nature of the protests (єврореволюція єврореволюция).

Unlike the first period, the English stream shared the least number of common topics between tweets and retweets; these shared topics mostly evolved around the use of force by police (polic shoot, polic beat) and the government's attempts to stabilise the situation (ukrain accord, ukrain govt). Similar to the Ukrainian stream, violence on Hrushevskogo remained mostly ignored together with the activists' deaths; instead, the large number of tweets called for sanctions against the Ukrainian government (sanction ievgzdr, sanction fixnrjtb). These calls, however, were rarely retweeted, because the respective messages were mostly produced in the course of Twitter storms. Consequently, many of these tweets were automatically generated and included a number of nonsensical terms, which made them less appealing for retweeters.

3) *Third period.* The third period (2-4 February) was distinguished by the predominance of Russian tweets. This change can be explained by the mass demonstration in Moscow that took place on February 2 in support of Euromaidan. Compared with the previous period the amount of retweets significantly decreased: in the Ukrainian and English streams the ratio of tweeted and retweeted content dropped to the level of the first period, whereas in the Russian stream it became the lowest for the whole period of observations. This significant shift towards tweeted content in the Russian stream can be attributed to the increased interest of Russophone users towards Euromaidan articulated by the extensive media coverage of the Moscow demonstrations.

In the Russian stream the majority of message concerned the mass demonstration in Moscow (марш москв), which was viewed in as an indicator of support of Euromaidan among Russians (москв поддежда, поддежда люд). Such an interpretation was predominant both among tweeted and retweeted content; similarly, both tweets and retweets referred to the tragic death of Serhiy Nigoyan, a Euromaidan activist, who was killed in Kyiv during the Hrushevskogo street riots (серге нигоя), and the growing unease in relations between Ukraine and Russia (войн украин, черн мор).

Unlike the Russian stream, the Ukrainian one did not mention the demonstration in Moscow; instead, both tweets and retweets here were focused on the victims of the Hryshchivskogo riots. In contrast to the Russian stream, which usually mentioned Serhiy Negoyan, messages in Ukrainian mostly referred to the Mikhail Zhiznevsky (жизневського куля, товариш жизневського), a Belorussian activist and a member of UNA-UNSO, a far-right Ukrainian party. Another common subject for both tweets and retweets involved calls for public mobilization (бийся українець, допомоги вромайдан).

In the English stream two topics were common for tweets and retweets: the suffering of Ukrainians and political/public figures. Messages on the first topic emphasized the criminal nature of Yanukovich regime (million afraid, tortur victim) and highlighted the goals of protesters (free ukrain, democrat world). The second subject was represented by messages which referred to pro- (irina berezhna, akhmetov hide) and anti-government personalities (vitali klitschko); among these personalities, Petro Poroshenko, the future President of Ukraine, was mentioned for the first time (poroshenko say, poroshenko address). Unlike retweets, tweets contained a number of calls for violent action (fighter resist, kill fight) and blamed supporters of government for sexual abuse of children (abus children, freemason pedophil).

4) *Fourth period.* Similar to the first period, the fourth period (5-17 February) was characterized by the low degree of activity on Twitter, which can be explained by the lack of significant changes in the course of the protests; instead, the major news makers during this period were Russian and American public figures, who blamed each other for destabilization of the situation in Ukraine. Similar to the first and the third periods, the number of tweets significantly exceeded the number of retweets; also, similar to the first weeks of January, the Ukrainian stream again became the most prolific one. The only exception from this pattern was related to two Twitter storms (February 6 and 13), which were organized by the #DigitalMaidan activist group in the English stream.

In the Russian stream the significant differentiation of topics between tweeted and retweeted content was observed during this period. The number of common subjects decreased: the only themes which appeared both in tweeted and retweeted content were the activity of Automaidan (активист автомайда) and the confrontation between protesters and special police forces (вромайда беркут). The worsening economic situation in Ukraine (євромайда кред) was the main subject of Russophone tweets, whereas retweets openly criticize the protests by claiming that the revolution depleted itself (кризис вромайда, себ исчерпа) and predicting that protesters soon would be defeated (зачистка майдан).

The Ukrainian stream remained more focused on the latest developments in relation to the protests, but it also showed increased differentiation between tweeted and retweeted. Only two common subjects for tweets and retweets were the activities of Pravyi Sector (правий сектор) and the standoff on Hryshkevskogo (барикади грушевського). Similar to the Russian stream, Ukrainian tweets demonstrated the decrease of interest towards protesters' agendas and instead focused on the developments in selected Ukrainian regions (львів революція) as well as confrontations with special police forces (беркут революція). By contrast, retweets tried to revive public interest by communicating recent developments, in particular, the meeting of Yanukovich with former Ukrainian presidents (президенти кравчук), and transmitting protesters demands to free Timoshenko (автомайдан Тимошенко).

The English stream experienced the least differentiation between tweets and retweets and continued reporting recent developments from Ukraine. Two common subjects for tweets and retweets during this period concerned the scandal related to the recorded conversation between two American diplomats, Victoria Nuland and Geoffrey Pyatt (frustrat nuland, alleg leak), and protests in different parts of Ukraine (euromaidan kyiv, odessa euromaidan). Two distinct subjects among tweets were calls for sanctions against the Ukrainian officials (same sanction), and calls for Western politicians to help Ukraine (johnkerri pass, ukrain help); like earlier, many of these messages were automaticall generated in the course of Twitter storms and were not particularly appealing for retweeting.

5) *Fifth period.* The fifth period (18 February-1 March) was distinguished by the rise of Twitter activity. Similar to the second period, this activity peak can be attributed to the outbreak of violence, in particular the beginning of clashes between protesters and police in Kyiv on February 18-21, which left dozens of people dead and wounded. This period was also marked by the growing activity in the Russian stream; unlike the Ukrainian and English streams, which become less active after the assignment of new pro-Western government and the flight of Yanukovich, the amount of Russophone tweets increased in the end of period.

In the Russian stream tweets and retweets were focused on the confrontation in Kyiv with a special emphasis on human losses (институтск расстрел, погибш ранен). The discussion of Ukrainian political figures – both the pro-government (министр захарченк) and the opposition ones (оробец вромайда) – was another common subject together the role of Russia in the ongoing events (господ российск, росс путин). A distinct topic among tweets was related to the activities of MMM (мммщик вромайда) and the alleged involvement of a popular blogger Yuri Khovanski (хованск вромайда) in the pyramid's activities. By contrast, retweets were focused on the Automaidan's activities (автомайда саф) and mourning of the confrontation's victims (траур сцен).

In the Ukrainian stream the set of common topics for tweets and retweets was similar to the one in the Russian stream. The confrontation between protests and police in Kyiv (загиблих поранених, пекло євромайдан) and the removal of Yanukovich (негайна відставка) together with the dismissal of potential compromises (ніяких угод) were common topics for tweets and retweets. Similar to the

Russian stream, Ukrainian tweets included references to MMM and Yuri Khovanski (мммщиком євромайдан); unlike it, a number of Ukrainophone messages shared the recent developments beyond Kyiv, in particular in Western Ukraine (рівняни євромайдан). The main subjects among retweets were potential deals between the government and the opposition (опозиція азаров) and calls for the further radicalisation of the protest and reprisals against Communists (комунякам гілляку).

Similar to the other two streams, the English one was focused on the confrontation in Kyiv (kyiv death, blood ukrainian); however, unlike the Russian one, the significant attention is also paid to the events in Crimea and the role of Putin (watch putin, start crimea). A number of retweets were used for promoting certain YouTube videos (youtub messag) and calling for sanctions against Ukrainian oligarchs, in particular Rinat Akhmetov (akhmetov sanction) Unlike retweets, a distinct group of tweets pointed to the involvement of Nazi in the protests (maidan nazi, nazi news), which can be interpreted as an attempt of anti-Maidan activists to communicate their views to the Anglophone audience. At the same time, the large number of tweets called for sanctions against Yanukovich (euhv yanukovich, stop dictatorship); like earlier, these messages were produced in the course of #DigitalMaidan campaigns.

6) *Sixth period.* The six period (2-14 March) was characterized by the predominance of Russian tweets, which can be related to the end of Euromaidan protests and the shift of the unrest's focus from the mainland Ukraine to Crimea.

In the Russian stream a common topic both for tweets and retweets was the criticism of the Russia's actions in Crimea (имперск лоботом, пикетирова посольств); however, this criticism was contrasted by references to the alleged threats to the Russophone population, including the Pravyi Sector (вромайда правыйсектор) and Tatars (крым татар), and activities of pro-Russian activists in Crimea (борис рожин). A number of tweets framed the Crimean crisis as a direct consequence of Euromaidan by referring to the rise of violence and chaos in Ukraine (хаос територ, захват пост); another common subject among tweets was the threat of potential Ukrainian attack against Cri (крым штурм, крым гашн).

The Ukrainian stream was focused on the crisis in Crimea, in particular the Russian involvement (сусідів ласих, окупанти погрожують); a number of messages drew parallels between events in Crimea and annexations of territories by Hitler (гітлером німецькомовних). Another common subject for both tweets and retweets in Ukrainian were the pro-Russian protests in Eastern Ukraine, especially in Donetsk (донецкий сепаратизм, губарева арештували). In both cases actions of Russian military forces and pro-Russian activists were viewed as an attempt to extract revenge for the victory of Euromaidan; consequently, the confrontation with Russians was viewed as a continuation of Euromaidan struggle.

The English stream involved two subjects, which were common for tweets and retweets: the Russian aggression in Crimea (kidnap crimea, russiainvadesukrain crimea) and perspectives of the pro-European rallies in Russia (russian euromaidan, moscow euromaidan). By contrast, retweets were mostly focused on dissemination of specific news items such as the ones on Euromaidan activists in Crimea (crimea

euromaid) or fake news about the golden toilet at the Yanukovich residence (toilet yanukovych). Among tweets, however, the subject of protests became less present; the majority of references to Euromaidan involved foreign reactions to the change of government in Kyiv (houseforeign euromaidan, carlbildt euromaidan).

IV. CONCLUSIONS

Similar to the earlier studies [10, 11, 12], which identified significant variation in the representation of protest activities in different languages; we found that various language streams on Twitter propagated varying interpretations of the protests in Ukraine. The Ukrainian stream, which was actively used for public mobilization, presented the Euromaidan protests as a Ukrainian revolutionary movement; these revolutionary frames were supplemented by a number of historical references to the national liberation movement, in particular the Ukrainian Insurgent Army. In general, this stream was the most positive towards Euromaidan and tended to provide rather binary image of protests (e.g. by emphasizing brutality of police and ignoring the use of violence by protesters).

The English stream, which was used for internationalization of the protests, was also mostly positive about Euromaidan; yet, unlike the Ukrainian stream, which framed the protests as a national – or even a nationalistic movement – it emphasized the struggle for universal values such as human rights and democracy. The Russian stream was the least influenced by the pro-Maidan online campaign and offered the most varied coverage of the protests which included both pro- and anti-Maidan voices. Consequently, the Russian stream propagated a more nuanced interpretation of the protests, including both expressions of mourning for human losses and concerns about the negative consequences of removing Yanukovych from power.

While our observations point to a number of distinct features of Euromaidan framing in each of the language streams, neither of these streams was homogeneous in its interpretation of the protests; instead, these interpretations experienced significant changes in the course of the protests. Some of these changes were episodic ones (e.g. the decrease of anti-Maidan messages in the Russian stream during the periods of violence), whereas others can be explained by the gradual evolution of the protests' image (e.g. the transformation of the Euromaidan's image from a peaceful protest to a forced resistance in the Ukrainian stream). The most illustrative example of the latter type of frame evolution is found in the Russian stream, where framing of Euromaidan evolved from a sympathetic image of peaceful struggle for human rights to the potential source of the economic decline and then of the existential danger to the Russophone population in Crimea.

Finally, we found a number of distinctions between frames, which were disseminated through tweeted and retweeted content. Initially, we expected that retweets would be more representative for staged communication (i.e. online protest campaigns), whereas tweets would demonstrate how frames evolve in natural communication (i.e. day-to-day communication among Twitter users); however, these expectations were not supported by our data. Instead, we found that those subjects, which were common both for tweets and retweets (e.g. frames of suffering in the Russian

stream), seem to be more representative for natural communication; by contrast, those subjects, which were prevalent only in one group of content, often were artificially propagated through retweeting networks or Twitter storms.

Together, our findings prompt for more nuanced understanding of social media framing of protest movements, including the Euromaidan protests in Ukraine. The velocity and reach of social media not only accommodates different interpretations of protest activity, but also makes these interpretations easier to adapt as the time goes by. While these features of framing protests on social media make them beneficial for short-term mobilization, they also create a dangerous possibility of forming divergent views on the protests among different audiences. Not only these contradictory interpretations can result in highly varying expectations of the protest consequences, which in the case of Euromaidan varied from the restoration of human rights and removal of corrupt government to the economic decline and existential threat to the Russophone population, but also facilitate the manipulative use of established protest frames after the end of protests like it happened in the case of the Crimea annexation.

REFERENCES

- [1] A. Fisher, *Bullets with Butterfly Wings: Tweets, Protest Networks, and the Iranian Election*. In Y. Kamalipour (Ed.), *Media, Power, and Politics in the Digital Age: The 2009 Presidential Election Uprising in Iran*, 2010, pp. 105–118. Lanham: Rowman and Littlefield, Inc.
- [2] E. Anduzaa, C. Cristanchoa, J. Sabucedo, *Mobilization through online social networks: the political protest of the indignados in Spain*. *Information, Communication & Society*, vol. 17 (6), pp. 750–764, 2014.
- [3] M. Tremayne, *Anatomy of Protest in the Digital Era: A Network Analysis of Twitter and Occupy Wall Street*. *Social Movement Studies*, vol. 13 (1), pp. 110–126, 2014.
- [4] G. Nikiporets-Takigawa, *Tweeting the Russian protests*. *Digital Icons: Studies in Russian, Eurasian and Central European New Media*, vol. 9, pp. 1–25, 2013.
- [5] O. Onuch, G. Sasse, *What does Ukraine's #Euromaidan teach us about protest?* *Washington Post*, 2014. Retrieved from <http://www.washingtonpost.com/blogs/monkey-cage/wp/2014/02/27/what-does-ukraines-euromaidan-teach-us-about-protest/>
- [6] I. Kozachenko, *How social media transformed pro-Russian nostalgia into violence in Ukraine*. *The Conversation*, 2014. Retrieved from <http://theconversation.com/how-social-media-transformed-pro-russian-nostalgia-into-violence-in-ukraine-33046>
- [7] J.A. Tucket, M. Metzger, D. Penfold-Brown, R. Bonneau, J.T. Jost, J. Nagler, *Protest in the Age of Social Media*. *Carnegie Reporter*, vol. 7 (4), 2014. Retrieved from <https://medium.com/@carnegiecorp/protest-in-the-age-of-social-media-7ae9fd940b06>
- [8] A. Gruzdz, K. Tsyganova, *Information Wars and Online Activism During the 2013/2014 Crisis in Ukraine: Examining the Social Structures of Pro- and Anti-Maidan Groups*. *Policy & Internet*, vol. 2 (7), pp. 121–158, 2015.
- [9] B. Etling, *Russia, Ukraine, and the West: Social Media Sentiment in the Euromaidan Protests*. *Internet Monitor Special Report Series*, 2014. Retrieved from <http://cyber.law.harvard.edu/publications/2014/euromaidan>
- [10] F. Jansen, *Digital Activism in the Middle East: Mapping Issue Networks in Egypt, Iran, Syria and Tunisia*. *Knowledge Management for Development Journal*, vol. 6 (1), pp. 37–52, 2010.
- [11] T. Poell, K. Darmoni, *Twitter as a Multilingual space: the articulation of the Tunisian revolution through #sidibouzid*. *NECSUS - European Journal of Media Studies*, vol. 1, pp. 14–34, 2012.
- [12] M. Lynch, D. Freelon, S. Aday, *Syria's Socially Mediated Civil War*. Washington, DC: U.S. Institute of Peace, 2014.

Big Data Automatic System of Analysis and Trading on Financial Markets

Serhii A. Rybalchenko
Faculty of Economics, Department of Economic Cybernetics
Taras Shevchenko National University of Kyiv
Kyiv, Ukraine
sergiy.rybalchenko@gmail.com

Abstract—The paper considers the main tendencies of the international currency markets and opportunities for individual investments. The methods of decision support of assets purchasing were provided. A mechanism of intellectual trade with the use of a significant data flow is proposed. Open and closed types of information products were combined to automate the system. Guided by the effectiveness of trade signals assessment, recommendations were made for the use of methods of artificial neural networks or logistic regressions. A set of criteria has been formed to support lifecycle of the automated trading system.

Keywords—*automated trading system, artificial neural networks, logistic regressions, international currency markets.*

I. INTRODUCTION

In modern conditions, the international financial market serves as the main indicator of changes in economic environment and economic development of certain regions and industries. All economic projects need financing. Most projects require cheap international investment. In turn, foreign investments come into efficient and safe environment where the balance between risk and yield is maintained.

The modern foreign exchange market began forming during the 1970s. At that time, the main participants of international currency transactions were: the International Monetary Fund (IMF), International Bank for Reconstruction and Development, and the General Agreement on Tariffs and Trade (GATT) [1, p. 169]. With the widespread adoption of the Internet in the 1990s, banks and small companies created online networks to produce automated quotes and allowed for instantaneous trading [2]. Over time Forex became available to individual investors as well. According to the Bank for International Settlements, the preliminary global results from the 2016 Triennial Central Bank Survey of Foreign Exchange and OTC Derivatives Markets Activity show that trading in foreign exchange markets averaged \$5.09 trillion per day in April 2016. This is down from \$5.4 trillion in April 2013 but up from \$4.0 trillion in April 2010. Measured by value, foreign exchange swaps were traded more than any other instrument in April 2016, at \$2.4 trillion per day, followed by spot trading at \$1.7 trillion [3]. These data allow us to perceive Forex as the main indicator and instrument of the international financial market.

The leading status of the market entices significant scientific efforts to develop methods of analysis and forecasting of exchange rates. In addition to some developments carried out or supported by commercial organizations, a lot of material is being developed by

educational institutions. This is due to easy access to the market, historical data and volume of information. The data of the international currency market reflect the real processes of the economy, while at the same time are suitable for student learning proven and the newest methods. Accordingly, students at the relevant departments annually conduct 10-15 courseworks and diploma projects on application of economics and mathematical methods for analyzing and forecasting the exchange rates of the international exchange market. Within this publication, the main directions of scientific research on optimization and automation of trading and other systems are singled out.

The object of study – automated trading and analysis systems. Subject of research – analysis and synthesis method, quantitative method of measuring and comparing, historical and simulation methods. The aim – to create effective automatic system of analysis and trading, based on artificial intelligent and simulation methods. The goal will be revealed through the following objectives:

- Determine the conditions and trends of the international currency markets;
- Outline the main opportunities for individual investments;
- Explain weaknesses of decision support methods;
- Summarize the experience of forming of analysis and trading systems;
- Show effective methods of Big data analysis with open-source statistical applications;
- Develop automatic system of analysis and trading, recommendations for further lifecycle system management.

II. OPPORTUNITIES FOR INDIVIDUAL INVESTORS AND RESEARCHERS

Prior to the propagation of information technologies to the household level, individual investors could only use Forex forwards and futures contracts. With the proliferation of computers, there was an expansion of opportunities for investing. A person with a laptop can easily connect to the international financial system by one of many broker options. In turn, brokers interact with banks and other financial institutions where they can carry out foreign exchange operations at their discretion and at customer's request. If we look towards the implementation of customer contracts, brokers can usually use the mechanism of "clearing". For the same currency pair, there will always be several different

offers from clients: someone wants to buy the first currency of the pair, someone the second one. Even by volume and opposite by direction customer orders are offset and the broker will not have to carry out any currency exchange transactions on the interbank market, and will not pay therefore any the commission. When executing customer transactions, the broker will copy off the loss of the first client to the win of the second, and write down some of the funds on his own account as a commission. If all transactions will not be offset with the opposite ones, the lack of demanded currency will be purchased on the interbank market. Obviously, with the "clearing" system the volume of real transactions can be less than 20%. Such a system could be neutral for the client, if not for abuse by dis-honest brokers. With a neutral system customer does not care about the mechanism for the implementation of transactions, he is set to receive speculative profits. But when the deal is not backed by real market performance, improbable temporary events in the fluctuation of the asset rate may occur. For example, drop by 30% for 1 second with a subsequent sharp return to the normal level. After adoption of auto-mated trading systems by large companies, such cases became a reality [4]. But exceptional and possible only when the system crashes for significant part of market participants. Such market participants are in high-level jurisdictions and customers are protected by the law and reputation of the company. Derived problems arise. Small common brokerage companies purchase informational blocks of currency quotations from banks or other financial institutions. With the proper manipulation of the courses in the "clearing" system brokers shift responsibility on the supplier of quotations. The client will have to enter into a legal dispute. In such circumstances, jurisdiction is important. Such models were common in the 2000s. Although the international currency market does not have a single regulator and is by all signs relatively decentralized, the consumers have strong influence on it. Hence, the market clears up for the benefit of the general good. At this stage, thriving brokerage companies have full execution of customer agreements, i.e. without internal "clearing". Most companies have a level of jurisdiction not lower than "C".

Let's consider the factors of income formation for a client of brokerage company. The basic law of commerce here as well is to buy cheaper and to sell with profit later. Therefore, speculative profit:

$$Income_i = (Price_i - Price_{i-k}) \cdot Volume_i, \quad (1)$$

where $Income_i$ is income in time period i , $Price_i$ is currency price in time period i , k is periods number of order (between open and close order), $Volume_i$ is assets (currencies) volume in time period t .

Under such formal conditions, the break-even point of a speculative investor is in equality:

$$\sum_{j=1}^n Price_i^j = \sum_{j=1}^n Price_{i-k}^j, \quad (2)$$

where j is order index. So,

$$E(Price_i) = E(Price_{i-k}). \quad (2)$$

The interpretation of formulas (1) - (3) consists in the need to correctly guess or predict the direction of price

movement in 50% of cases under neutral conditions ($Price \neq const, Volume_j = const$). This is an acceptable task for both scientific methods and for any coin (situation with coin flip). We call such a situation a point of zero informational advantage. That is, when we know nothing about the history and prospects of the market. All that is needed is the current asset price and the ability to fulfill the condition. It is possible to draw the equivalence between the available information and the probability of correctly predicting the direction of the trend of the asset price.

$$P_{0.5} \approx I_0. \quad (3)$$

In reality, of course, this is not enough. When executing customer order, and closing the order at the same price, the client receives a loss from the executed order. The nature of the negative difference at an even price lies in the broker commission, currency spread and other payments.

Brokerage commission is the main source of income for financial intermediation. Clearly exhibited commission rates have been the subject of fierce competition between brokers. And 0% became minimum completion of lowering of explicit commission for a client. Most FOREX brokers do not demand commission fees for individual investors with small volumes. Trading in amounts greater than \$100,000 is possible with fixed fees for each transaction, relative to the amount of the transaction, periodic as well for the implementation of individual services. For investors with small amounts, the source of the broker's profit is transferred to the spread. The spread is the difference in the buy and sell price of any asset or currency pair [5]. Buying an asset at price P_{-1} we can sell an asset in the same moment at lower price P_{-2} . The difference between the prices of P_{-1} and P_{-2} is the spread of the broker. It can be fixed or relative, depending on the time of day and market volatility.

Brokers measure spread in pips. Pip is the minimal change in the internationally accepted format of currency pairs. For example, for EUR/USD pip is 0.0001, for USD/JPY - 0.01. For most major currencies for brokers, the relative spread ranges from 0.01% to 0.1% (Fig. 1.), but such a low spread-commission is misleading. Since it is impossible to execute an order on the international interbank market for a client with a small amount, the broker provides the client with a financial leverage (1:50-500). That is, to buy a minimum USD lot on the market (\$ 100,000), at the most popular leverage 1: 100 client needs to spend \$ 1,000. Recently, brokers commonly provide clients with opportunities to enter into deals for a part of the minimum market lot. All losses or profits due to the rate change of purchased USD (asset) are not shared in proportion to the invested funds, but are fully transferred to the client.

The principle of transferring the result entirely to the client is the reason for the high volatility of client accounts. With the growth of USD only by 1%, client funds invested in 1 lot increase by 100%. This is an attractive situation, but it works both ways. That is, for the loss of all funds fluctuation of the rate by 1% is sufficient. Risk-driven investors can be attracted by such a system, but it is often overlooked that commission is paid exclusively by client funds. Accordingly, the range of relative spreads is also multiplied by the financial leverage, and for the standard leverage 1:100 it will be respectively 1% -10%. With the use of coin flip to open orders for EUR/USD, the client's funds will transfer to the broker on average after 100 transactions executed. It is worth

noting that the financial leverage leads not only to the increase of risk and relative spread, but also to the possibility of diversification.

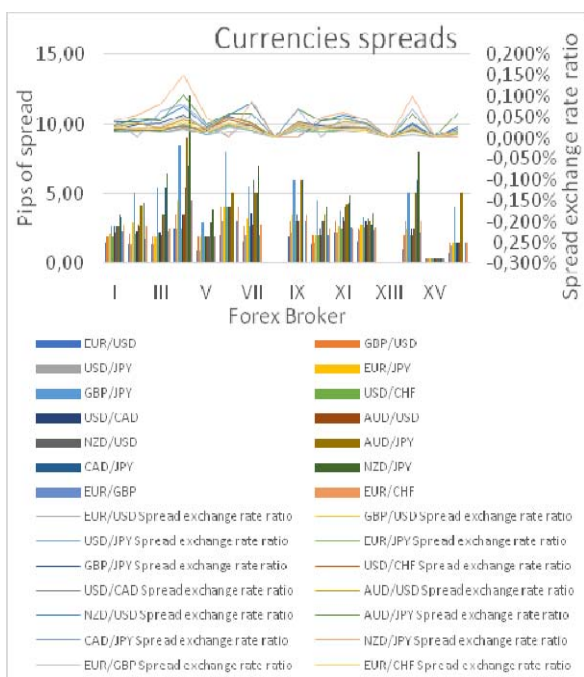


Fig. 1. Currencies spreads distribution among forex brokers [6]

Thus, owning \$ 1,000 and using 1:100 leverage we can buy only 1 lot, and with 1:500 leverage 5 diversified lots. Therefore, the choice of leverage is an investor's own deal. Also, if we keep an order beyond the date of purchase, the swap will occur. For each currency overnight interest rate is set by the central bank of the country. For example, if the annual rate for EUR is 5% and 4% for the USD, then the holding of the purchase transaction in EUR / USD pairs will result in the investor receiving in-come by volume $(5\% - 4\%) \cdot \text{Volume} / 365$ every night. With the sale order - to the same losses every night. Therefore, in a balanced trading system with the equal share of purchase and sale transactions such expenses are compensated. If broker adds his own fee to the swap, then it should be included in the relative spread. We are interested in the shift of break-even point of the trading system.

$$P_x \cdot \Delta Price \cdot Volume = (1 - P_x)(1 + s) \cdot \Delta Price \cdot Volume, \quad (4)$$

where P_x – the probability of correctly predicting a change in the price of an asset (>0.5), $\Delta Price$ – change in asset price, s – relative spread when buying an asset through a broker.

The equation (5) can be transformed:

$$\frac{P_x}{1 - P_x} = 1 + s \approx \frac{I(x)}{I(0)}, \quad (5)$$

where $I(x)$ – information available for investor, providing the required level of probability P_x .

As we see from (6), an investor can offset commission costs by adding additional information about the asset and market activity. Information should grow primarily in quality rather than volume. The established rule will allow us to measure how much really useful information the trading system provides. To do this, it's enough to get a share of

profitable transactions in the system and divide it by 0.5 ($P_{0.5}$).

III. MODERN DECISION SUPPORT METHODS ON FX MARKET

Traditionally, methods for analyzing a currency market and decision making are divided into fundamental and technical ones. Forex fundamental traders evaluate currencies, and their countries, like companies and use economic announcements to gain an idea of the currency's true value [7]. For making decisions in accordance with the methodology of fundamental analysis, macroeconomic indicators, that give an equivalent valuation of the currency, and the price dynamics of the currency themselves are important. The reaction of the currency market to changes in the macroeconomic environment shows up in the medium and long-term trends. The percents of swap become the source of profit of investor in such situations. In a competitive international market, low percentage rates are offset by fluctuations in the value of the currency itself. When macroeconomic indicators are changed not by objective laws, but subjective decisions of public authorities, the market takes into account the change instantaneously (within a few minutes). Therefore, it is not enough to rely solely on information of the fundamental analysis in modern conditions. The maximally near placing of client servers to the servers of exchanges becomes the determinative factor of fundamental methods, to have time to carry out agreements before "window of possibilities" closed (a few minutes taking subjective change of macroindexes into account by a market). For individual investors with small sums this is not relevant. Especially in speculative trading system, swap income in the long run will be mutually compensated.

Technical analysis is a trading tool employed to evaluate assets and attempt to forecast their future movement by analyzing statistics gathered from trading activity, such as price movement and volume [8]. The basis of technical analysis is the efficient-market hypothesis [9]. A current competitive market price effectively takes into account all available information about the asset. That is, observing fluctuations in prices can get a sense of and knowledge of all the necessary information regarding the financial instrument.

The main tools of technical analysis traditionally include historical price charts, indicators, financial ratios. Recently, automated systems using artificial intelligence replacing traditional technical indicators. Deutsche Börse AG have announced the use of a global index based on the selection of financial instruments to it by artificial intelligence methods [10]. Use of artificial neural networks does not guarantee success, but it becomes necessary for a successful trading system. These methods allow to process large amounts of information quickly and get the desirable information advantage, and in consequence reduce the risk.

We will focus solely on price indices and volume transaction characteristics. An information advantage means getting more qualitative data, available in all market history, about patterns in the past, that have already begun to operate in the present and will continue for some time in the future.

The general trend is the dominance of international banks, but at the same time in-cresing the role of brokers who use high-tech information systems investment (XTX Markets) [11]. All top companies can afford expensive systems using artificial intelligence, but their disadvantage is

the scale and long implementation of the latest methods. But even with the rapid ability to implement analytical systems, which use artificial intelligence, by individual investors, top10 traders will always be better informed about their own future actions, due to asymmetry. That is, individual investors, using the Pareto rule, should focus on information systems that can reach the probability of successful transactions 0.8. So provide information advantage over 80% of the market participants, which may indirectly influence the market.

IV. FORMING OF ANALYSIS AND TRADING SYSTEM WITH BIGDATA FLOWS

The first important step in the formation of price dynamics analysis system is a in-formation base. In addition to traditional historical data on price and volume of transactions, the database should include all possible technical indicators, derivative data and proportional, regression levels. A reference point is the information used by other participants in the currency market. The principle of quantitative advantage in information is provided by wider accumulation of all possible data streams. The principle of qualitative advantage in information consists in the effective selection of useful information from a vast set to the next stage of the analysis. Historical asset price data can be found on many online resources. But if the analysis is not static, and includes dynamic principle, we should use one of the online terminals. The most popular is MetaTrader4. Its advantage is the extensive use by brokers and their clients. Therefore, when we install it, we will have the opportunity to immediately open a demo-educational or valid cash account. The terminal includes the program language MQL4. The main problem is the closed structure and the impossibility of online data export to other software applications in a client free version. There is a simple solution to this limitation. The online community provides many useful solutions in the form of open source. One such example is the library mt4R.mqh by Bernd Kreuss (2010) [12], which covers the rights of free use under the terms of the GNU General Public License. By integrating this library into the MetaTrader4 terminal, we will be able to send all the necessary data in the R environment and return the result back to the Metatrader4 software environment. R is an open-source statistical package that is constantly expanding the tools of statistical, regression, neural network and other analyzes. The environment of R-scripts is convenient for forming the core of online analysis systems: information filtering, classification of information, model construction, forecasting and signal calculations.

In order to get possibility to exchange data between MetaTrader4 and R, we should add the following at the beginning of the expert advisor .mq4 program code:

```
#define RPATH "C:/Program Files/R/R-3.4.1/bin/i386/Rterm.exe --no-save"
#define RDEBUG 2
#include <mt4R.mqh>
```

In the above lines in order: connect the environment R, indicating the location on the computer of the installed folder with R; set the level of display of events in the register of the operating system; connection of program commands from the library mt4R. At the initial tuning of code, it is recommended that the RDEBUG parameter be equal to 2 and track all events, for example, through the DebugView application. When the automated system working stable

without any interruptions, the parameter can be reduced in the value to the minimum.

The main task of the program expert on the MetaTrader4 side will be to send initial online information and trade orders when receiving processed trade signals back from the R script. As part of the MT4 procedure onOpen () - executes each time when a new time period starts (minute, hour, day ...); we can generate a code:

```
Rm("rHigh", High[], bars, 1);
```

In the above line we used a programm command "Rm" from the library mt4R, forming a matrix in R of dimension bars * 1 with the name in " " - rHigh, rewriting data from a one-dimensional array High[] of current currency. Having formed such 5 blocks, we transfer all the basic information – prices on open, close, high, low and volume of trade – to the environment of the R-script. It is possible to calculate all known technical indicators in the environment of R, using specific libraries. But there are differences in the formulas for calculating many indicators, so it is recommended to use the same data that appear on the screens of traders around the world. That is, the calculated indicators in MT4 should also be transferred to R by inserting similar to the previous program blocks and using the indicators iMA (...), iMFI (...), iSAR (...), customized ZigZag iCustom (...), ZigZag, ...) and others. Overall moved an additional 171 technical indicator with averaging for 2-3 periods of 10 years of daily information. It is important to expand the information by all possible methods, especially by adding derivative indicators used by individual traders. Popular is the swing theory of trend definition derived by Gann (1936) [13]. We can calculate the swing in R or by running from R the macro in Excel. Having a swing identification of a short motion and a trend in general, at each moment of time allows us to calculate even the derivative indicators: the duration of the current short and long-term movement in days / minutes; speed of change in prices in the short and long-term. We get the same structure information from the ZigZag indicator. Some traders use Gann theory, others - ZigZag indicator. Our goal is to aggregate each indicators information in the database. After defining all types of wave-like movements (swings), we calculate all possible levels of the Fibonacci. As a result, received 8 basic prices, time, volume streams and 256 streams of indicators.

V. AUTOMATIC SYSTEM OF ANALYSIS AND TRADING

We will continue to process existing information flows in R, taking into account the repeatability of the script execution in each new time period: day, hour, minute. Since the swing calculation of swings lasts on average 3 minutes, then it is not recommended to use a system with a complete set of data for minute time intervals. Trading systems within the minute time interval are classified as high-frequency. For such systems it is recommended to include only the basic 8 variables in the dataflow. Based on the received data, we can execute a regression forecast for price changes and accordingly decide on the opening deal. We can also evaluate the price movement indicator (up = 1, down = -1) by a neural network classifier or logit regression. This brief analysis will require on average 15-20 seconds and can be successfully used in high-frequency trading. Going further, we will concentrate on the positional trading systems of hour and day time intervals, with the ability to analyze the full data set in order to obtain information advantage.

All 264 variables should be checked for stationarity by `adf` or `kps` tests (library `tseries`). To avoid in-process failures of the automated system, we recommend to use the `try ()` function. Verify the result of the function `try(kps (...))`. If the result is not numeric, then repeat the procedure as long as not obtain a numerical result. It is enough to use the `while {}` cycle with the maximum limit of the number of passes equal to 100. When identifying a non-stationary series, you must use the differences method to transform it into a stationary data. The next step is a factor selection to predict the direction of price movement. There are sufficient mechanisms for choosing the optimal set of factors. For example, function `regsubsets` (library `leaps`). Because the function algorithm made by browsing all possible options, the choice of factors from a large-scale database will last for years. Unacceptable option. There is also a function "step" (library `stats`), but have chosen function "omit" from other open-source statistical package - `Gretl`. A well-established library "Rgretl" and its function "grmod" integrate those statistical packages [14]. The following code will generate a list of optimal factors for the selected price stream:

```
try(grmod("ols y const X*
omit --auto=0.01", data = dataGr, output="$xlist"))
```

`dataGr` – pre-prepared matrix with data.

In addition, in order to predict future swings and trends in the price of short and long periods, it is necessary to choose the optimal set of factors as well:

```
try(grmod("logit y const X*
omit --auto=0.01", data = dataGr, output="$xlist"))
```

The choice of factors is carried out, but in the classification model - `logit`. This solves the fundamental problem of the uncertainty of the current price movement. Now, from the time point of the beginning of the uncertainty of the movement direction, we predict the probability of up / down swing in `R` on the received sets of factors. We have received a completely filled database in any situation, at each moment. In the internal environment `R`, we forecast price levels using dynamic `ARIMA`-factors-lagged regressions, neural networks, or other models. We compare the received future prices with the current and get not only information about a possible direction of movement, but also about the horizon and the limits of this movement. If we say in traders' words, we got the type of buy/ sell transaction, the entry point (at which price will open the deal) and the exit point (at which price the transaction will be closed). Results are transmitted from `R` back to `MT4`.

`NormalizeDouble(Rgd("Hp"),5)`

Function `Rgd ()` sends the numeric value of the variable `Hp` (High price forecasted in `R`) from `R` to `MT4`. The average duration of a single run the entire program takes 8-9 minutes on an ordinary computer. With the integration of all safety devices against failures of the automated work, the average duration of work is 25-40 minutes.

VI. CONCLUSIONS

The international currency market is developing towards individual investors. Conditions for entry into the market, even with insignificant capital, have been created. At the same time, investors should pay close attention to trading conditions provided by brokers.

The relationship between the success of the trading system, investors' decisions and information flows is shown. Available fundamental and technical indicators do not meet the information needs of the full extent of investors with little capital.

Modern systems should use a broad stream of online data. Within an integrated data stream, a classification, cluster, and factor analysis should be performed in order to obtain new useful information. Thus, the information advantage of an investor with a small capital can be formed.

The current automated system was activated during 01.03.16-10.11.16 on actual market `EUR / USD` and as a result made 80 daily deals. The probability of profitability of the transaction in the system was 51.25%, the total profit was \$ 219.51 at the initial balance of \$ 96.4. Profitability is 127.7%, which is far more than the average Ukrainian deposit rate – 14%. The results indicate the achievement of the information advantage by the system, during working in real conditions and taking into account all commissions.

Strengthening the system is possible at choosing the optimal set of prediction factors, applying evolutionary methods to determine the optimal parameters of orders, classification market conditions.

REFERENCES

- [1] C. Geisst, *Encyclopedia of American Business History*. Infobase Publishing, 2009.
- [2] History of the Foreign Exchange (Forex) Market. (<http://www.nasdaq.com/forex/education/history-of-retail-forex-market.aspx> Accessed 27 Jan 2018)
- [3] Triennial Central Bank Survey of foreign exchange and OTC derivatives markets in 2016. (<https://www.bis.org/publ/rpfx16.htm>).
- [4] N. Popper, "Knight Capital Says Trading Glitch Cost It \$440 Million," *The New York Times*, August 2, 2012.
- [5] The cost of trading forex. Tradimo Interactive ApS. (<https://learn.tradimo.com/dont-go-broke-protect-your-capital/the-cost-of-trading-forex>)
- [6] Forex brokers spread comparison. (<https://www.100forexbrokers.com/compare-forex-brokers-spreads>)
- [7] C. Murphy, "Forex Tutorial: Fundamental Analysis & Fundamentals Trading Strategies," *INVESTOPEDIA*. (<https://www.investopedia.com/university/forexmarket/forex6.asp>)
- [8] A. MacEachern, "What is the best method of analysis for forex trading," *INVESTOPEDIA*. (<https://www.investopedia.com/ask/answers/forex/best-method-of-forex-analysis.asp>)
- [9] B. Malkiel, E. Fama, "Efficient capital markets: a review of theory and empirical work," *The Journal of Finance*, vol. 25(2). pp. 383–417, 1970. doi:10.1111/j.1540-6261.1970.tb00518.x
- [10] R. Evans, K. Leinz, It's man vs. machine in a battle to pick a better stock index. *Bloomberg*. (<https://www.bloomberg.com/news/articles/2018-01-22/the-latest-pence-calls-report-on-trump-affair-baseless>)
- [11] J. McGeever, "Citi tops Euromoney global FX poll again, but big banks lose grip," *Reuters*, 2016. (<https://www.reuters.com/article/global-forex-euromoney/citi-tops-euromoney-global-fx-poll-again-but-big-banks-lose-grip-idUSL5N18M29O>)
- [12] B. Kreuss, `R` for `MetaTrader`. 7bit. 2010 (<https://sites.google.com/site/prof7bit/r-for-metatrader-4>)
- [13] J. Kuepper, "Introduction to Swing Charting," *INVESTOPEDIA*. 2017. (<https://www.investopedia.com/articles/technical/04/080404.asp>)
- [14] O. Komashko, (2017): Package 'Rgretl'. (<https://cran.r-project.org/web/packages/Rgretl/Rgretl.pdf>)

Development of a New Algorithm Based on Simulation – Optimization Algorithms for Big Data Mining to Improve Prediction of Future Electricity Prices in the Iranian Electricity Market

Mesbaholdin Salami
*Department of Industrial Engineering
Central Tehran Branch, Islamic Azad
University
Tehran, Iran*

Farzad Movahedi Sobhani
*Department of Industrial Engineering
Science and Research Branch
Islamic Azad University
Tehran, Iran
Fmovahedi@iau.ac.ir*

Mohammad Sadegh Ghazizadeh
*Department of Electrical Engineering
Abbaspour School of Engineering
Shahid Beheshti University
Tehran, Iran*

Abstract—Abstract The structure of the electricity market in the world is rapidly changing. Due to the current structure of the Iranian electricity market and its changes, a high volume of daily data is generated and stored at a very high rate, leading to big databases. One of the most important problems with big data is their analysis. Despite numerous problems with big data, it provides unparalleled opportunities for the more accurate analysis and prediction of important variables of the electricity market. Traditional data mining techniques do not provide quality and time-efficient big data analysis. New smart and metaheuristic methods are needed to be used in the electricity market to compensate for weaknesses of traditional data mining techniques in big data analysis. In this article, a smart method called K-Means Simulation Optimization Algorithm (KMSOA) was presented. The performance of this algorithm on the Iranian electricity market databases showed positive results in data clustering including a more accurate prediction of the major variables of this market, e.g. the electricity supply and demand and faster data analysis. After estimating these two variables, they were calculated along with other parameters with the aim of calculating a fair electricity price. The predicted price should be offered by the electricity market manager for each period from a multi-objective model defined to minimize costs for actors in this market. The speed and accuracy of predicting fair prices are very important in this market, bringing huge savings as mentioned in the results of this article.

Keywords—*electricity market, Big data, Fair electricity price, Simulation-Optimization, KMSOA, Multi-Objective model*

I. INTRODUCTION

The growth rate of data generation has increased dramatically in recent years such that more than 2.5 exabytes (equivalent to 2,500 million gigabytes) of data were generated every day in 2016. This volume of data and the increasing rate of data generation have shaped the concept of big data [1, 2]. The rapid growth of data has caused a storm leading to great challenges in data acquisition, storage, security and analysis [3]. One of the most important challenges is data analysis. Institutions that provide platforms for big data analysis can benefit from the value of such data in decision-makings involved in organizational processes [4]. The energy sector, specifically the electricity market, is one of the areas in which high data rates are

constantly generated. The electricity market is experiencing a high daily data generation rate. This large volume of data generation affects decision making in the electricity market management such as setting a fair price for electricity on a daily basis. The fair electricity price is proposed by the electricity market manager and is used as the basis for electricity trading between buyers and sellers at any given time [6]. Therefore, solving associated problems with big data mining in the electricity market is an area for attractive pioneering research. The large volume of big data including semi-structured or non-structured data [3] has seriously limited data analysis using traditional data mining methods, making time-efficient data mining practically impossible using such methods [3]. Therefore, new methods should be developed to guarantee both quality and time-efficiency. Methods presented in recent years have solved these problems in various aspects [4]. Some of these methods such as machine learning adjust and explore algorithms that enable computers and systems to learn [7]. Reinforcement learning is another type of machine learning inspired by behaviorist psychology and aiming at maximizing rewards [8]. Simulation-optimization methods have received much attention in recent years. The aim of this study is to develop a new algorithm called K-Means Simulation Optimization Algorithm (KMSOA) for big data mining based on simulation- optimization methods. The proposed algorithm will be used to solve problems associated with big electricity market data analysis. It can be used in various fields as a sustainable method for big data analysis. In this method, big data features including volume, rate, value, accuracy and validity are used to design and adjust the algorithm parameters. The designed algorithm was implemented on the big data of the Iranian electricity market. The data volume and generation rate in the Iranian electricity market are high due to the daily interactions between purchasers and sellers in the Iranian electricity market to buy and sell electricity. Therefore, high-precision data mining in this area will bring huge savings as mentioned in the results of this article for the actors in this field including producers and consumers, i.e. power plants and electricity distribution companies, respectively. The accuracy and runtime of this algorithm were compared with other methods to predict the future price for electricity. Comparison was also made with one of the most important newly-introduced reinforcement learning

methods for big data mining. This demonstrated the advantages of and need for data analysis using the new algorithm. Finally, the predicted variables derived from data mining were considered as input parameters to the multi-objective model to determine the fair price for electricity by the electricity market manager to be used by buyers and sellers in future electricity exchanges.

II. RESEARCH BACKGROUND

The aim of this study is to propose an algorithm for big data mining in the electricity market based on simulation-optimization methods. Thus, it is necessary to review prediction of future electricity prices, simulation – optimization algorithms and big data mining, separately.

A. Electricity market

There are numerous studies on the electricity market. A number of articles with a similar topic to this article concern prediction of future electricity prices. Methods have been proposed in these articles to determine the electricity price in the market with the least error and taking the constraints resulting from the actual electricity exchanges into account. A number of articles address the use of time series methods in the prediction of electricity prices. For example, Cuaresma et al. (2004) presented a time series multivariate model for predicting electricity prices. Other articles address the use of neural network methods to predict electricity prices. Yamashita et al. (2004) presented a neural network model of parameters affecting prediction of electricity demand and prices. Pao et al. (2007) investigated the prediction of future electricity prices in European countries using neural network methods. Some articles offer a variety of single or multi-objective models with constraints designed to predict future electricity prices. Borovkova et al. (2017) modeled electricity prices considering random time changes. Islyayev et al. (2015) presented a model for predicting future electricity prices in contracts taking into account random parameters. Florian et al. (2015) proposed a model for EPEX prediction taking into account renewable energies in the model. Another category of articles concern combined modeling and statistical methods for predicting electricity prices. For example, Borovkova et al. (2017) modeled electricity prices with random variations of time. A number of papers address the simultaneous use of the time series and data mining techniques. For example, Lu et al. (2005) investigated prediction of electricity prices using data mining techniques. None of these papers mentioned the trend of structural changes in the electricity market and generation of large volumes of data, analysis of big data and prediction of future electricity prices in these conditions.

B. Big data mining and the role of machine learning algorithms

Various methods, tools and even software pieces have been used to analyze big data. A type of algorithms known as "the big data analysis methods" includes the following applications: the DBDC method in Parallel clustering ((Januzaj et al. (2004)), PKM for map-reduction-based k-means clustering (ZhaoW et al. (2009)), CloudVista in Cloud computing for clustering (Xu H et al. (2012)), DHTRIE in the Applied frequentpattern algorithm to cloud platform (Yang L et al. (2011)), MSFCUDA in the GPU for clustering (Cui X et al. (2013)), BDCAC in Fidelman's Ant on grid

computing environment forclustering (Ku-Mahamud et al. (2013)), Corest in tree construction for generating the coresets in parallel for clustering (Feldman et al. (2013)), SOM-MBP in Neural network with CGPclassification (Hasan S et al. (2013)), CoS in concerning Rebalancing of Parallel Computing for Classification (Tekin C et al. (2013)), QuantumSVM in Quantum Computing for Classification (Rebentrost and Both Eran (2014)), DPC in Map-Reduction Model for frequent pattern mining (Lin MY et al. (2012)), MFPSAM in Concerned The specific Interest Constraints and Applied Map-Reduction Model (Leung CS et al. (2014)) and SVMGA in using GA to reduce the number of dimensions (Lee J et al. (2014)). Simulation-optimization methods have not been directly used in any of the machine learning algorithms for big data mining.

C. Simulation-optimization algorithms

Machine learning is one of the most important parts of artificial intelligence (AI). Machine learning consists of several parts, one of which is reinforcement learning. Simulation-optimization algorithms, which have been used for numerous applications in recent years, are one of the topics in reinforcement learning. Most articles delve into the application of these algorithms in maintenance as well as issues related to reliability. For instance, Alrabghi et al. (2017) conducted a case study on the maintenance system. Shahrabi et al. (2017) used this method in the job shop dynamics planning. Yegul et al. (2017) investigated configuration enhancement of complex production lines by this method. This type of methods has not been used directly or indirectly in the electricity market. In addition, this method has not been used as part of a big data mining algorithm. Therefore, the advantages of this type of algorithms are exploited in this article to predict electricity prices.

III. ELECTRICITY MARKET STRUCTURE

Unlike the past, electricity is currently exchanged as a commodity similar to other commodities in the supply and demand system. This product needs to be exchanged with an exact planning between buyers and sellers to not deal with such problems as power outage or excess electricity generation in the electricity grid. A small disruption in the exchange of this vital commodity will cause an irreparable damage to electricity consumers including large and small industries and electricity generators, namely power plants. In these conditions, returning to normal conditions will impose enormous costs on the national economy [18]. Unlike the old structure in which the production manager was also responsible for power transmission and distribution, the systems operate independently within the new structure of the Iranian electricity industry. In this market, electricity is exchanged between buyers and sellers on a daily basis. Therefore, buyers and sellers should be informed of the production capacity and the consumption load, respectively. The electricity market manager should be informed of both of these data to predict the fair electricity price. Naturally, these predictions do not always match with real conditions, as price prediction is always associated with some error. Therefore, predicting the buyer demand and the production rate of power plants with the least error and thereby the exact prediction of a fair electricity price may bring considerable savings for electricity buyers and sellers [19]. Major actors of the electricity industry include the electricity market

management, the producers or power plants, electricity transmission companies, distributors and ultimately consumers or electricity subscribers. The actors are briefly described below.

A. Electricity market management (Independent System Operator)

In the new structure of the electricity industry, an independent entity called the independent system operator (ISO) has been established for controlling exchanges between buyers and sellers. To have a competitive market, the ownership of the grid should be separated from the controlling entity. Therefore, an entity independent of producers, distributors, and owners of the lines should determine the transmission price, maintain the system safety, coordinate maintenance planning and monitor the measurement of performance data of each electricity market actor. This entity should provide free and non-discriminatory access for all system users [12]. Based on stored data, ISO can place some or all of the system production units in or remove them from the grid. To maintain the system safety (due to transmission constraints, the supply-demand balance, and desired frequency maintenance), ISO can cut off some loads. The ISOs of these units encourage them to make effective use of existing facilities and invest in the necessary resources to avoid possible future deviations. In the restructured electricity market, the goal of these companies is to maximize their profits. To this end, these firms may be involved in any kind of market (energy market and ancillary services) and make any transaction (cash, risk coverage, speculation). The firm is responsible for considering possible risks. Therefore, producers should plan in such a way to maximize their profits, i.e. the difference between incomes and expenses. Sometimes the costs of power plants, such as unload costs (the cost needed to keep the unit operational) and startup costs are quite large so that producers prefer to be in constant operational mode. On the other hand, these producers need to estimate their production capacity by instantaneous evaluation to adjust their production capacity. Consistent operation of power plants and imposition of enormous and never-ending costs require correct estimation of the buyer demand in the electricity market. Power production units may have to continue operation at non-peak hours resulting in loss. Manufacturing companies occasionally have to buy energy from other firms instead of power generation to fulfill their obligations and sales commitments.

B. Electricity transmission companies

The mission of transmission companies is transmission of high-voltage power in the transmission grid from manufacturers to customers and distributors. The facilities of the transmission company are used under regional ISO surveillance. The equipment for electricity transmission including cables and conversion posts is owned by these companies. The owners of the transmission grid are typically the same as those of vertically integrated transmission grids [18]. In all electricity markets in the world, the transmission sector is operated under the exclusive monopoly of these companies. Basically, competition in the transmission grid is not practical and cost-effective. Therefore, to ensure fair conditions in the electricity market, transmission companies operate under strict supervision and regulations.

C. Electricity distribution companies

Each distributor distributes electricity among its customers in a specific region through its facilities. These companies operate under the existing regulations. The grids owned and operated by distributors transmit and distribute high-voltage electricity. They are responsible for maintaining the voltage and providing ancillary services in their area. At the same time, distribution companies may be retailers. In this case, specific provisions are needed to prevent the creation of a monopoly in the retail market as the ownership of the distribution grid is naturally monopolistic and creation of multiple distribution grids is unrealistic; although there are some examples of multiple distribution grids. In this article, electricity distribution companies are considered as purchasers in the electricity market.

D. Electricity subscribers

Subscribers are the final consumers of electricity. If subscribers are present in a retail market, consumers are connected to the distribution network. If they are present in the wholesale market, they are connected to the transmission grid. In the reformed electricity system, customers will not necessarily be forced to buy electricity from the regional electricity company, but they can buy electricity from the best service provider in terms of safety, reliability and price by participating in the market. For example, a customer may buy electricity from suppliers with lower prices at non-peak hours. They also buy electricity from a retailer which provides electricity from producers observing environmental criteria. Currently, consumers can be classified into two main categories of micro and macro consumers. Micro consumers buy electricity at a fixed rate or, ultimately, at a limited number of fixed rates. In the current situation of the electricity market, they do not buy electricity from power plants directly. However, after establishment of the nationwide smart power grid in Iran (until 2030, according to the schedule), micro consumers including domestic consumers will play a role in the electricity market as a member [10]. Due to fixed electricity rates, they are not sensitive to fluctuations in electricity prices. Based on empirical evidence, although consumption decreases in the short-run as electricity prices rise, the impact is insignificant. In other words, the price elasticity of power demand is low. However, the macro consumers of the power industry including large industries are the main buyers of the power industry. They are sensitive to fluctuations in electricity prices so that buying electricity at a lower spot price is desirable for them. But macro consumers and retailers (as intermediaries of electricity wholesale and retail markets) have to purchase electricity at a variable rate with high sensitivity (for different hours and days) and sell to micro consumers at a fixed rate. For this reason, the major buyers lose when electricity price is high and gain profits when the cost of energy is lower. In this case, the major buyers attempt to estimate their customers' consumption carefully before engaging in spot markets to conclude the purchase of electricity from the wholesale market. Competitive market in the demand side may increase productivity with the initiatives of these major buyers leading to a more competitive market. In the future structure of the electricity market, by establishing a communication structure between buyers and sellers, all consumers, even micro consumers and all manufacturers including private and public producers will be directly connected to the electricity market management.

In this way, the ISO will exchange data between buyers and sellers as an intermediate [11].

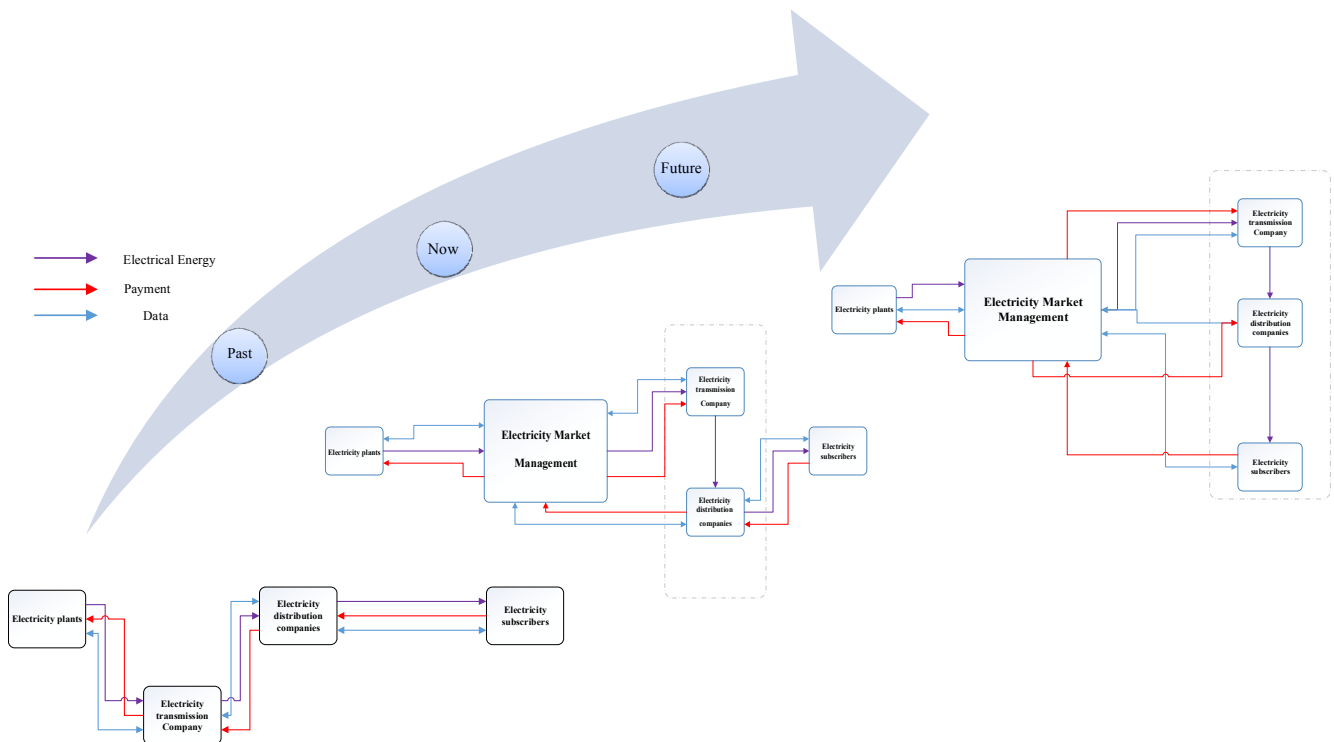


Fig. 1. Structural changes in the Iranian electricity market

Therefore, in the new structure, the electricity market management (Fig. 1) is responsible for exchanging information between buyers and sellers as a new member and an independent entity. This independent entity prepares the ground for data recall from actors as well as effective data storage and processing in this market. It is expected that in the new structure, the production and use of data will play a more significant role in the electricity market. The future structure of the electricity industry maintains the current structure while creating a data transmission network infrastructure for the presence of retailers, even domestic electricity subscribers in the electricity market. In this way, the retailers may contribute to electricity purchase from power plants, electricity transmission by transmission companies and power distribution by distribution companies. With the implementation of the smart grids, data is generated and transmitted very quickly and used as criteria for making decisions in the smart power grid [9]. In this case, the market will be quite competitive and data will assume greater significance than ever [27]. Therefore, fundamental solutions must be proposed for storing and analyzing big data and future changes in the electricity market. Algorithms should be developed to guarantee the future of data analysis in this area [17].

IV. 4V'S ELECTRICITY MARKET DATA

Electricity market data has been generated and stored since its establishment in Iran for buying and selling electricity. This data is generated based on the activities of sellers and buyers in this area. The electricity market management as an intermediary receives data from buyers and sellers to predict and approve the fair electricity price for the coming day after analysis. More than 300 parameters are involved in this area, representing the performance of buyers

and sellers. In this study, 38 parameters of about 300 electricity market parameters have been investigated. The total volume of data generated since the establishment of the electricity market by 2015 is 10,158 gigabytes. It is expected to reach 21,116 gigabytes by 2022 (regardless of the smart grid implementation). The velocity of data in Iran's electricity market is very high in a way that the main data related to various parameters are stored at minute intervals and the rest are stored and analyzed on a daily basis [23]. Given the changes in the electricity market to create a smart grid and to prepare the ground for the presence of retailers and sellers in this market, the data generation rate is much higher than in the past so that the data will be recorded instantly [14]. With the implementation of the national smart grid, it is expected that 2.5 terabytes of data be produced per day leading to an incredible increase in the number of data sources in such a way that any house or any building in the city will become a source of data for the electricity market [14]. Therefore, the future of the electricity market is a future with a large volume of data in such a way that the volume of daily data generation may be as large as the data generated in a non-smart electricity market during a month. Therefore, a suitable structure should be prepared for data storage and analysis for massive changes in data generation in the electricity market in the near future (Fig. 2). Despite the current non-smart structure of the Iranian electricity market data, it is indicative of the existence and production of a large volume of data on its own (Fig. 3).

The veracity of the data generation by different sources of data was validated by comparing the actual and estimated past average data. The results indicate that the generated data is uncertain. The data uncertainty is a feature of big data [3]. The data used in the next step for data mining, i.e. electricity market data shows a degree of variety. This variety includes

nominal, sequential, text, numeric, date and time data. Due to data variety, an algorithm is required for big data mining with a high data variety. This data variation must be converted to sequential numerical data before analysis by the methods outlined in Section 5.2.

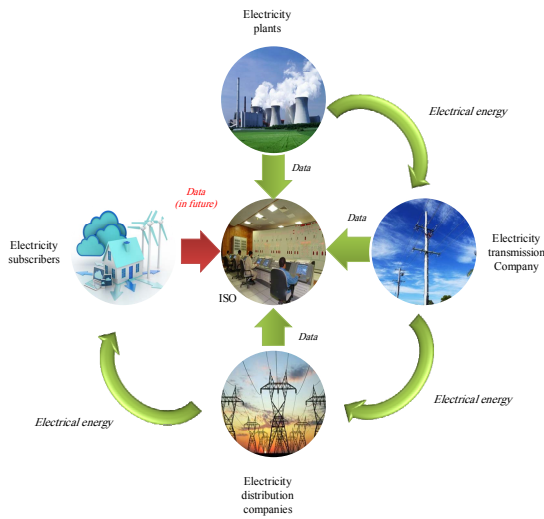
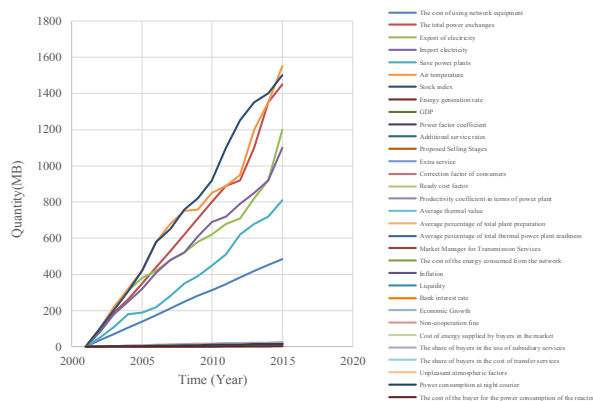
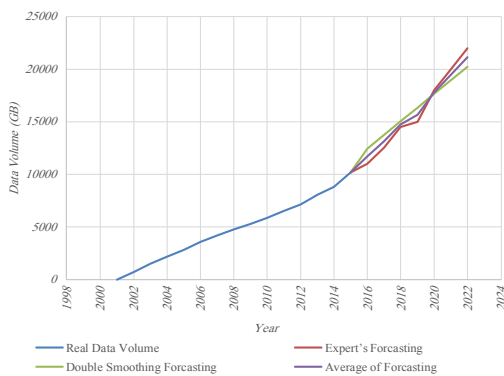


Fig. 2. The actors in the Iranian electricity market

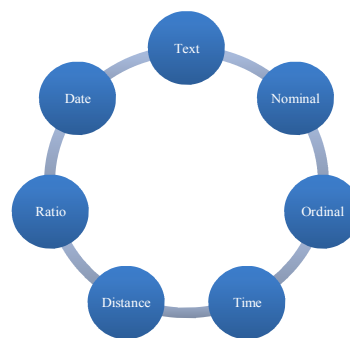
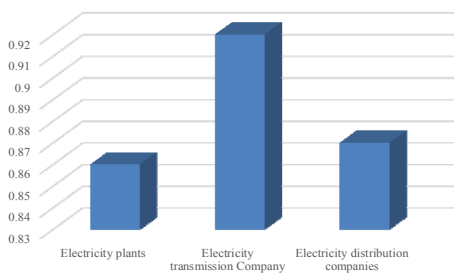
Demand and power generated by power plants are two very important parameters that should be calculated based on existing data by clustering big data. The predicted values for these two parameters should form the basis for decisions. These two parameters must be predicted with high accuracy to predict the fair electricity price exchange between buyers and sellers in the electricity market. On the other hand, it is necessary to make predictions in the electricity industry instantaneously so that buyers and sellers can compete in this market. Thus, data mining is required at a significant speed which requires methods for rapid analysis of big data [3]. The use of traditional data mining methods is obviously inadequate, as the time and quality of data mining and in particular clustering will face serious challenges. Therefore, it is necessary to provide a high-quality and time-efficient algorithm for data mining in this large volume of data. The results should be available to decision makers in this area or all areas dealing with big data. A small percentage of improvements in precision and forecast time will result in incredible savings for both sellers and buyers [20].



Volume

4 V'S

Velocity



Veracity

Variety

Fig. 3. 4V'S in the Iranian electricity market

V. DEVELOPMENT OF A SIMULATION-OPTIMIZATION ALGORITHM FOR BIG DATA CLUSTERING

As discussed in the literature review (Section 2), time series combined with concepts such as neural networks have been recently used to predict electricity prices in the

electricity market [26]. As mentioned in a number of other related articles, it is not simply possible to rely on past data for predicting the electricity price [30], because in a situation where consumption or production is imbalanced for any reason, prediction by the time series cannot provide an accurate estimate [29]. In other words, when the distribution

network deals with fluctuations in supply or demand, the prices deviate from the predictable trend. In this case, prices need to be predicted by other techniques, such as data mining methods [21]. It should be noted that the electricity market management does not calculate the fair electricity price based on past data to offer electricity prices for the next period; instead it should consider the profit and loss of buyers and sellers with a fair approach. Therefore, modeling is needed to minimize the cost of actors in this area. Figure 4 shows the steps for forecasting supply and demand variables.

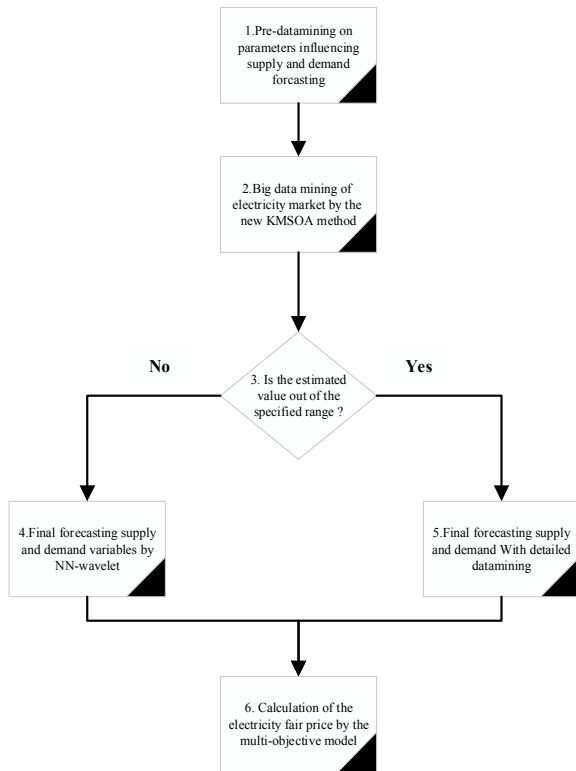


Fig. 4. The flowchart for predicting supply and demand in the electricity market

As shown in Figure 4, the necessary measures to prepare data for prediction are initially employed. This stage is called pre-data mining. Then, the new algorithm is used for predicting the supply of all power plants as power suppliers and demand for all distribution companies as consumers, individually. If the predicted value is greater or lower than $\mu \pm \alpha \sigma$, where α is considered an adjustable parameter in the algorithm, μ is the mean, and σ is the standard deviation of the variable, future predictions need to be made by data mining. Otherwise, predictions are made based on the NN-wavelet method according to Zhang et al. (2001) [22].

A. Big data mining with the new KMSOA method

Machine learning methods are among the main methods of big data analysis [15]. In this category, all methods are based on the application of the initiative strategies to solve problems. One of the important issues in machine learning is reinforcement learning [24]. Reinforcement learning is a branch of machine learning inspired from behaviorism. This approach focuses on behaviors which a machine must do to maximize rewards [16]. Articles on the use of machine learning to solve big data problems were reviewed in Section 2. Simulation-optimization methods are among the methods

used in reinforcement learning to solve big data problems. The algorithm proposed for this purpose is shown in Fig. 5. Using the new method presented in this paper, data is analyzed with the Simulation-Optimization approach. The aim of optimization is to obtain the best cluster with the least error using the k-means clustering algorithm and random data from the entire data set. Each cluster set is considered as a solution and evaluated. The cluster set is generated from a large volume of past data. Then, simulation is carried out by the Monte Carlo method. The value of E_f as the evaluation criterion (Section 5.2.3) is calculated and is used as the decision making criterion during simulation. The use of k-means algorithm with kd time complexities is a NP-hard problem given the complexity of the problem considering the constancy of dimensions, where k is the number of parameters, and d is the number of records, assuming both of these values are growing rapidly in the process described in Section 4 [15]. So it is necessary to use meta-heuristic methods combined with simulations. The flowchart for the simulation- optimization algorithm, called KMSOA is shown in Fig. 6. Table 1 lists the parameters considered from the electricity market databases for data mining to estimate demand and supply. These data include structured, semi-structured, and unstructured data:

TABLE I. PARAMETERS CONSIDERED FROM THE IRANIAN ELECTRICITY MARKET DATABASES FOR PREDICTION OF SUPPLY AND DEMAND

Row	Data	Row	Data
1	The cost of network equipment use	20	Power consumption at night peak
2	Cost of energy consumption of the grid	21	Total power exchange
3	Cost of overseas exchanges	22	Non-cooperation fines
4	Cost of energy supplied by buyers in the market	23	Non-cooperation amount
5	The share of buyers in using ancillary services (IRR)	24	Liquidity
6	The cost of purchasers of reactive power consumption (IRR)	25	Inflation
7	Purchasers' share of transmission service costs	26	Air temperature
8	Cost of overseas exchanges	27	Market management costs for transmission services (IRR)
9	Power plants storage	28	Encouragement cost for consumption test
10	Energy generation rate	29	Average percentage readiness of all power plants
11	Power plant productivity	30	Average percentage readiness of all thermal power plants
12	Cost of ancillary services	31	Unfavorable social factors
13	Sales offer stages	32	Unfavorable atmospheric factors
14	Corrective coefficients of consumers	33	Readiness return cost
15	Productivity coefficient according to the power plant type	34	Transaction support
16	Average thermal value	35	Stock index
17	Readiness cost factor	36	Interest rate
18	Electricity imports	37	Economic growth
19	Electricity exports	38	GDP
20	Ancillary services (data analysis as characters (Section 5.1.1.2))		

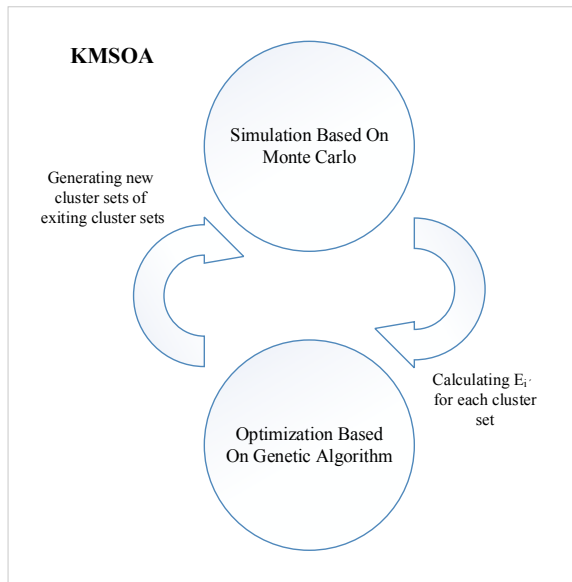


Fig. 5. The flowchart of the new algorithm

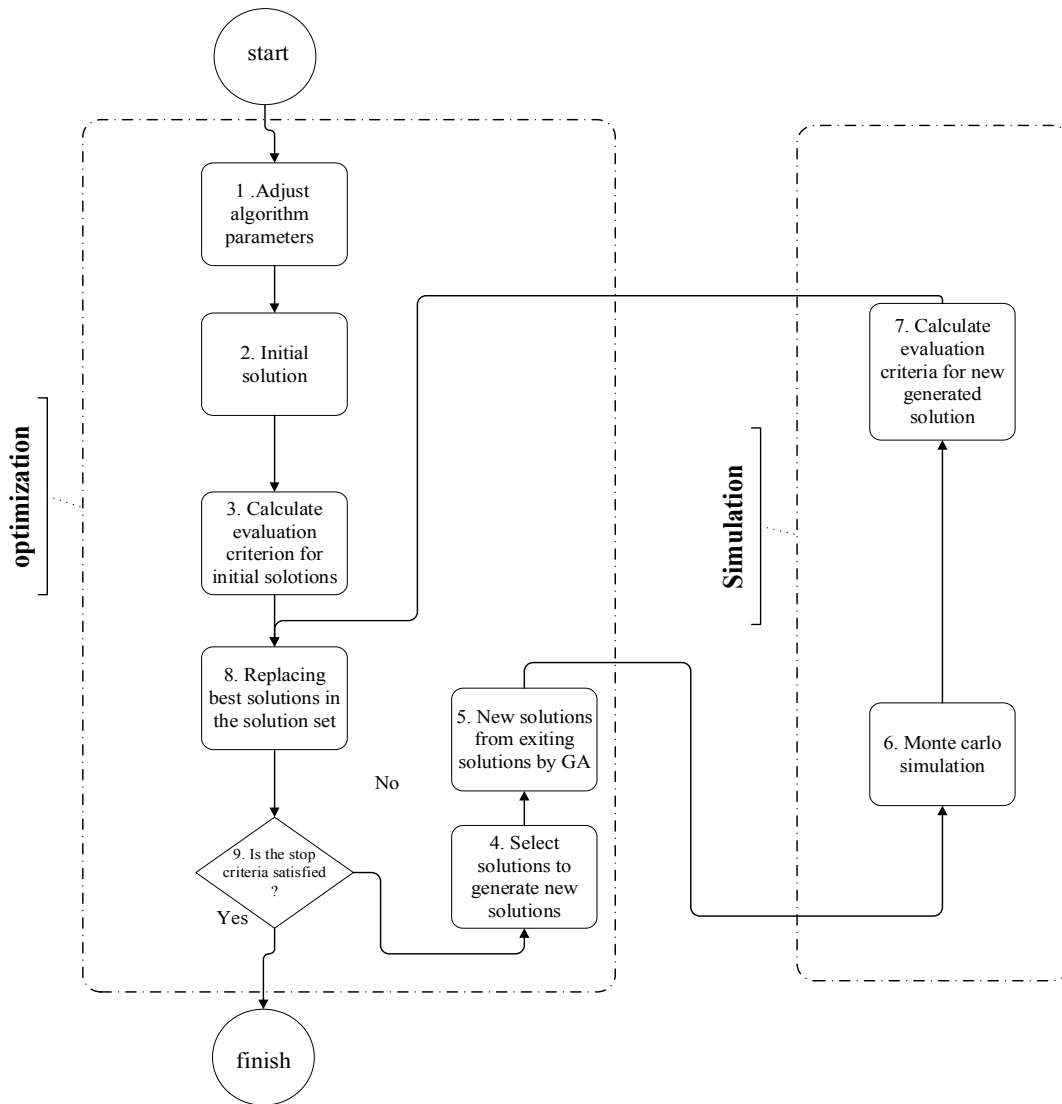


Fig. 6. The KMSOA flowchart

This algorithm is implemented by Matlab2016® for the Electricity Distribution Company No. 3 (in this paper, the

applicant for electricity) on February 8, 2017 (Fig. 7). The procedure is described below:

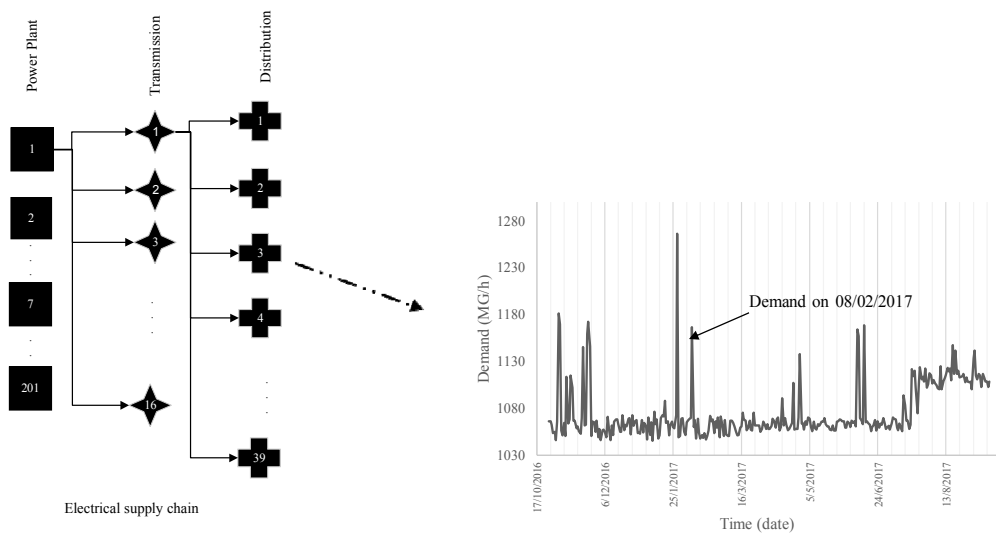
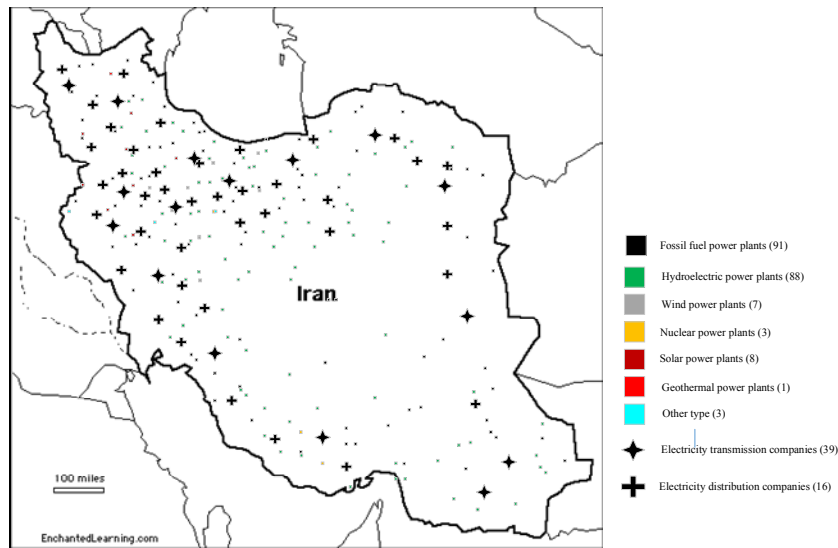


Fig. 7. Dispersion of the Iran's electricity supply network and electricity demand in the Electricity Distribution Company No. 3

B. Adjusting the parameters of the algorithm

The parameters of the new algorithm should be adjusted after each run. There are two common methods to adjust the parameters of the algorithm in Table 2 including Taguchi methods and responsive surface method (RSM). In this

paper, the RSM is used to adjust the problem parameters [35]. This factor includes the following parameters and the response variable includes the algorithm runtime, the evaluation criterion, or E_r (Section 5.2.3) and the error rate of the electricity price prediction calculated for running the algorithm on 02/08/2012:

TABLE II. PARAMETERS OF THE KMSOA ALGORITHM

Parameter	Description	Range	Best value by RSM	Parameter	Description	Range	Best value by RSM
P_0	Number of parameters selected for data mining	[20,102]	73	$\varphi^T(\cdot)$	Significance of the third objective function	[0,100]	33
l_0	Number of clusters in each cluster set	[10,50]	15	n_0	Stop criterion in terms of number of iterations	[1000,1500]	1239
r_0	Size of each random sample	[100,150]	122	F	Mutation parameter in GA	[0.5,0.8]	0.7
q_0	Number of random samples	[20,60]	25	D_0	Number of simulation runs	[30,80]	62
U_0	Stop criterion in terms of evaluation criterion	[65,75]	70	M_0	Number of random points in each simulation run	[100,150]	120
l_1	Number of clusters in each cluster subset	[5,15]	12	C_0	Capacity of solution set	[10,15]	12
φ	Significance of the first objective function	[0,100]	31	α	A coefficient of standard deviation	[2,3]	2.4
φ'	Significance of the second objective function	[0,100]	35	β	Significance of two evaluation indicators	[0,100]	65

C. Generation of initial solutions

In this section, a random sample is selected based on simple random sampling method. In this method, samples are selected based on the storage time and have an equal probability to be selected. Thus, probability rules determine which units or groups of the mother population will be selected. Mother population is the same set of chromosomes recorded in time including 38 parameters and the supply or demand variable. The number of random samples is q_0 , the

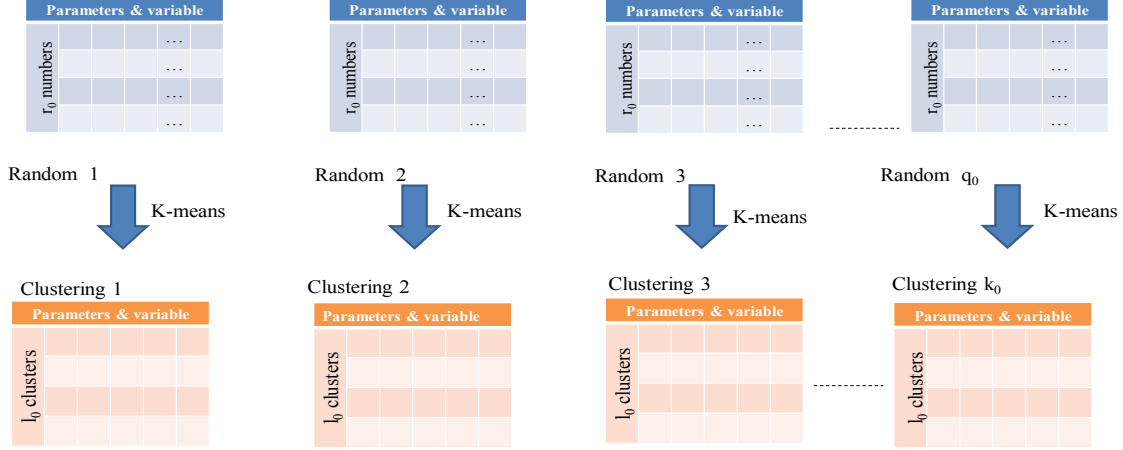


Fig. 8. Generation of the cluster set from random points

D. Evaluation of initial solutions

In this section, the set of initial solutions is evaluated by two important criteria. Each of the solutions with the best values of the combination of these two criteria will be selected. The best solutions generated in each run are replaced with the worst solutions. The first indicator, $RSMD_i(\hat{\theta})$ is used to measure the accuracy of the algorithm

number of selected parameters for data mining is P_0 which is a random number between 20 and 102. The chance of parameters with the highest correlation with the supply or demand variable based on the results of past forecasts is higher. Then, we proceed with clustering using k-means algorithm. The size of each random sample is r_0 , and the number of clusters of each cluster set is l_0 , except for the adjustable parameters of the algorithm at this stage. The production of the initial solution is shown in Fig. 8.

through testing the cluster set obtained with a set of random data and actual demand and supply values. The other indicator, $d_i(p, q)$, represents the distance between the points forming the center and the center of the cluster. E_i is the weighted average of these two indices which is considered as an indicator for evaluating the produced cluster set. Table 3 describes the variables of this evaluation criterion:

TABLE III. INDICES, SETS AND PARAMETERS FOR ERROR CALCULATION

$\hat{\theta}_{i'j'k'}$	The estimated value of the random sample j' selected for estimating the parameter k' in the cluster set i'	i'	Cluster set index
$RSMD_i(\hat{\theta})$	The average error for the cluster set I'	I'	The index for all cluster sets
$p_{i'l'}$	coordination of the random point l' for generating the cluster set I'	j'	The index for samples selected for error evaluation
$q_{i'l'}$	The nearest center of the cluster set i' with point l'	J'	The index for all samples selected for error evaluation
$d_i(p, q)$	Total Euclidean distance of points to centers of cluster sets i	l'	The index for points selected for clustering
α	Significance of $d_i(p, q)$	L'	The index for all points selected for clustering
β	Significance of $RSMD_i(\hat{\theta})$	k'	Parameters predicted in data mining
m	Number of random samples selected for calculating the index $RSMD_i(\hat{\theta})$	K'	All cluster sets
E_i	The evaluation criterion for the cluster set i'	$\theta_{i'j'k'}$	The actual value of the random sample j' selected for estimating the parameter k' in the cluster set i'

$$RSMD_i(\hat{\theta}) = \left(\sum_{j' \in J'} \sqrt{\sum_{k \in K'} (\hat{\theta}_{i'j'k'} - \theta_{i'j'k'})^2} \right) / m, \forall i' \in I' \quad (1)$$

$$d_i(p, q) = \sqrt{\sum_{l' \in L'} (p_{i'l'} - q_{i'l'})^2} \quad \forall i' \in I', \quad (2)$$

$$E_i = (1 - \beta) \cdot d_i(p, q) + \beta \cdot RSMD_i(\hat{\theta}) \quad \forall i' \in I'. \quad (3)$$

E. Choosing a solution for optimization

To select the best answer for improvement during algorithm runtime, a fixed number of three cluster sets should be selected and entered into the optimization algorithm. In this step, the roulette cycle with more chance is used to choose solution with the best evaluation criterion. Figure 11 shows the initial solution set. Each line in this figure represents a cluster and each curve shows a cluster set

or a solution in the algorithm (37 parameters of sequential numerical type and 1 parameter (ancillary services) which is

analyzed as 36 characters). The total number of parameters is 73 parameters.

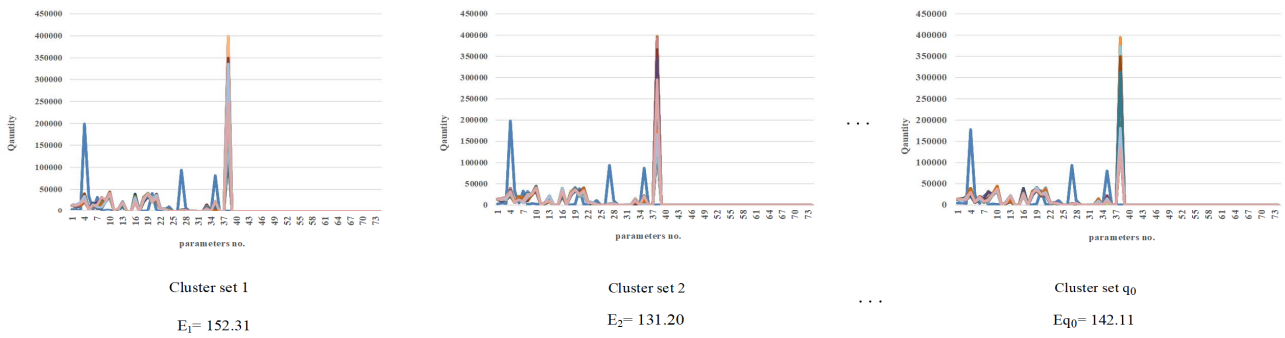


Fig. 9. The set of initial solutions (each line represents a cluster of a set of solutions) for the Electricity Distribution Grid No. 3 on 02/08/2012.

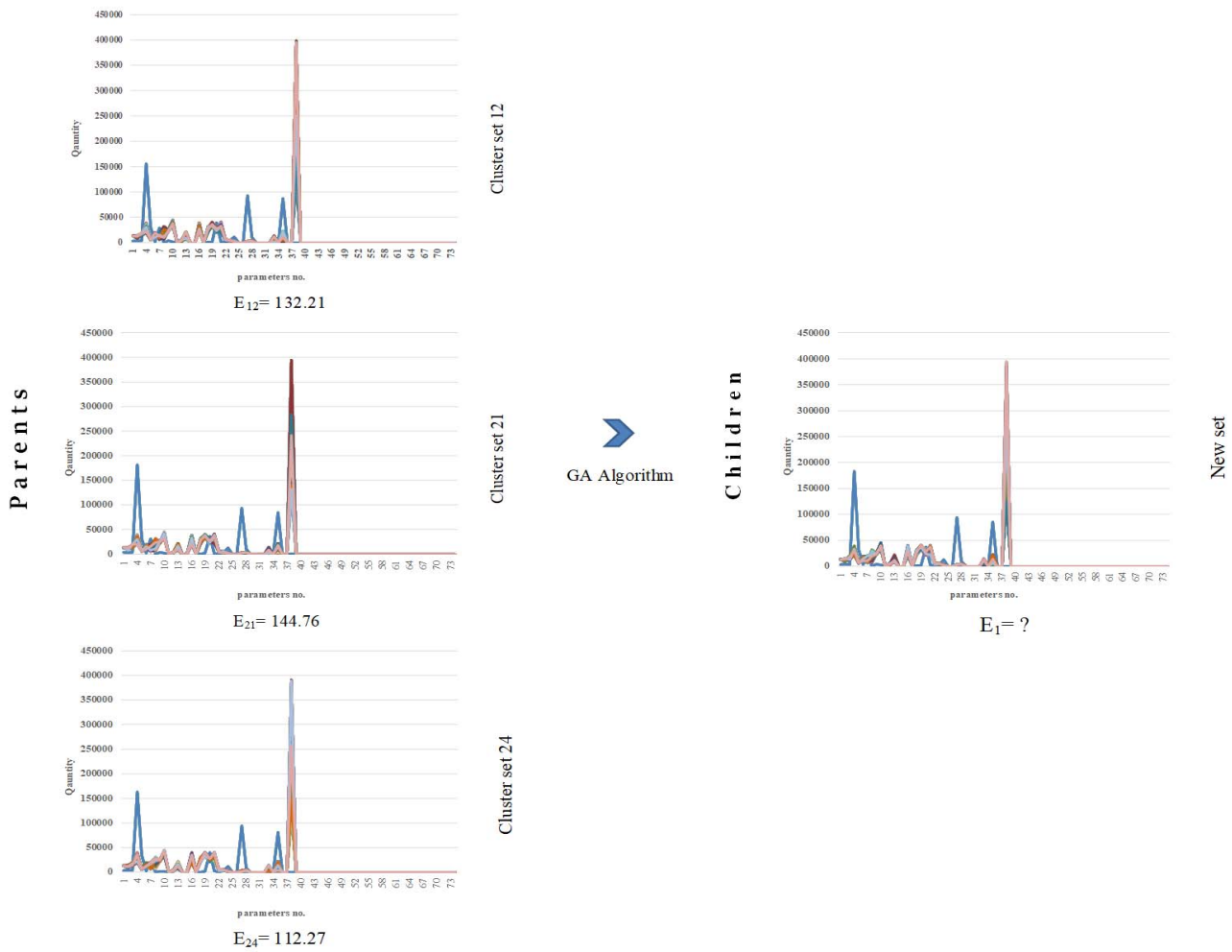


Fig. 10. Production of a new answer from the answers in the first iteration for the Electricity Distribution Grid No. 3 to estimate demand on 02/08/2012

F. Improvement of random solution based on selected random samples

1) *Crossover operator.* This operator is used to concentrate and integrate the solutions and to produce new children from parents. In this algorithm, a two-point operator is used as crossover operator. For data examined as characters, the crossover operator involves changing the number of each type of letter or number.

2) *Mutation operator.* This type of operator produces the new solution by selecting and combining three solutions. The value of F is selected randomly. The relation used for this operator is as follows: $X_{new} = X_1 + F(X_2 - X_3)$. For data examined as characters, the mutation involves changing the type of data. In other words, the mutation in this type of data changes the letter or number codes in the column.

3) *Selection (Monte Carlo simulation during optimization).* After performing mutation and crossover

operators and generating new solutions, the random numbers generated in the simulation process need to be assigned to the nearest new cluster to obtain the value of the objective function or the value of $E_{i'}$ for new solutions (Fig. 10)). The total distance of random points from point centers is considered as the fitness function. The fitness function of new points, calculated by the aforementioned method, will replace the previous solutions if it is better than the objective function of the old solutions.

G. Monte Carlo Simulation

In this section, Monte Carlo simulation is used. The Monte Carlo is a computational algorithm that uses random sampling in calculations. Due to reliance on iterative calculations and random or pseudo-random numbers, the Monte Carlo methods are adjusted to be run by computers. The Monte Carlo methods are often used when it is impossible to calculate the exact solutions by certain algorithms. In this problem, randomized simulation is repeated D_0 times and M_0 points are generated in each run (Fig. 11).

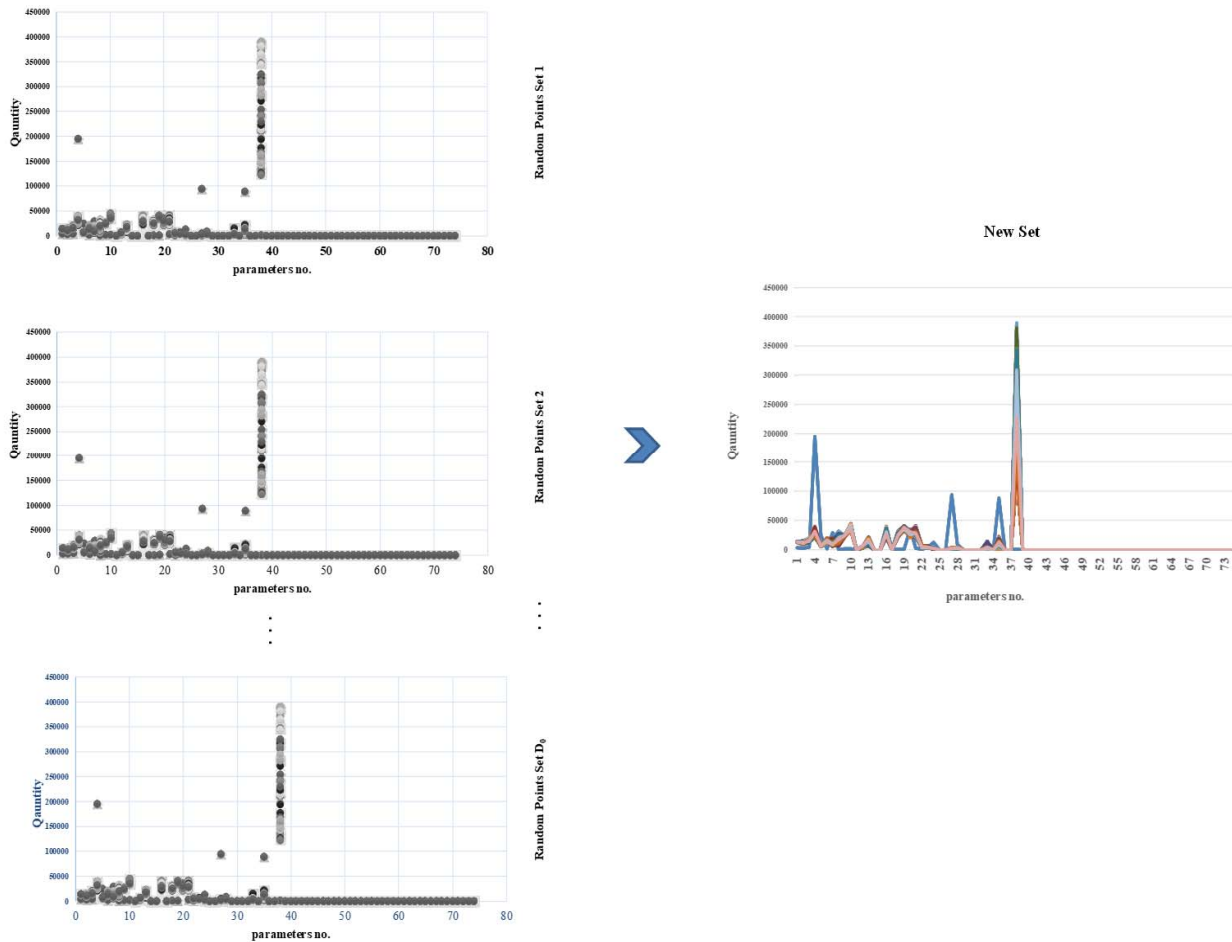


Fig. 11. Simulation to estimate the error rate for the answer generated in the first repetition to estimate the demand of electricity in the Distribution Grid No. 3 on 02/08/2012

H. Error calculation

For each of the simulated answers, the minimum error is calculated by comparing the simulated values with the cluster set. The histogram shown in Fig. 12 compares the values obtained from the simulation with the cluster set produced in the previous steps. The average error rate for the solutions generated in the first iteration is 87.70 which is considered as the estimated value for $d_i(p, q)$. $RSMD_i(\hat{\theta})$ is estimated to be 15.71 by calculating the prediction error with random samples. So $E_{i'}$ equals 103.41.

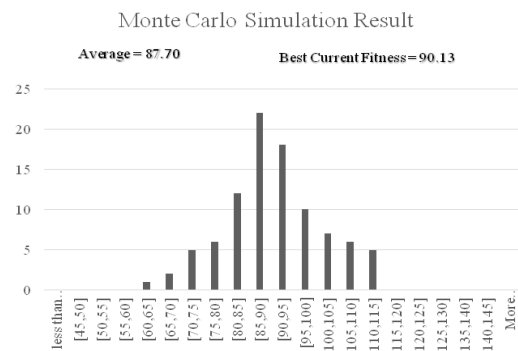


Fig. 12. Simulation results for the generated solutions in the first repetition for the Electricity Distribution Grid No. 3 on 08.02.2017

I. Adding existing solutions to new solutions

The answer set lead to better evaluation criterion (E_i') during simulation. By eliminating least desirable answers, the best answers are kept at each simulation stage. The capacity of solution set is equal to the constant parameter C_0 in the range of 10 to 25. This is optimized and adjusted during the simulation. For example, in the sample mentioned in the previous steps, since the new solution is better than the parent solution, it replaces the worst available solution, i.e. the cluster set 21.

J. Stop criterion

If the number of repetitions is greater than n_0 or the lowest value of evaluation criterion E_i' from the set of

solutions is less than U_0 , $\text{Min}(E_i') \leq U_0$, the algorithm stops. Otherwise the algorithm will continue.

K. Selecting prediction method

After calculating the best classification by running simulation-optimization algorithm on random samples (for example, for Electricity distribution company No. 3 on 08.02.2017 (Fig. 13)), it is necessary to estimate the demand and production quantities of power plants.

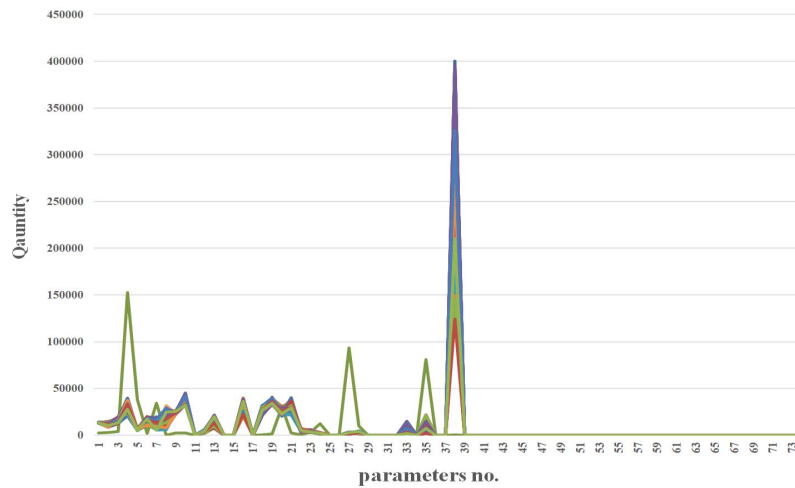


Fig. 13. The best cluster generated by KMSOA algorithm

Since these values can falsely announced upper and lower by buyers and sellers for more profit, the management of the electricity market predicts them based on other electricity market data optimized clustering to provide fair electricity prices. If the predicted value by the new data mining algorithm is greater or less than the specified limit μ

$\pm \alpha \sigma$, prediction is made by data mining. Otherwise, the NN-wavelet method is used to predict these values. For the period (08/21/2017), the predicted demand is out of this range by 1146.53 (Fig. 14). Thus, data mining is used for predictions on this date (Section 4.5).

Current Real Quantity	Parameters			000	73	Demand	$d_i(p, q)$	
	1	2	3					
	[3989	8800	15801					38
								1]
clusters	1	2	3	000	73	Demand	$d_i(p, q)$	
1	13456	8756	14987	000	2	1100	252.42	
2	13086	8997	15164	000	1	980	181.67	
3	13199	9626	16297	000	0	850	177.21	
.	
.	
.	
12	13706	8723	16205	000	0	1146.53	79.56	Best*
.	
.	
42	13630	13861	18313	000	1	1550	223.42	

Fig. 14. Estimated demand for the Distribution Grid No. 3 on 08/20/2017 from the KMSOA algorithm

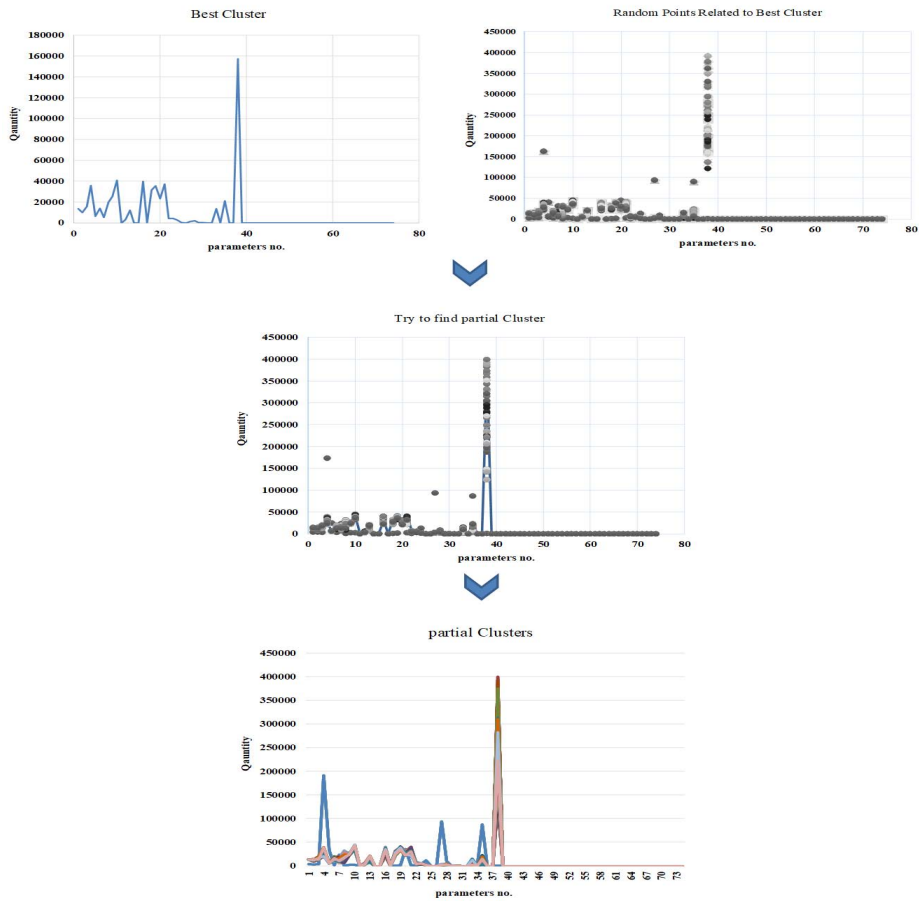


Fig. 15. Calculating the subset of the cluster to estimate the demand for the Distribution Grid No. 3 on 02/08/2017

L. Data mining forecasting

In the data mining forecasting, the cluster with the smallest distance with the values of the presumed parameters is considered. Then the random data that formed this cluster are identified and clustering is repeated for this group of samples with specific clusters. In this way, partial clusters of these data are obtained. After obtaining a subset of new clusters, the prediction of the supply or demand is performed based on the new cluster set (Fig. 15). The output from new clustering to forecast the demand of the distribution company No. 3 on 08/07/2012 is 1345.87 megawatts per hour.

M. Time series prediction

If the predicted value of supply or demand from the new algorithm is within range, predictions will be based on the NN-wavelet method as proposed by Niimura et al. (2002). The results of this method to estimate the demand of the electricity distribution company No. 3 from 22/07/2012 to 22/07/2012 are listed in Table 4. As is clear from the difference in average of the execution or non-execution of the new algorithm, an average improvement of 0.01 has been achieved for these 20 time periods. This method is used to calculate the demand of all electricity distribution companies and the supply of power plants.

TABLE IV. THE RESULTS OF THE KMSOA ALGORITHM FOR PREDICTING THE DEMAND OF THE ELECTRICITY DISTRIBUTION COMPANY NO.3 IN 20 TIME PERIODS FROM 2/8/2017

Date	Predicted value by KMSOA	$\mu \pm \alpha\sigma$	Demand prediction by NN-wavelet	Prediction by data mining	Actual demand	Error without implementing the algorithm	Error by implementing the algorithm
2/8/2017	1146.53	1131.12	1350.54	1345.87	1102.23	0.816	0.819
3/8/2017	1138.21	1132.21	1345.67	1256.76	1119.23	0.832	0.891
4/8/2017	1136.67	1121.21	1654.54	1563.43	1145.21	0.692	0.732
5/8/2017	1124.32	1134.56	1234.23	1234.23	1112.31	0.901	0.901
6/8/2017	1129.67	1128.11	1128.89	1003.32	1127.21	0.999	0.890
7/8/2017	1118.89	1129.21	1119.5	1119.5	1121.31	0.998	0.998
8/8/2017	1134.56	1132.12	1122.775	987.98	1116.21	0.994	0.885
9/8/2017	1121.53	1132.21	1115.45	1115.45	1114.32	0.999	0.999
10/8/2017	1123.23	1132.54	1112.275	1112.275	1114.34	0.998	0.998
11/8/2017	1112.21	1135.87	1108.8	1108.8	1112.23	0.997	0.997
12/8/2017	1112.23	1132.85	1100.25	1100.25	1103.23	0.997	0.997
13/8/2017	1132.21	1126.67	1124.87	1232.44	1136.23	0.990	0.922
14/8/2017	1123.23	1129.87	1100.025	1100.025	1101.34	0.999	0.999
15/8/2017	1124.56	1136.67	1101.37	1101.37	1109.32	0.993	0.993

Date	Predicted value by KMSOA	$\mu \pm \alpha\sigma$	Demand prediction by NN-wavelet	Prediction by data mining	Actual demand	Error without implementing the algorithm	Error by implementing the algorithm
16/8/2017	1112.21	1134.54	1104.32	1104.32	1105.67	0.999	0.999
17/8/2017	1129.89	1136.34	1106.25	1106.25	1105.54	0.999	0.999
18/8/2017	1137.72	1132.23	1110.55	1110.55	1111.98	0.999	0.999
19/8/2017	1128.21	1135.51	1106.15	1106.15	1108.67	0.998	0.998
20/8/2017	1125.21	1134.43	1115.32	1115.32	1113.87	0.999	0.999
Average						0.948	0.958

N. Modeling to calculate the fair electricity price

To predict the price of electricity in subsequent periods, it is purely predicted on the basis of past data in many articles. However, the management of the electricity market takes the best decision or pricing considering the terms of buyers and sellers and the costs incurred by both parties and other existing restrictions based on the obtained information [31].

Therefore, after forecasting the demand of all electricity distribution companies and the total production of power plants using the algorithm, these data along with other assumed parameters in Table 5 are entered into the following model. The fair electricity price is calculated by cost minimization and is used as the criterion of electricity exchange in the electricity market. The proposed model is calculated for 20 periods from 02/07/2012 to 08/22/2012.

TABLE V. INDICES, PARAMETERS AND VARIABLES OF THE MULTI-OBJECTIVE MODEL

Indices		The load generated by the power plant i in the time period t (from KMSOA)	q_{it}
Power plant	$i \in I$	The fixed cost of transmission by the transmission company j at t	S'_{jt}
Electricity transmission companies	$j \in J$	Spinning reserve of the unit k in the time period t	R_{kt}
Electricity distribution companies	$k \in K$	Maximum up time of the power plant i	T_i^{up}
Period	$t \in T$	Minimum down time of the power plant i	T_i^{down}
Sets		Maximum down time of the power plant i	T_j^{up}
Power plants	I	Minimum up time of the transmission company j	T_j^{down}
Electricity transmission companies	J	Payment by the unit k in the time period t	A'_{kt}
Electricity distribution companies	K	On time duration of the unit i at t	$X_{on,i}^{(t)}$
Period	T	Off time duration of the unit i at t	$X_{off,i}^{(t)}$
Parameters		On time duration of the unit j at t	$X_{on,j}^{(t)}$
Scheduled periods	T	Off time duration of the unit j at t	$X_{off,j}^{(t)}$
Total number of power plants	N	Percentage profit from power production in the power plant at t	\varnothing_{it}
Total number of electricity transmission companies	M	Percentage profit from transmission by transmission company at t	ϵ_{jt}
Total number of electricity distribution companies	L	The variable cost of load transmission by the company j from power plant i to the distribution company k at t	C'_{ijkt}
Maximum capacity of the power plant i	$q_{max\ i}$	The fixed cost of the transmission by the transmission company j at time t	S'_{jt}
Minimum capacity of the power plant i	$q_{min\ i}$	The fixed production cost in the power plant i at time t	S_{it}
Maximum capacity of the transmission company i	$q'_{max\ j}$	Variables	
Minimum capacity of the transmission company i	$q'_{min\ j}$	If the unit i is on in the period t , it equals 1 otherwise zero	u_{it}
Demand for the unit k at t (from KMSOA)	d_{kt}	The load transmitted by the transmission company j from the power plant i to the distribution company k in period t	q'_{ijkt}
The maximum number of power plants allowed for electricity supply by electricity transmission companies at t	A_t	Electricity prices for the power plant i in period t	P_{it}
The maximum number of transmission companies allowed to supply electricity by electricity distribution companies at t	B_t	If the unit j is on in the period t , it equals 1 otherwise zero	u'_{jt}
The variable cost of generating a load unit in the power plant i at time t	C_{it}	If the load of the distribution company k is supplied from the power plant i by the transmission company j in the period t , it equals 1 otherwise zero	ω_{ijkt}

1. The total operation costs are minimized as follows:

$$\text{Min } Q = \sum_{i=1}^N \left[\sum_{t=1}^T u_{it} [C_{it}(q_{it}) + S_{it}] \right] + \sum_{i=1}^N \sum_{j=1}^M \sum_{k=1}^L \sum_{t=1}^T v_{jt} [C'_{ijkt}(q'_{ijkt}) + S'_{jt}]$$

2. The total electricity cost of power plants is minimized as follows:

$$\text{Min } R = \sum_{i=1}^M \sum_{t=1}^T P_{it}$$

3. The cost of non-supply by electricity distribution companies is minimized as follows:

$$\text{Min } S = \left| \sum_{i=1}^N \sum_{j=1}^M \sum_{t=1}^T \omega_{ijkt} q'_{ijkt} - \sum_{t=1}^T d_{kt} \right| \forall k \in K.$$

Thus, the overall objective function is defined as follows where φ , φ' and φ'' respectively represent the first to third objectives so that $\varphi + \varphi' + \varphi'' = 1$.

$$\text{Min } U = \varphi Q + \varphi' R + \varphi'' S.$$

4. The total electricity demand and supply is calculated from data mining: $\sum_{j=1}^M \sum_{k=1}^L q'_{ijkt} \leq q_{it} \forall i \in I, \forall t \in T$ and

$$\sum_{i=1}^N \sum_{j=1}^M q'_{ijkt} \leq d_{kt} \forall k \in K, \forall t \in T.$$

5. The minimum spinning reserve at time t is as follows:

$$\sum_{i=1}^N \sum_{j=1}^M \omega_{ijkt} q'_{ijkt} > d_{kt} + R_{kt} \forall k \in K, \forall t \in T.$$

6. This constraint includes minimum and maximum production and the minimum up and down time for production and transmission units. This is among operation constraints of production units:

$$u_{it} q_{\min i} \leq \sum_{j=1}^M \sum_{k=1}^L \omega_{ijkt} q'_{ijkt} < u_{it} q_{\max i} \forall i \in I, \forall t \in T,$$

$$(X_{on,i}^{(t)} - T_i^{up})(u_{it}) \geq 0 \forall i \in I, \forall t \in T,$$

$$(X_{off,i}^{(t-1)} - T_i^{down})(1 - u_{it}) \geq 0 \forall i \in I, \forall t \in T.$$

6 - This constraint includes minimum and maximum production and the minimum up and down time for production and transmission units. This is among operation constraints of production units:

$$u'_{jt} q'_{\min j} \leq \sum_{i=1}^N \sum_{k=1}^M \omega_{ijkt} q'_{ijkt} < u'_{jt} q_{\max j} \forall k \in K, \forall t \in T,$$

$$(X_{on,j}^{(t)} - T_j^{up})(u'_{jt}) \geq 0 \forall j \in J, \forall t \in T,$$

$$(X_{off,j}^{(t)} - T_j^{down})(1 - u'_{jt}) \geq 0 \forall j \in J, \forall t \in T.$$

7. Payment dependence constraint to ensure that all winning units will cover proposed unloading, commissioning and energy generation cost is shown as follows:

$$A'_{kct} \geq \left(\sum_{i=1}^N [u_{it} [C_{it}(q_{it}) + S_{it}]] + \sum_{j=1}^M \sum_{k=1}^L u'_{jt} [C'_{ijkt}(q'_{ijkt}) + S'_{jt}] \right) \left(\sum_{i=1}^N \sum_{j=1}^M \omega_{ijkt} \right)$$

8. Supply chain equilibrium: The total load transmitted by each transmission company from different power plants to different distribution companies is at most equal to the total production of power plants at any time.

$$\sum_{i=1}^N u_{it} q_{it} \geq \sum_{i=1}^N \sum_{k=1}^L \omega_{ijkt} q'_{ijkt} \forall j \in J, \forall t \in T$$

9. The relationship between profits, costs and revenue is shown as follows:

$$P_{it} = (\mathbf{1} | \boldsymbol{\theta}) \left(\sum_{i=1}^N [u_{it} [C_{it}(q_{it}) + S_{it}]] \right) | (\mathbf{1} | \boldsymbol{\epsilon}) \left(\sum_{i=1}^N \sum_{j=1}^M u'_{jt} [C'_{ijkt}(q'_{ijkt}) + S'_{jt}] \right)$$

10. Transmission companies can communicate at most with a specific number of power plants. On the other hand, distribution companies can communicate at most with a specific number of the transmission companies:

$$\sum_{i=1}^N \sum_{k=1}^L \omega_{ijkt} \leq A_t \forall j \in J, \forall t \in T \text{ and}$$

$$\sum_{i=1}^N \sum_{j=1}^M \omega_{ijkt} \leq B_t \forall k \in K, \forall t \in T.$$

11. If the power plant i is on at time t, it should be on in the next two periods. If the transmission company j is on at time t, it should be on in the next two periods:

$$u_{it} < u_{i(t+1)} + u_{i(t+2)} \forall i \in I, \forall t \in T,$$

$$u'_{jt} < u'_{j(t+1)} + u'_{j(t+2)} \forall j \in J, \forall t \in T.$$

VI. OUTPUTS OF THE PROPOSED ALGORITHM AND MODEL

The proposed model is solved with the help of genetic algorithm in Matlab2016® considering 201 power plants, 16 electricity transmission companies and 39 electricity distribution companies. The output of the cost minimization model for 20 times is shown in Figure 16. Electricity supply condition for the Electricity Distribution Company No. 3 on 2/8/2017 is shown in Figure 17.

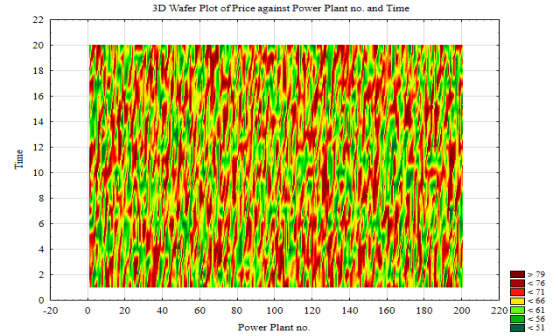


Fig. 16. The fair prices of electricity for 201 power plants in 20 periods from 02/08/217

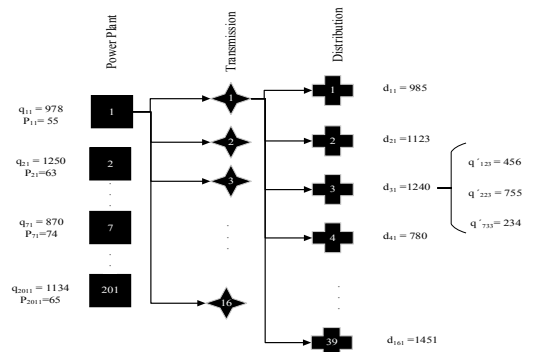


Fig. 17. Electricity supply for the Electricity Distribution Company No. 3 on 2/8/2016

VII. CONCLUSION AND RELATED WORK

The process of producing data in various fields of industry and commerce is moving towards the creation of a

database of big data. The electricity market is no exception to this trend and the current and future structure of this market (which was studied in Iran as a case study) represents a move toward big data generation. Big data requires new algorithms to solve associated problems with this type of data such as analysis, storage, display, and so on. In this study, a new KMSOA algorithm was introduced to solve big data analysis problems. This method allows analysis of structured, semi-structured, and unstructured data. The electricity market big data were clustered by the proposed algorithm to predict important input variables of the model including the amount of electricity supply and demand. The multi-purpose model defined with the aim of minimizing costs imposed on buyers and sellers was implemented to calculate a fair electricity price. The electricity market manager will announce this fair price for the upcoming period. The proposed algorithm shows a reasonable speed and accuracy for big data analysis. The accuracy of the new algorithm was compared with existing algorithms for estimating the price of electricity and

the results are shown in Table 6. The performance of this algorithm is compared with the SVMGA algorithm presented by Lee J *et al.* (2014) [28] in terms of runtime at different data volumes. Since both methods are GA-based algorithms, the results indicate the positive performance of the new algorithm as shown in Figure 18.

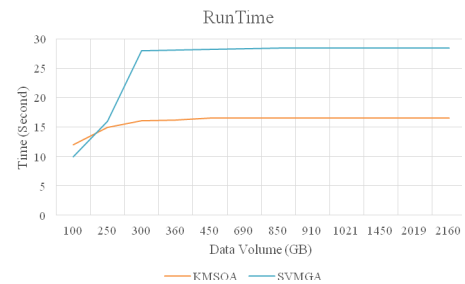


Fig. 18. Comparing the new algorithm with SVMGA in terms of runtime with increasing data volume

TABLE VI. COMPARING THE PREDICTION ACCURACY FOR 20 PERIODS

Time	Time series data mining by NN-wavelet method	Time series data mining and data mining by Lu <i>et al.</i> method (2005) [13]	Time series data mining and KMSOA data mining	Time	Time series data mining by NN-wavelet method	Time series data mining and data mining by Lu <i>et al.</i> method (2005) [13]	Time series data mining and KMSOA data mining
1	0.92	0.92	0.92	11	0.91	0.92	0.94
2	0.91	0.93	0.94	12	0.94	0.94	0.94
3	0.85	0.90	0.95	13	0.81	0.84	0.90
4	0.92	0.95	0.98	14	0.86	0.93	0.94
5	0.86	0.86	0.86	15	0.92	0.92	0.92
6	0.89	0.90	0.91	16	0.93	0.94	0.95
7	0.93	0.93	0.93	17	0.89	0.92	0.91
8	0.88	0.91	0.94	18	0.91	0.91	0.91
9	0.93	0.93	0.93	19	0.90	0.92	0.94
10	0.87	0.92	0.95	20	0.90	0.90	0.90
Total Average Error					0.89	0.91	0.93

As mentioned, a small percentage of improvement in the forecast of electricity prices will lead to substantial savings for buyers and sellers [25], because of high volume of daily electricity exchanges. For example, for 20 predicted periods in Table 6, the error rate is reduced by 0.02 on average. This may lead to \$ 26231040 saving in 20 time periods and \$ 478716480 a year for buyers and sellers:

$$E (\text{savings for 20 periods in the electricity market})$$

$$= E (\text{improvement of the predicted values of the electricity price for 20 periods}) * E (\text{average price of electricity for 20 periods}) * E (\text{average electricity purchased and sold for the 20 periods}) = 873672 * \$ 75 * 0.02 = \$ 26,231,040$$

$$E (\text{average savings for one year in the electricity market})$$

$$= E (\text{average saving for 20 periods}) * (365/20) = \$ 71750313$$

This huge saving is achieved by improvements in the prediction of a fair electricity price. This also shows the necessity of smart methods and the use of meta-heuristic algorithms for the analysis of big data in this type of market (with a very high exchange rate and data volume). On the other hand, decisions need to be made at a very high speed in a smart grid. The new algorithm showed positive results in high-speed clustering as a requirement for rapid decision-making by actors in the electricity market. This result shows that, these results can use in short term forecasting of demand and supply of electricity market.

REFERENCES

- [1] Zikopoulos, Paul; deRoos, Dirk; Parasuraman, Krishnan; Deutsch, Thomas; Giles, James; Corrigan, David Harness, the Power of Big Data The IBM Big Data Platform, McGraw-Hill Osborne Media, 2012
https://www.ibm.com/developerworks/community/blogs/SusanVisser/entry/flashbook_harness_the_power_of_big_data_the_ibm_big_data_platform24?lang=en
- [2] G.Halevi, H.Moed, The evolution of big data as a research and scientific topic: Overview of the literature, Research Trends,(2012) 3-6.
- [3] C.K.Emani , N.Cullot , C.Nicolle, Understandable Big Data: A survey, Computer Science Review, 17 (2015) 70-81.
- [4] K.Kambatla , G.Kollias , V.Kumar, A.Grama, Trends in big data analytics , Journal of Parallel and Distributed Computing, 74 (2014) 2561-2573.
- [5] K.Zhoua, C.Fua, S.Yanga, Big data driven smart energy management: From big data to big insights, Renewable and Sustainable Energy Reviews,56 (2016) 215-255.
- [6] N.Amjady , F.Keynia, Day ahead price forecasting of electricity markets by a mixed data model and hybrid forecast method, International Journal of Electrical Power & Energy Systems, 30 (9) (2008) 533-546.
- [7] O.Y.AI-Jarraha, P.D.Yoob, S.Muhaidatc, G.K.Karagiannidisd, K.Tahaa, Efficient machine learning for big data: A Review, Big Data Research, 2 (3) (2015) 87-93.
- [8] J.Shahrabi, M.A.Adibi, M.Mahootchi, A reinforcement learning approach to parameter estimation in dynamic job shop scheduling, Computers & Industrial Engineering, (2017).
<http://www.sciencedirect.com/science/article/pii/S0360835217302309>

- [9] P.Ringler, D.Keles, W.Fichtner, Agent-based modelling and simulation of smart electricity grids and markets – A literature review, *Renewable and Sustainable Energy Reviews*, 57 (2016) 205-215.
- [10] K.Zhoua, C.Fua, S.Yanga, Big data driven smart energy management: From big data to big insights, *Renewable and Sustainable Energy Reviews*,56 (2016) 215-255.
- [11] P.D. Diamantoulakish, V.M.Kapinasb, G.K. Karagiannidis, Big Data Analytics for Dynamic Energy Management in Smart Grids, *Big Data Research*,2 (3) (2015) 94-101.
- [12] M.G.Pollitt, Lessons from the history of independent system operators in the energy sector, *Energy Policy*, 47(2012) 32-48.
- [13] X.Lu, Z.Y.Dong, X.Li, Electricity market price spike forecast with data mining techniques,*Electric Power Systems Research*, 73(2005) 19-29.
- [14] <http://folk.uib.no/secea/databank/reservcapacity/Electricity%20market%20price%20spike%20forecast%20with%20data%20mining%20techniques%20Electric%20Power%20Systems%20Research,%20Volume%2073,%20Issue%201,%20January%202005,%20Pages%2019-29.pdf>
- [15] R.BayindirA,ColakbG.FullicK.Demirtasd, Smart Grid Technologies and application, *Renewable and Sustainable Energy Reviews*, 66(2016) 499-516.
- [16] K.P.Subbu, A. V.Vasilakos, Big data for context Aware computing – Perspectives and challenges , *Big Data Research*, 10(2017) 33-43. <http://www.sciencedirect.com/science/article/pii/S2214579616300077>
- [17] M.Drugan,M.Wiering,P.Vamplew,M.Chetty, Special issue on multi-objective reinforcement learning, *Neurocomputing*, 263, (2017),1-2.
- [18] P.Ringler,D.Keles,W.Fichtner , Agent-based modelling and simulation of smart electricity grids and markets – A literature review, *Renewable and Sustainable Energy Reviews*,57 (2016) 205-215.
- [19] Electricity Market Forecasting via Low-Rank Multi-Kernel Learning, *Ieee Journal Of Selected Topics In Signal Processing*, 8 (2016) 1182-1193.
- [20] V.Kekatos, G.B. Giannakis, R.Baldick Online Energy Price Matrix Factorization for Power Grid Topology Tracking, *IEEE TRANS. ON SMART GRID*, 2015 <http://www.faculty.ece.vt.edu/kekatos/papers/TSG2015.pdf>
- [21] A.J. Conejo, M.A. Plazas, R. Espinola , Day-ahead electricity price forecasting using wavelet transform combined with ARIMA and G, *Applied Energy*, 87 (2010) 3606–3610.
- [22] V.Fanelli, L.Maddalena,S.Musti , Modelling electricity futures prices using seasonal path dependent volatility, *Applied Energy*, 173 (2016) 92-102.
- [23] Y.Zhang, Ce.Li, L.Li, Electricity price forecasting by a hybrid model combining wavelet transform-ARMA and kern, *Applied Energy* ,190 (2017) 291–305.
- [24] N.Nezamoddini,Y.Wang, Real-time electricity pricing for industrial customers Survey and case studies in-the United State, 195(2017) 1023-1037.
- [25] M.Naimur, R.A.Esmailpour, J.Zhao, Machine Learning with Big Data An Efficient Electricity Generation Forecasting System, *Big Data Research*, 5(2016) 9-15.
- [26] S.Islyae,P.Date , Electricity futures price models: Calibration and forecasting, *European Journal of Operational Research*, 247(1) (2015) 144-154.
- [27] H.T.Pao , Forecasting electricity market pricing using artificial neural networks, *Energy Conversion and Management* , 48 (2007) 907–912.
- [28] H.M.I.Pousinho, V.M.F.Mendes, J.P.S.Catalão, Short-term electricity prices forecasting in a competitive market by a hybrid intelligent approach, *Energy Conversion and Management*, 52(2) (2011) 1061-1065.
- [29] Lee J, Hong S, Lee JH. An efficient prediction for heavy rain from big weather data using genetic algorithm. In: Proceedings of the international conference on ubiquitous information management and communication; 2014, 25:1- 25:7.
- [30] M.Bennedsen ,A rough multi-factor model of electricity spot price, *Energy Economics*, 63(2017) 301-313.
- [31] A.Mirakyan, M.M.Renschhausen, A.Koch, Composite forecasting approach, application for next-day electricity price forecasting, *Energy Economics*, 2016.
- [32] R.Miller,L.Golab,C.Rosenberg Modelling weather effects for impact analysis of residential time-of-use electricity pricing, *Energy Policy*, 105(2017) 534-536.
- [33] R.W.Allmendinger,C.R.Siron,C.P.Scott, Structural data collection with mobile devices: Accuracy, redundancy, and best practices, *Journal of Structural Geology*, 102(2017) 98-112.
- [34] F.Zhan, N.Yao ,On the using of discrete wavelet transform for physical layer key generation, *Ad Hoc Networks*, 64(2017) 22-31
- [35] J.Wang, D.Ding, O.Liu, M.Li, A synthetic method for knowledge management performance evaluation based on triangular fuzzy number and group support systems, *Applied Soft Computing*, 39 (2016) 11-20.
- [36] X.Tang, J.Luo, F.Liu, Aerodynamic shape optimization of a transonic fan by an adjoint-response surface method, *Aerospace Science and Technology*, 68 (2017) 26-36.
- [37] D.Yamashita, A.M.Isa, R.Yokoyama, T.Niimura, Forecasting of Electricity Price and Demand Using Autoregressive Neural Networks, Proceedings of the 17th World Congress The International Federation of Automatic Control, 2008. <http://folk.ntnu.no/skoge/prost/proceedings/ifac2008/data/papers/3789.pdf>
- [38] J.C.Cuaresma, J.Hlouskova, S.Kossmeier, M.Obersteiner, Forecasting electricity spot-prices using linear univariate time-series models, *Applied Energy*, 77(1) (2004) 87-106
- [39] H.T.Pao, Electricity price modeling with stochastic time change, *Energy Economics*, 48(3) (2007) 907-912.
- [40] S.Islyae,P.Date, Electricity futures price models: Calibration and forecasting, *European Journal of Operational Research*, 247(2015) 144-154.
- [41] J. Shahrabi, M. Amin Adibi, M. Mahootchi ,A reinforcement learning approach to parameter estimation in dynamic job shop scheduling, *Computers & Industrial Engineering*, (2017), <http://dx.doi.org/10.1016/j.cie.2017.05.026/>
- [42] Beckmann M, Ebecken NFF, de Lima BSLP,Costa MA. A user interface for big data with rapidminer. *RapidMiner World*, Boston, MA, Tech. Rep.; 2014. [Online]. <http://www.slideshare.net/RapidMiner/a-user-interface-for-big-data-with-rapidminer-marcelo-beckmann>
- [43] E. Januzaj, H.P. Kriegel, M. Pfeifle DBDC: Density based distributed clustering. In: Proceedings of the advances in database technology, vol. 2992; 2004. pp. 88–105.
- [44] W. Zhao, H. Ma, Q. He Parallel k-means clustering based on mapreduce. *Proc Cloud Comp*. 2009;5931:674–9.
- [45] L. Yang, Z. Shi, L. Xu, F. Liang, I. Kirsh DH-TRIE frequent pattern mining on hadoop using JPA. In: Proceedings of the international conference on granular computing; 2011. pp. 875–878.
- [46] X Cui, J.S. Charles, T. Potok GPU enhanced parallel computing for large scale data clustering. *Future Gener Comp Syst*. 2013;29(7):1736–41.
- [47] KR. Ku-Mahamud Big data clustering using grid computing and ant-based algorithm. In: Proceedings of the international conference on computing and informatics; 2013. pp. 6–14.
- [48] S. Hasan, S. Shamsuddin, N. Lopes Soft computing methods for big data problems. In: Proceedings of the symposium on GPU computing and applications; 2013. pp. 235–247.
- [49] CS. Leung, R. MacKinnon, F. Jiang Reducing the search space for big data mining for interesting patterns from uncertain data. In: Proceedings of the international congress on big data; 2014. pp. 315–322.
- [50] J. Lee, S. Hong, JH. Lee An efficient prediction for heavy rain from big weather data using genetic algorithm. In: Proceedings of the international conference on ubiquitous information management and communication; 2014. pp. 25:1–25.

Topic #3

Hybrid Systems of Computational Intelligence

Hybrid Multidimensional Wavelet-Neuro-System and its Learning Using Cross Entropy Cost Function in Pattern Recognition

Olena Vynokurova
IT Step University
Lviv, Ukraine
vynokurova@gmail.com

Semen Oskerko
IT Step University
Lviv, Ukraine
semosker@gmail.com

Dmytro Peleshko
IT Step University
Lviv, Ukraine
dpeleshko@gmail.com

Viktor Voloshyn
IT Step University
Lviv, Ukraine
voloshyn_v@itstep.org

Yuriy Borzov
Department of Project Management,
Information Technologies and
Telecommunications
Lviv State University of Life Safety
Lviv, Ukraine
uob1968@gmail.com

Abstract— In this paper, the hybrid multidimensional wavelet-neuro-system for pattern recognition tasks is proposed. Also learning algorithm for all its parameters (synaptic weights, the centers, and widths of wavelet activation functions) based on cross entropy cost function was proposed. The proposed system is characterized by high learning speed and high approximation properties in comparison with well-known approaches. The efficiency of the proposed approach has been justified based on different benchmarks and real data sets.

Keywords— patterns recognition, hybrid wavelet-neuro-system, learning algorithm, wavelet transform, cross entropy cost function.

I. INTRODUCTION

Machine grouping, classification, and recognition of patterns are important problems in a variety of engineering and scientific areas, such as artificial intelligence [1-2], computer vision [3], internet of things (IoT) [4], biology, medicine [5], marketing, etc. The patterns could be the handwritten cursive words and symbols, the biometrical images, or a speech signal.

Nowadays the machine learning methods (especially artificial neural networks) [6-15] are widely spread for solving the pattern recognition and images classification tasks due to their universal approximating properties and their learning abilities. Since there's a number of practical tasks when a learning sample volume is restricted, a learning rate factor goes in the forefront.

However, not all approaches (first of all, based on multilayer architectures, which are learned using the error backpropagation procedure) satisfy to the conditions of the real tasks because of a low speed of a learning process and a possible overfitting effect.

Therefore, hybrid systems are the most effective systems in machine learning, especially neuro-fuzzy and wavelet-neuro-fuzzy systems that combine neural networks' universal approximation ability, fuzzy inference systems' interpretability and detection of the local features of patterns using wavelet transform.

Today, a lot of machine learning approaches are

proposed for pattern recognition and classification, among them in [16] authors have proposed the spiking neural network for pattern recognition and a learning algorithm based on the relative ordering of output spikes, in [17] approach for a face recognition based on recurrent regression neural network is proposed, in [18] authors have proposed a novel convolutional neural network for prediction of the emotion, in this case, the proposed model has two parts: classification network for a positive-or-negative emotion recognition and a deep neural network for specific emotion recognition, in [19] an efficient face feature extraction method based on local Gabor binary pattern histogram sequence and wavelet neural network for classification have been proposed, in [20] authors present a single image super resolution technique in which we estimate wavelet detail coefficients of a desired high resolution image using a convolutional neural network on the given low resolution image, in [21] a novel hybrid approach called switching particle swarm optimization-wavelet neural network has been proposed.

For most proposed approaches the choice of the type and parameters of the activation functions is the problem, which is solved by the empirical fit. To solve this problem, the hybrid system has to adjust all his parameters in process of training.

Therefore, in this paper, the architecture of hybrid multidimensional wavelet-neuro-system and its learning algorithm of all its parameters based on cross entropy cost function are proposed. The proposed hybrid system has only one layer of information processing and is characterized by high learning speed and increased approximation properties.

II. THE ARCHITECTURE OF HYBRID MULTIDIMENSIONAL WAVELET-NEURO-SYSTEM

The structure element of proposed hybrid multidimensional wavelet-neuro-system is one-dimensional wavelet neuron, which had been proposed in [22, 23]. Thereafter, in [24] the wavelet neuron with adaptive learning algorithm for the activation functions parameters using quadratic criterion had been proposed. But such learning algorithm is not effective for pattern recognition tasks, especially for the image classification tasks.

In generally, wavelet neuron consists of the wavelet synapses WS_i with the wavelet activation functions ($i=1,2,\dots,n$), where the synaptic weights $w_{ji}(k)$ are the adjustable parameters.

In many cases solving the real problems is needed the processing of multidimensional data, which are fed from some sources at one time. For this case, we have introduced multidimensional wavelet neuron [25] with the learning algorithm, which was constructed based on quadratic error function and only synaptic weights were adjusted.

Using the entropy cost function [26] and tuning the activation function parameters in process of learning systems will improve the approximating properties of such a system and improve the quality of information processing.

In this case, we propose hybrid multidimensional wavelet-neuro-system, which consists of two part: the first part is the generation subsystem of features vector and the second part is the pattern recognition subsystem based on the multilayer wavelet neuron. The architecture of the proposed system is shown on Fig.2.

The generation subsystem of features vector consists of the discrete two-dimensional wavelet transform [27] of the pattern and forming the input feature vector, which will be fed to the inputs of multidimensional wavelet neuron in the form

$$x(k) = [cA', cH', cV', cD'] \quad (1)$$

where cA', cH', cV', cD' are deployed by columns matrices of coefficients cA (the approximation coefficients matrix) and cH, cV and cD (horizontal, vertical, and diagonal detail coefficients matrices, respectively).

The subsystem of information processing consists of the multidimensional wavelet neuron, which has the one-layered architecture with n inputs, m outputs and h wavelet activation function for each input.

The input observation vector is fed to the input layer of multidimensional wavelet neuron in the form

$$x(k) = (x_1(k), x_2(k), \dots, x_n(k))^T \quad (2)$$

where $x_i(k)$ is the wavelet coefficients, k is a number of image in the data set.

The output of system can be written in the from

$$y_j(k) = \frac{1}{1 + \exp(-\gamma u_j)} \quad (3)$$

$$u_j(k) = \sum_{l=1}^h \sum_{i=1}^n \varphi_{li}(x_i(k)) w_{lij}(k) \quad (4)$$

where $w_{lij}(k)$ is j -th synaptic weight of l -th wavelet activation function of i -th input of neuron, $i=1\dots n$, $j=1\dots m$, $l=1\dots h$, y_j is j -th output of multidimensional wavelet neuron, $\varphi_{li}(x_i(k))$ - l -th wavelet activation function of i -th input of neuron, γ is a rise rate of sigmoidal activation function.

Here, we will use the one-dimensional wavelet activation function:

$$\varphi_{li}(x_i(k)) = (1 - t_{li}^2(k)) \exp\left(-\frac{t_{li}^2(k)}{2}\right) \quad (5)$$

where $t_{li}(k) = (x_i(k) - c_{li}(k)) \sigma_{li}^{-1}(k)$; $c_{li}(k)$ is the parameter that defines a location of the function center; $\sigma_{li}(k)$ is the parameter that defines the function width, which will be adjusted.

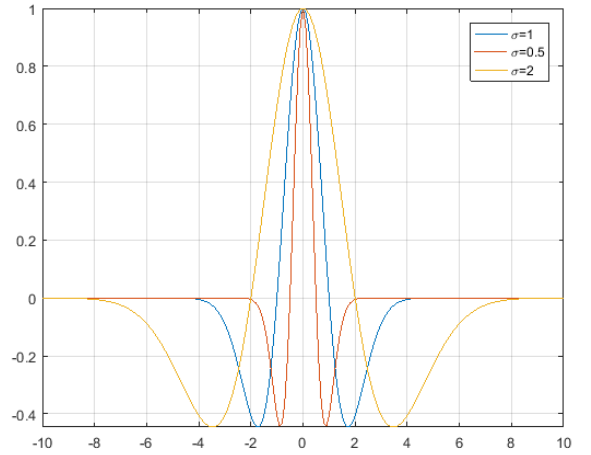


Fig. 1. Wavelet activation functions with different values of the width parameter σ_{ii}

For the optimization of computational implementation let's rewrite the input of multidimensional wavelet neuron in the form

$$y(k) = W(k) \varphi(x(k)) \quad (6)$$

where

$$\bar{\varphi} = (\bar{\varphi}_1, \dots, \bar{\varphi}_p, \dots, \bar{\varphi}_{hm})^T = (\varphi_{11}(x_1), \varphi_{12}(x_2), \dots, \varphi_{1n}(x_n),$$

$\varphi_{21}(x_1), \dots, \varphi_{hn}(x_n))^T$ is $(hn \times 1)$ dimension vector of wavelet activation functions,

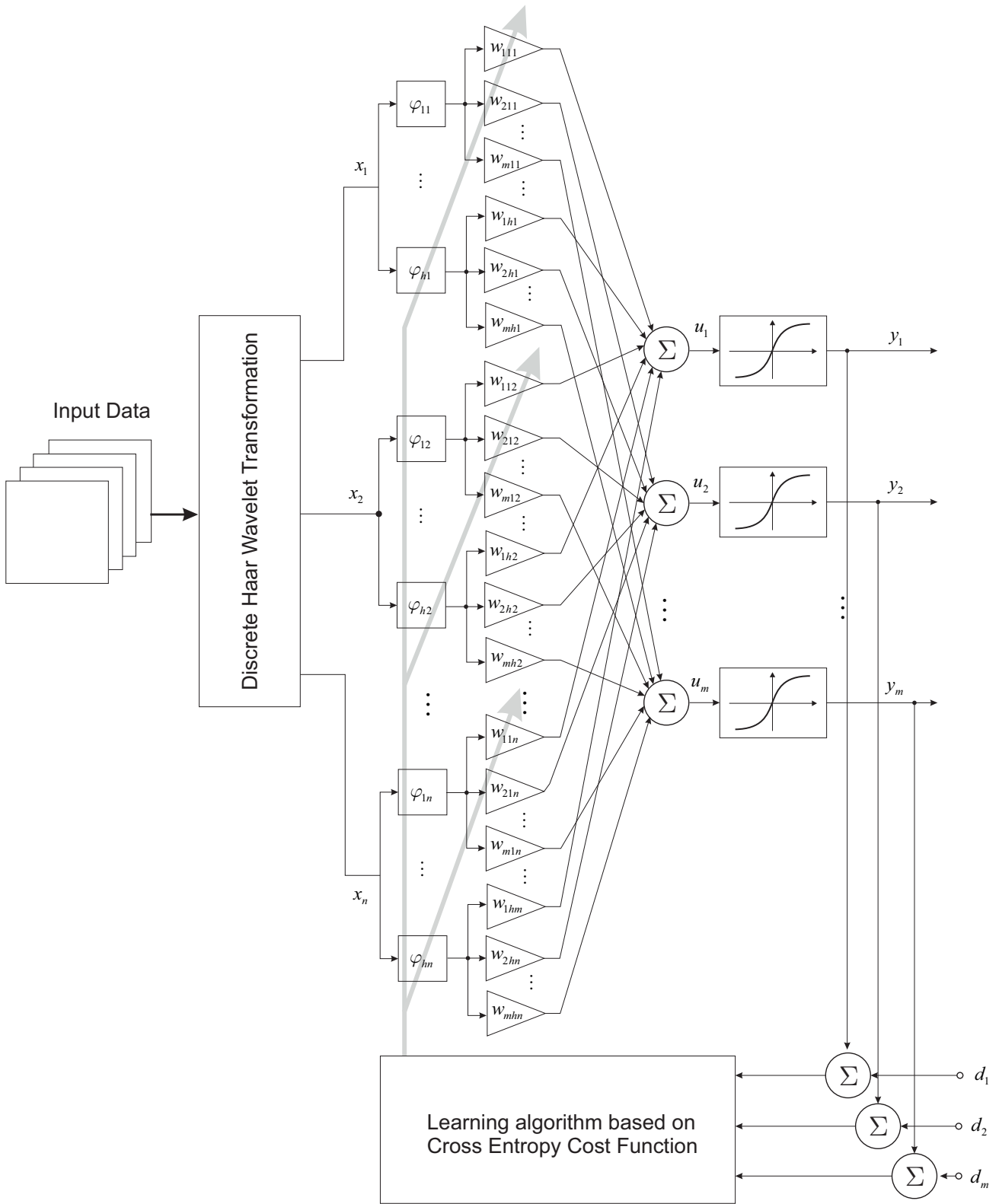


Fig. 2. The architecture of hybrid multidimensional wavelet-neuro-system

$$W(k) = \begin{pmatrix} w_{111} & w_{112} & \cdots & w_{11n} & w_{121} & \cdots & w_{1hn} \\ w_{211} & w_{212} & \cdots & w_{21n} & w_{221} & \cdots & w_{2hn} \\ \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ w_{m11} & w_{m12} & \cdots & w_{m1n} & w_{m21} & \cdots & w_{mhn} \end{pmatrix}$$

or for optimization of notation

$$W(k) = \begin{pmatrix} \bar{w}_{11} & \cdots & \bar{w}_{1p} & \cdots & \bar{w}_{1(h-n)} \\ \bar{w}_{21} & \cdots & \bar{w}_{2p} & \cdots & \bar{w}_{2(h-n)} \\ \vdots & \cdots & \ddots & \cdots & \vdots \\ \bar{w}_{m1} & \cdots & \bar{w}_{mp} & \cdots & \bar{w}_{m(h-n)} \end{pmatrix}$$

is $(m \times hn)$ matrix of synaptic weights.

III. THE LEARNING ALGORITHM OF HYBRID WAVELET-NEURO-SYSTEM

Due to the synaptic weights of multidimensional wavelet neuron depend on the output systems linearly, we can use the stochastic approximation algorithms, which minimize cross entropy cost function in the form

$$E(k) = \sum_{j=1}^m -d_j(k) \ln y_j(k) - (1-d_j(k)) \ln(1-y_j(k)) \quad (7)$$

Minimizing the criterion (7) by synaptic weights $\bar{w}_{jp}(k)$

$$\frac{\partial E(k)}{\partial \bar{w}_{jp}(k)} = -e_j(k) \gamma \varphi_p(k) \quad (8)$$

we can write learning algorithm in the form

$$\begin{aligned} \bar{w}_{jp}(k+1) &= \bar{w}_{jp}(k) - \eta \frac{\partial E(k)}{\partial \bar{w}_{jp}(k)} = \\ &= \bar{w}_{jp}(k) + \eta \gamma e_j(k) \varphi_p(k) \end{aligned} \quad (9)$$

where η is learning rate parameter ($0 < \eta \leq 1$).

For minimizing the criterion (7) by the centers and widths of the wavelet activation functions to introduce into consideration the notations in the form

$$\bar{c} = (\bar{c}_1, \bar{c}_2, \dots, \bar{c}_p, \dots, \bar{c}_{hn})^T = (c_{11}, c_{12}, \dots, c_{1n}, c_{21}, \dots, c_{hn})^T$$

is $(hn \times 1)$ dimension centers vector of wavelet activation functions and

$$\bar{\sigma} = (\bar{\sigma}_1, \bar{\sigma}_2, \dots, \bar{\sigma}_p, \dots, \bar{\sigma}_{hn})^T = (\sigma_{11}, \sigma_{12}, \dots, \sigma_{1n}, \sigma_{21}, \dots, \sigma_{hn})^T$$

is $(hn \times 1)$ dimension widths vector of wavelet activation functions.

Thus, we can write partial derivations by the centers and the widths parameters of wavelet activation functions in the form

$$\frac{\partial E(k)}{\partial \bar{c}_p(k)} = -\sum_{j=1}^m e_j(k) \gamma \bar{w}_{jp}(k) \frac{\partial \bar{\varphi}_p(k)}{\partial \bar{c}_p(k)}, \quad (10)$$

$p = 1 \dots hn$

$$\frac{\partial E(k)}{\partial \bar{\sigma}_p(k)} = -\sum_{j=1}^m e_j(k) \gamma \bar{w}_{jp}(k) \frac{\partial \bar{\varphi}_p(k)}{\partial \bar{\sigma}_p(k)}, \quad (11)$$

$p = 1 \dots hn$

and after that, we can write the learning algorithm for parameters of wavelet activation function

$$\begin{aligned} \bar{c}_p(k+1) &= \bar{c}_p(k) - \eta \frac{\partial E(k)}{\partial \bar{c}_p(k)} = \\ &= \bar{c}_p(k) + \eta \gamma \frac{\partial \bar{\varphi}_p(k)}{\partial \bar{c}_p(k)} \sum_{j=1}^m e_j(k) \gamma \bar{w}_{jp}(k), \end{aligned} \quad (12)$$

$$\begin{aligned} \bar{\sigma}_p(k+1) &= \bar{\sigma}_p(k) - \eta \frac{\partial E(k)}{\partial \bar{\sigma}_p(k)} = \\ &= \bar{\sigma}_p(k) + \eta \gamma \frac{\partial \bar{\varphi}_p(k)}{\partial \bar{\sigma}_p(k)} \sum_{j=1}^m e_j(k) \bar{w}_{jp}(k) \end{aligned} \quad (13)$$

where

$$\frac{\partial \bar{\varphi}_p(k)}{\partial \bar{c}_p(k)} = \tau_p(k) \bar{\sigma}_p^{-1}(k) (3 - \tau_p^2(k)) \exp(-\tau_p^2(k)/2)$$

$$\frac{\partial \bar{\varphi}_p(k)}{\partial \bar{\sigma}_p(k)} = \tau_p(k) (x_i(k) - \bar{c}_p(k)) \bar{\sigma}_p^{-2}(k) (3 - \tau_p^2(k)) \exp(-\tau_p^2(k)/2)$$

IV. EXPERIMENTS

The recognition task of handwritten digits was solved based on MNIST database [28]. A training set consists of the 60.000 images, and a test set consists of the 10.000 images. The initial bi-level images from NIST database were normalized. The multidimensional wavelet neuron has 784 inputs, 10 outputs, and 15 wavelet function for each input. The initial synaptic weights values were taken randomly in the interval $[0,1]$ and learning rate parameter was taken $\eta = 0.99$.

Table I shows the comparison results of the proposed systems with the existed approaches.

TABLE I. THE CLASSIFICATION RESULTS OF MNIST DATASET

Neural Network Classifier	Preprocessing Image	Test accuracy (%)
Hybrid multidimensional wavelet-neuro-system (1-layer NN , 15 wavelet activation function for each input, tuning synaptic weights, centers and width of wavelet activation function, cross entropy criterion)	none	96,5%
Hybrid multidimensional wavelet-neuro-system (1-layer NN , 15 wavelet activation function for each input), tuning synaptic weights, cross entropy	none	95,3%
Multidimensional wavelet neuron (1-layer NN , 15 wavelet activation function for each input, tuning synaptic weights, quadratic criterion)	none	94.5%
Linear 1-layer NN classifier	none	88.0%
Linear 1-layer NN classifier	deskewing	91.6%
Pairwise linear classifier	deskewing	92,4%
2-layer perceptron (300 sigmoidal activation functions in the hidden layer and 10 ones in the output layer)	none	96,4%
Spiking neural network [16]	none	90,3%

As the quality criterion was taken the percentage of the false classified objects based on the testing data image set.

As it may be inferred from the obtained results, the proposed multidimensional wavelet-neuro-system has the best quality of classification among 1-layer neural network classifiers. The 2-layer perceptron has the same quality of classification but has 3.5 times more adjustable parameters than the multidimensional wavelet neuron.

V. CONCLUSION

In this article, the architecture of hybrid multidimensional wavelet-neuro-system is proposed. Such system can be used as a classifier of the multidimensional data sets. The main advantage of the proposed hybrid system is a simplicity of implementation in the hardware of IoT applications because such system has the one-layered architecture of information processing. Also, the learning algorithm of hybrid multidimensional wavelet-neuro-system based on entropy cost function is proposed. This learning algorithm allows tuning not only synaptic weights but the parameters of wavelet activation functions.

The proposed system can be used for solving the problems in Big Data Processing, Computer Vision, IoT applications and Data Stream Mining.

The computational experiments are performed using benchmarks and real data sets. The obtained results have confirmed the advantages of the proposed approach in comparison with the existed methods.

REFERENCES

- [1] S. Babichev, M.A. Taif, V. Lytvynenko, Inductive model of data clustering based on the agglomerative hierarchical algorithm. In: Proc. of 2016 IEEE 1st International Conference on Data Stream Mining and Processing, DSMP 2016, pp. 19-22.
- [2] S. Babichev, M.A. Taif, V. Lytvynenko, V. Osypenko. Critical analysis of gene expression sequences to create the objective clustering inductive technology. In: Proc. of 2017 IEEE 37th International Conference on Electronics and Nanotechnology, ELNANO, 2017, pp. 244-248.
- [3] D. Peleshko, M. Peleshko, N. Kustra, I. Izonin, Analysis of invariant moments in tasks image processing. In: Proc. of 2011 11th International Conference the Experience of Designing and Application of CAD Systems in Microelectronics (CADSM), Polyana-Svalyava, 2011, pp. 263-264.
- [4] M.S. Mahdavinejad, M. Rezvan, M. Barekatin, P. Adibi, P. Barnaghi, A.P. Sheth.: Machine learning for Internet of Things data analysis: A survey. Digital Communications and Networks, 2017 (in press).
- [5] I. Pliss and I. Perova "Diagnostic Neuro-Fuzzy System and Its Learning in Medical Data Mining Tasks in Conditions of Uncertainty about Numbers of Attributes and Diagnoses" Automatic Control and Computer Sciences, , 51(6), 391-398, 2017.
- [6] Ye. Bodyanskiy, O. Vynokurova, I. Pliss, D. Peleshko, Hybrid adaptive systems of computational intelligence and their on-line learning for green it in energy management tasks In: Kharchenko V., Kondratenko Y., Kacprzyk J. (eds) Green IT Engineering: Concepts, Models, Complex Systems Architectures. Studies in Systems, Decision and Control, Springer, Cham, vol 74., pp. 229-244, 2017.
- [7] L. Rutkowski, Computational Intelligence: Methods and Techniques. Springer-Verlag, Berlin, 2008.
- [8] C.M. Bishop, Pattern Recognition and Machine Learning. Springer, Berlin, 2006.
- [9] K.P. Murphy, Machine Learning: a Probabilistic Perspective. MIT press, 2012.
- [10] Ye.V. Bodyanskiy, O.A. Vynokurova, A.I. Dolotov: Self-learning cascade spiking neural network for fuzzy clustering based on group method of data handling. Journal of Automation and Information Sciences, v. 45, pp.23-33., 2013.
- [11] Zh.Hu, Ye.V. Bodyanskiy, O.K. Tyshchenko, V.M. Tkachov: Fuzzy Clustering Data Arrays with Omitted Observations. International Journal of Intelligent Systems and Applications (IJISA) 9(6), pp.24-32, 2017.
- [12] Ye. Bodyanskiy, O. Vynokurova, G. Setlak, D. Peleshko, P. Mulesa.: Adaptive multivariate hybrid neuro-fuzzy system and its on-board fast learning. Neurocomputing, v. 230, pp. 409-416, 2017.
- [13] Zh. Hu, Ye.V. Bodyanskiy, O.K. Tyshchenko, V.O. Samitova: Possibilistic Fuzzy Clustering for Categorical Data Arrays Based on Frequency Prototypes and Dissimilarity Measures. International Journal of Intelligent Systems and Applications (IJISA), 9(5), pp.55-61, 2017.
- [14] Ye. Bodyanskiy, O. Vynokurova, I. Pliss, G. Setlak, P. Mulesa, Fast learning algorithm for deep evolving GMDH-SVM neural network in data stream mining tasks. In.: Proc. of 2016 IEEE 1st International Conference on Data Stream Mining and Processing, pp. 257-262, 2016.
- [15] Ye. Bodyanskiy, G. Setlak, D. Peleshko, O. Vynokurova, Hybrid generalized additive neuro-fuzzy system and its adaptive learning algorithms. In: Proc. of 2015 IEEE 8th International Conference on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications, IDAACS, pp. 328-333, 2015.
- [16] Zh. Lin, De Ma, J. Meng, L. Chen Relative ordering learning in spiking neural network for pattern recognition. Neurocomputing Volume 275, 31 January 2018, pp. 94-106
- [17] Ya. Li, W. Zheng, Zh. Cui, T. Zhang, Face recognition based on recurrent regression neural network. Neurocomputing, 2018 (in press).
- [18] Xu. He, W. Zhang, Emotion recognition by assisted learning with convolutional neural networks. Neurocomputing Vol. 291, 24 May 2018, pp. 187-194
- [19] P. Sharma, K.V. Arya, R.N. Yadav, Efficient face recognition using wavelet-based generalized neural network. Signal Processing, Vol. 93, Is. 6, June 2013, pp. 1557-1565.
- [20] Neeraj Kumar, Ruchika Verma, Amit Sethia Convolutional neural networks for wavelet domain super resolution Pattern Recognition Letters Volume 90, 15 April 2017, Pages 65-71
- [21] Ya. Lu, N. Zeng, Yu. Liu, N. Zhang, A hybrid Wavelet Neural Network and Switching Particle Swarm Optimization algorithm for face direction recognition. Neurocomputing, Vol. 155, 1 May 2015, pp. 219-224.
- [22] T. Yamakawa, A novel nonlinear synapse neuron model guaranteeing a global minimum - Wavelet neuron. Proc. 28 th IEEE Int. Symp. on Multiple-Valued Logic. Fukuoka, Japan: IEEE Corp. Soc., 1998, pp. 335-336.
- [23] T.Yamakawa, E. Uchino, T. Samatu, Wavelet neural networks employing over-complete number of compactly supported non-orthogonal wavelets and their applications. IEEE Int. Conf. on Neural Networks, Orlando, USA, 1994, pp. 1391-1396.
- [24] Ye. Bodyanskiy, N. Lamonova, I. Pliss, O. Vynokurova An adaptive learning algorithm for a wavelet neural network. Expert Systems. 2005. №22(5). pp. 235-240.
- [25] O. Vynokurova, D. Peleshko, M. Peleshko, S. Oskerko, V. Lutsan Multidimensional Wavelet Neuron in Pattern Recognition Tasks for Internet of Things Application. In: Eds. by Z.B. Hu S. Petoukhov, I. Dychka, M. He (2018) Advances in Computer Science for Engineering and Education. Advances in Intelligent Systems and Computing book series (AISC, volume), Springer, Cham (in press).
- [26] A. Cichocki, R. Unbehauen. Neural networks for optimization and signal processing. Stuttgart: Teubner, 1993.
- [27] Y. Meyer Wavelets: Algorithms and Applications - Philadelphia, PA: SIAM., 1999.
- [28] MNIST Homepage <http://yann.lecun.com/exdb/mnist/>, last accessed 19.11.2017

Development of Hybrid Computational Intelligence by Knowledge Genesis Method

Sergej Koryagin
Engineering and Technical Institute
Baltic Federal University of Immanuel
Kant
Kaliningrad, Russia
SKoryagin@kantiana.ru

Pavel Klachek
Engineering and Technical Institute
Baltic Federal University of Immanuel
Kant
Kaliningrad, Russia
pklachek@mail.ru

Irina Liberman
Engineering and Technical Institute
Baltic Federal University of Immanuel
Kant
Kaliningrad, Russia
ILiberman@kantiana.ru

Abstract— The article presents the principles of a promising method of hybrid computational intelligence - "Knowledge Genesis" method intended to create a special class of mathematical models able, on the basis of flexible computing, to integrate accurate, inaccurate and uncertain knowledge within a single system, to change (adapt) their structure in the process of evolution of simulation object, allowing to simulate at a new level complex processes and phenomena including nonlinear ones.

Keywords— hybrid computational intelligence, mathematic simulation, system analysis, military-industrial sphere, innovative circuitry solutions.

I. INTRODUCTION

In the middle of 2000 a number of well-known scientists in various subject domains (military-industrial sphere [3], agro-ecosystems and water ecosystems [4], machine building [5], oil and gas industry [6], socio-economic sphere [6] etc.) defined the need for a fundamental revision of traditional approaches to the construction of mathematical models of complex, including weakly formalized systems. Besides they set a task of creating new directions, approaches and methodologies in the field of mathematical modeling and complex systems synthesis providing at a qualitatively different level the possibility of evolution of mathematical models of the simulation object. In fact, there was talk of creating new classes of mathematical models with a variable cognitive structure [1] as a special class of mathematical models that can change (adapt) their structure in the process of evolution of the observed simulation object. Despite the importance of this task, the works in this area have not received active development. In the year 2012 at the International Conference "Hybrid and Synergistic Intelligent Systems: theory and practice (HSIS-2012)" (Russia, Kaliningrad, the conference was held with support of the Russian Foundation for Fundamental Research) authors of the article for the first time proposed a concept of "soft" mathematical modeling of complex systems, based on heuristic correction of basic analytical dependencies by heuristic knowledge and soft computing [3] which received wide acclaim.

II. NOVELTY

The great theoretical and practical experience in various subject domains (agriculture, oil and gas sector, machine building, military-industrial sector, etc.) [3] accumulated so far in this approach, allowed the authors of the article to

formulate the main provisions (first introduced in this article) of the "knowledge genesis" method, intended to create a special class of mathematical models able, on the basis of flexible computing [3], to integrate accurate, inaccurate and uncertain knowledge within a single system, to change (adapt) their structure in the process of evolution of simulation object, allowing to simulate at a new level complex processes and phenomena including nonlinear ones. According to the authors, the proposed methodological and applied principles of soft mathematical modeling based on the knowledge genesis method may become the basis for creation of promising interdisciplinary areas in the field of hybrid computational intelligence, enabling to create new ways to gain, present and use knowledge and to synthesize advanced and innovative applied tools (for example, new classes of intelligent decision support systems, new generation of circuitry solutions, etc.) in various subject domains.

III. PRINCIPLES OF THE KNOWLEDGE GENESIS METHOD

In Fig. 1. the architectural-technological scheme of the knowledge genesis method is shown. Symbolically the architectural-technological scheme of the knowledge genesis method (Fig. 1) can be represented as follows [3]:

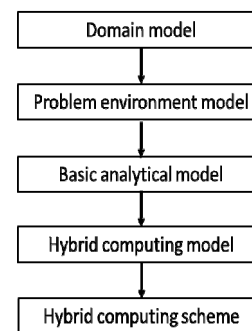


Fig. 1. The architectural-technological scheme of the knowledge genesis method.

$$(E^u, E^L, dm) \xrightarrow{met^u} \dot{m} \quad (1)$$

where: $E^u = \langle E^L, \Pi^u, \Pi^h, \Psi_1, \Psi_2 \rangle$ - is the problem environment model, $E^L = \langle \hat{S}, L^P, \Phi \rangle$ - is the domain model [6], $\hat{S} = \{S_1, S_2, S_3\}$ - is the strata set of problem-system (TABLE 1), L^P - is the language family for describing derived relations, Φ - is the correspondence of the form

$\Phi \subseteq L^P \times \hat{S} \mid \Phi \neq \emptyset$, $\Pi^u = \{\pi_1^u, \dots, \pi_{N_u}^u\}$ - is the decompositions set of problem-system π^u , $\Pi^h = \{\pi_1^h, \dots, \pi_{N_h}^h\}$ - is the subproblems set included in the initial problem-system π^u ($\forall \pi_i^u \exists \Pi^h = \{\pi_1^h, \dots, \pi_{N_h}^h\}$, where $i=1, \dots, N_{\Pi}$, $\forall i (N_h = \text{var } y)$, $\pi_i^u \in \Pi^u$), Ψ_1 - is the correspondence of the form $\Psi_1 \subseteq \Pi^u \times \hat{S}$, Ψ_2 - is the correspondence of the form $\Psi_2 \subseteq \hat{\Pi}^h \times \hat{S} \mid \hat{\Pi}^h = \bigcup_i^{N_{\Pi}} \Pi_i^h$, more than one π^h can belong to $S_i, i=1, \dots, 3$; $dm = \{dm_1^u, \dots, dm_{N_u}^u\}$ - is the basic analytical model with a correspondence of the form $\Psi_3 \subseteq \Pi^u \times dm$; $\dot{m} = \{\Xi^u, met^{\Delta}, R^A\}$ - is the hybrid computing model, Ξ^u - is the multi-valued mapping of the hybrid computational model, obtained on the basis of application of met^{Δ} - is the algorithm of functions of "knowledge genesis" [3], $met^{\Delta} = \{\alpha^1, \dots, \alpha^7\}$ - is a set of functional components (autonomous operators, see Table I), R^A - is the integration relations in set $met^{\Delta} = \{\alpha^1, \dots, \alpha^7\}$ (see the next section).

TABLE I. FUNCTIONAL COMPONENTS

class designation group	restrictions	strata
Analytical Computing α^1	Algebraic and Differential Equations	Parametric S_3
Neurocomputational α^2	Feed-Forward Neural Networks, Kohonen Maps	
Fuzzy Computing α^3	Mamdani, Takagi-Sugeno	
The arguments based on the experience α^4	The reasoning based on precedents	Situational S_1
Evolutionary Computing α^5	Classical genetic algorithms with tournament selection, Pareto GA with niches	Parametric S_3 Situational S_1
Statistical Computing α^6	Monte Carlo method	Streaming S_2
Brain teaser Reasoning α^7	Production expert systems	Situational S_1

IV. THE PRINCIPLES OF FORMAL DESCRIPTION OF THE HYBRID COMPUTATIONAL SCHEME

For a formal description of the hybrid computational scheme (HCS) built on the basis of hybrid computational model \dot{m} , we introduce the concept of "element" - the resource of HCS. The elements set - $RES^{\Delta} \subseteq RES$. The HCS element simulates decision of subtask, of the initial problem-system or performs auxiliary operations. It is constructed in accordance with the functional components

set met^{Δ} and it has properties $PR^{\Delta} \subseteq PR$, the most important of which are the "input" $pr^{\Delta i}$, "output" $pr^{\Delta o}$ and "state" st^{Δ} . To simulate the system behavior at different times, we introduce the concept of delay $z(-k)$, where k - is the delay factor, i.e. $R(x(t), x(t))^{z(-k)} = R(x(t), x(t+k))$. Then the conceptual model of the HCS element res_n^{Δ} , $n = 1, \dots, N$, where N - is the number of elements met^{Δ} and can be represented as follows:

$$\begin{aligned}
\alpha_n^{\Delta} &= R_1^{resmet}(res_n^{\Delta}, met^{\Delta}) \circ \\
&R_1^{respr}(res_n^{\Delta}, pr^{\Delta i}) \circ \\
&\circ R_1^{respr}(res_n^{\Delta}, met^{\Delta o}) \circ \\
&R_1^{resst}(res_n^{\Delta}, met^{\Delta}) \circ \\
&\circ R_1^{stst}(st^{\Delta}(t), st^{\Delta}(t))^{z(-1)} \\
&\circ R_1^{prst}(pr^{\Delta i}(t), st^{\Delta p}(t))^{z(-1)} \\
&\circ R_1^{stpr}(st^{\Delta}(t), pr^{\Delta o}(t))
\end{aligned} \tag{2}$$

where: $(R_1^{stst})^{z(-1)}$ and $(R_1^{prst})^{z(-1)}$ - are the "state-state" and "input-state" relations, (R_1^{stpr}) - is the "state - output" relation. The conceptual model of HCS built on the basis of hybrid computing model \dot{m} :

$$\begin{aligned}
\alpha^u &= R_1^{resmet}(res_A^u; \dot{m}) \circ \\
&R_1^{respr}(res_A^u, pr^{ui}) \circ \\
&\circ R_1^{respr}(res_A^u, pr^{uo}) \circ \\
&R_1^{resst}(res_A^u, st^u) \circ \\
&\circ R_1^{stst}(x^{up}(t), x^{up}(t))^{z(-1)} \\
&\circ R_1^{prst}(pr^{ui}(t), st^{up}(t))^{z(-1)} \\
&\circ R_1^{stpr}(st^{up}(t), pr^{uo}(t))
\end{aligned} \tag{3}$$

where: res_A^u is the HCS-aggregate as a resource of the solution of problem-system π^u ; RES^{Δ} - is a set of at least two elements res^{Δ} , constructed in accordance with the scheme (1); $(R_1^{stst})^{z(-1)}$, $(R_1^{prst})^{z(-1)}$, R_1^{stpr} are the relations of HCS functioning; R_1^{resres} are the relations of elements integration from the set RES^{Δ} ; R_1^{prpr} is the relations between the HCS inputs and the element inputs.

Relations of HCS functioning $(R_1^{stst})^{z(-1)}, (R_1^{prst})^{z(-1)}, R_1^{stpr}$ are not assigned a priori, but are fixed during the aggregate operation.

V. RESULTS OF EXPERIMENTS

A. Mathematical modeling of the launcher complex (anti-aircraft missile system "Buk M1")

The work [8] on example of program storage device (PSD) 9A39M1 of short-range anti-aircraft missile system (AAMS) "Buk-M1" analyses the launch preparation characteristics of launching complex and missiles, and also provides a set of mathematical models that implement inclined launch of semi-active homing AAMS (see figure. 2), taken as the basic analytical models in (1). The Table II lists as an example: the hybrid computing model of calculation of the predicted target impact point (representing the systemic foundation of the entire models set that implement the AAMS inclined launch) developed in compliance with (1), and basic analytical model proposed in [8].

TABLE II. MODELS OF CALCULATION OF THE PREDICTED TARGET IMPACT POINT

Basic analytical model proposed in [8]	Hybrid computational model obtained by the knowledge genesis method through heuristic correction of the basic analytical model
$\Psi_y = \Psi + \Delta\beta_y - 2\pi\text{sign} \Psi ,$ $\varepsilon_y = \varepsilon + \Delta\varepsilon_y,$ $D_{y1} = 0,7t_{y1} + 2\text{Cos}\varepsilon_y - 2,5t_{y1} =$ $= t_{y1}(1 - \frac{B}{D}\text{Cos}\Psi \cdot \text{Cos}\varepsilon),$ $\varepsilon_{1y} = \varepsilon_y + \Delta\varepsilon_{otcy1} + 0,33t_{y1},$ $D_{y1} = 0,7t_{y1} + 10\text{Cos}\varepsilon_{y1} - 10,$ $\varphi_{gy}^* = \Psi_y - \varphi_2 + \Delta\Psi_{otcy},$ $\varphi_{gy} = \varphi_{gy}^* - \pi\text{Sign}\varphi_{gy}^*.$	$\Psi_y = \Psi + \Delta\beta_y - 2\pi\text{sign} \Psi ,$ $\alpha_{bml}^3 = \langle X(t_{y1}, \varepsilon_{1y}, v, \gamma),$ $Y(D_{y1}), KB, F^\mu, F^{TS}, F^y, I^f \rangle,$ $I^f \rangle,$ $\varepsilon_y = \varepsilon + \Delta\varepsilon_y,$ $D_{y1} = t_{y1}(1 - \frac{B}{D}\text{Cos}\Psi \cdot \text{Cos}\varepsilon) \alpha_{bml}^3,$ $\varepsilon_{1y} = \varepsilon_y + \Delta\varepsilon_{otcy1} + 0,33t_{y1},$ $D_{y1} = 0,7t_{y1} + 10\text{Cos}\varepsilon_{y1} - 10,$ $\varphi_{gy}^* = \Psi_y - \varphi_2 + \Delta\Psi_{otcy},$ $\varphi_{gy} = \varphi_{gy}^* - \pi\text{Sign}\varphi_{gy}^*.$

Where: t_{y1} - is the predicted time, ε_{1y} - is the site angle, D_{y1} - is the range, φ_{gy} - is the azimuth of the predicted impact point, Ψ_y - is target angle, ε_y - is the site angle, D_y - is the range to the predicted target impact point, B - is the distance between the TELAR and the PSD, bD - is the range,

$$\alpha_{bml}^3 = \langle X(t_{y1}, \varepsilon_{1y}, v, \gamma), Y(D_{y1}), KB, F^\mu, F^{TS}, F^y, I^f \rangle,$$

$$\alpha_{bml}^3 = \langle X(B, bD, v, \gamma),$$

$$Y(\text{degree}), KB, F^\mu, F^{TS}, F^y, I^f \rangle,$$

are the classical fuzzy systems [3], where: X, Y - are the spaces of input and output of linguistic variables; KB - is the knowledge base as a set of rules sometimes called a

linguistic model; $F^\mu = \{F_x^\mu, F_y^\mu\}$ - is the membership function set of inputs and outputs, respectively; F^{TS} - is the functional relationship set as inferences in the Sugeno algorithm; F^y - is the set of analytical expressions and (or) algorithms of defuzzification; T_r - is the transposition operation, I^f - is the fuzzy system interpreter which can be represented by a set of five processes: $I^f = \langle I^{f1}, I^{f2}, I^{f3}, I^{f4}, I^{f5} \rangle$, where: I^{f1} - is the fuzzification; I^{f2} - is the aggregation; I^{f3} - is the activation; I^{f4} - is the accumulation; I^{f5} - is the defuzzification.

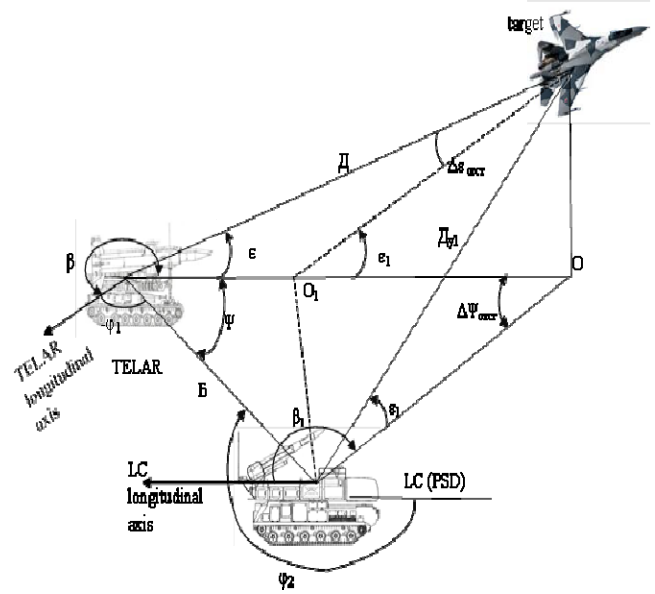


Fig. 2. The scheme of relative position of the transporter-erector-launcher-radar (TELAR), the launch complex (LC) and the target.

To compare the results of computer modeling on the basis of a hybrid computational model and a basic analytical approach (see Table 2), a computer program was developed (Fig. 3) simulating the interception of an aerodynamic target by a missile. The interception is depicted on the chart with the missile release point and the altitude plotted along the axes. The red line below shows the missile trajectory. On the right the gray line shows the target trajectory to the impact point. In the same area, the blue line indicates the missile speed. The violet line shows the estimated time to the impact point. Just above there are 5 more quantities characterizing the missile motion in time. The blue line shows the missile roll. The green and blue lines show the control action value vertically and horizontally, respectively.

A comparative analysis of computer modeling results on the basis of the hybrid computational model and the basic analytical approach, using the example of the modernized anti-aircraft missile system "Buk-M1-2", is shown in Fig. 4.

The results of the comparative analysis indicate broad possibilities of computer simulation based on the hybrid computational model (through heuristic correction of basic analytical models), expanding abilities of classical

mathematical models as well as a range of tasks allotted to short-range and medium-range AAMS.

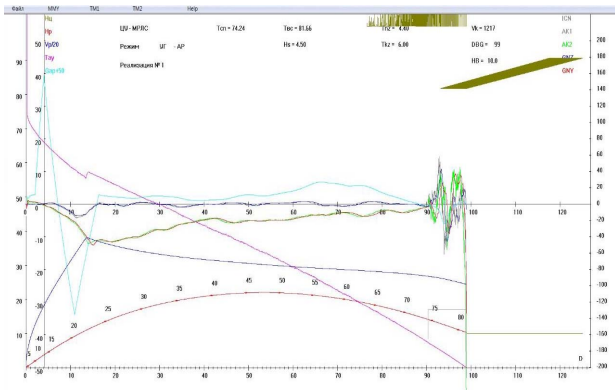


Fig. 3. A computer program simulating the interception of an aerodynamic target by a missile.

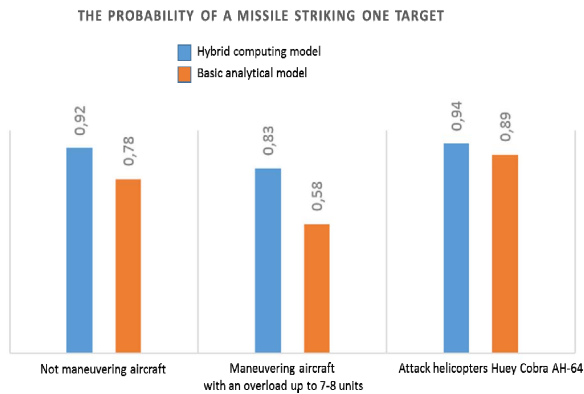


Fig. 4. Results comparison of computer modeling on the basis of the hybrid computational model and the basic analytical approach (Table. 2).

B. Simulation of the aircraft vortex trail, allowing to estimate the safe distance between aircraft.

When flying in the atmosphere an aircraft creates a vortex trail that poses a danger to other aircraft (Fig. 5).

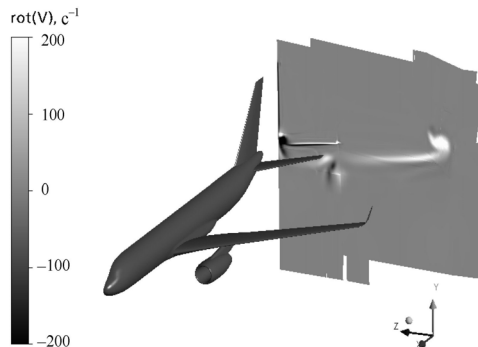


Fig. 5. Aircraft model and vortex trail in the control section $x = 0,5L$.

When taking off and landing, it is the restriction on the vortex trail that determines the safe distance value between aircraft. Reducing this distance increases the airport capacity, but the safety of the flight must be guaranteed. Currently, there are ICAO recommendations (a matrix indicating the safe distance value depending on the classes of the previous and subsequent aircraft), which accumulate all aviation experience. However, an airport controller is often guided by his own experience and not by ICAO recommendations reducing the safe distance value. As

indicated in the summarizing article [9], flight incidents associated with vortex trail encounter often occurred when landing on the controller instructions. When flying at the flight level a similar problem arises. The problem becomes more urgent as the possibility of reducing the distance between the flight levels is being considered now. Therefore, it is very important to have a reliable mathematical model that allows us to estimate the safe distance between aircraft, depending on their types and weather conditions. A unified vortex destruction model in a turbulent atmosphere is presented in [9], which combined vortex destruction model and atmosphere turbulence model and allowed to take into account the effect of weather conditions on the vortex trail lifetime, taken as a basis for creating the hybrid computational model of the aircraft vortex trail based on the model (1). The table 3 lists as an example: the hybrid computational model of aircraft vortex trail (representing the systemic foundation of the entire vortex models set), and NASA basic analytical model proposed in [9].

TABLE III. AIRCRAFT VORTEX TRAIL MODELS

The basic analytical model proposed in [9]	Hybrid computational model obtained by the knowledge genesis method through heuristic correction of the basic analytical model
$V_{\tau} = 1,4 \frac{G}{2\pi r} (1 - \exp(-10(\frac{r_c}{L})^{0,75})) \times (1 - \exp(-1,2527(\frac{r}{r_c})^2)),$ $r > r_c;$ $V_{\tau} = 1,4 \frac{G}{2\pi r} (1 - \exp(-10(\frac{r_c}{L})^{0,75})) \times (1 - \exp(-10(\frac{r}{r_c})^2)),$ $r > \Gamma = \frac{G}{\lambda_{\omega} \rho V_0}.$	$V_{\tau} = 1,4 \frac{G}{2\pi r} (1 - \exp(-10(\frac{r_c}{L})^{0,75})) \times (\exp(\frac{r}{r_c})^{\alpha_{rm1}^3}),$ $r > r_c;$ $V_{\tau} = 1,4 \frac{G}{2\pi r} (1 - \exp(-10(\frac{r_c}{L})^{0,75})) \times (1 - \exp(-10(\frac{r}{r_c})^2)),$ $r > \Gamma = \frac{G}{\lambda_{\omega} \rho V_0}.$

Where: V_{τ} - is the speed tangential component, λ_w - is the vortex range (distance between vortices), Γ - is the vortex circulation, G - is the aircraft weight, L - is the wingspan, r_c , λ_w parameters are determined experimentally from some averaged (by the aircraft type) dependence V_{τ} from the aircraft wingspan (or weight),

$$\alpha_{rm1}^3 = \langle X(V_{\tau}, G, \rho, r_c), Y(\Gamma, \text{degree}), KB, F^{\mu}, F^{TS}, F^y, I^f \rangle$$

- is a classical fuzzy system (see the previous section).

Fig. 6. shows the results of a comparison of the vortex field obtained on the basis of the hybrid computational model and the basic analytical model (see Table. III).

As it is seen from Fig. 6. the hybrid computational model allows to describe the laminarization of the flow and the evolution of the vortex trail more accurately and in more detail, in comparison with the basic analytical model.

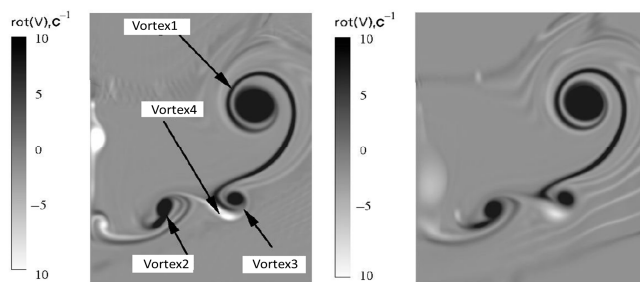


Fig. 6. Comparison of the vortex field obtained on the basis of the hybrid computational model (left) and the basic analytical model (right).

To obtain the integral characteristics of aircraft when it encounters the vortex trail from the preceding aircraft the computations for a number of aircraft-generators and assorted vortex trails were performed. It turned out that the value of the vortex circulation which is dangerous for the second aircraft slightly depends on the radius of the vortex core. Fig. 7 represents a generalized graph of the vortex Γ circulation of the aircraft-trail generator, permissible for the subsequent aircraft depending on its wingspan b . These results are obtained through estimation of the maximum permissible roll moment induced by a vortex from the preceding aircraft to the second aircraft flying along the vortex axis. The vortex field induced by the aircraft-track generator, described with the hybrid computational model (see Table 3), allowed to estimate on a fundamentally new level a permissible interval for ensuring a safe flight of the second aircraft (Fig. 7).

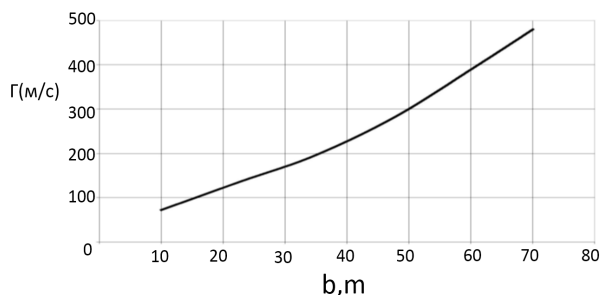


Fig. 7. The maximum allowable circulation of the leader aircraft for safe vortex trail encounter of aircraft with a wingspan b , flying along the vortex axis.

Thus, the results confirm the broad possibilities of computer simulation of aircraft vortex trail based on a hybrid computational model (through heuristic correction of basic analytical models), including applied aspects, allowing to assess more accurately the permissible interval for ensuring a safe flight of the second aircraft.

VI. COMPARISON

The results of the comparative analysis indicate broad possibilities of computer simulation based on the hybrid computational model (through heuristic correction of basic analytical models), expanding abilities of classical mathematical models. Currently the soft mathematical modeling of complex systems based on the knowledge genesis method is successfully applied when solving applied problem complexes in the military-industrial sector, agriculture, engineering, aircraft industry, water ecosystems, etc. [3-7].

VII. CONCLUSION

According to a number of leading scientists in the field of hybrid computing intelligence a new era is beginning for application of artificial intelligence, including computational one, characterizing primarily their applied potential. According to the authors, a place for new promising approaches and successful applications of hybrid computing intelligence technologies can be found in the direction of the development of flexible computing that can integrate accurate, inaccurate and uncertain knowledge within a single system, allowing to simulate at a new level complex processes and phenomena including nonlinear ones. Methodological and applied principles of soft mathematical modeling of complex systems based on the knowledge genesis method are presented for the first time. This forward-looking interdisciplinary approach is designed by the authors on the "border" of: hybrid intelligent systems, synergetic artificial intelligence, hybrid computational intelligence, systems analysis, - enabling to create new ways to gain, present and use knowledge and to synthesize advanced and innovative applied tools (for example, new classes of intelligent decision support systems, new generation of circuitry solutions, etc.) in various subject domains.

REFERENCES

- [1] P. M. Klachek, S. I. Koryagin, and O. A. Lizorkina, "To a new paradigm of mathematical modeling of complex systems," in Conf. Rec. HISIS int. 1 International Conference Hybrid and synergetic intellectual systems: theory and practice, vol 1, pp. 17-27, 2012.
- [2] B.S. Mordukhovich, "Variational Analysis and Generalized Differentiation," Springer, 2005.
- [3] P. M. Klachek, S. I. Koryagin, and O. A. Lizorkina, "Intellectual Systems Engineering, Kaliningrad," Publishing house of the Baltic Federal University of Immanuel Kant, 2015.
- [4] P. M. Klachek, S. I. Koryagin, A. V. Kolesnikov, and E. S. Minkova, "Hybrid adaptive intelligent systems: Theory and development technology," Kaliningrad: Publishing house of the Baltic Federal University of Immanuel Kant, 2011.
- [5] P. Klachek, S. Koryagin, and E. Minkova, "Intellectual evolution of decision support systems in enterprises of the 21st century on the basis of artificial intelligence," in Conf. Rec. RANint. Twelfth Russian RAN Symposium, pp. 189-197, 2011.
- [6] S. Koryagin, P. Klachek, E. Koryagin, and A. Kulakov, "The development of hybrid intelligent systems on the basis of neurophysiological methods and methods of multi-agent systems," IEEE int. First International Conference on Data Stream Mining & Processing, Lviv, Ukraine, pp. 95-102, August 2016.
- [7] S. Koryagin, P. Klachek, and V. Vasileva, "Development Of Bionic Approaches In The Microelectromechanical Systems Design Based On Cognitive Knowledge Bank," IEEE int. 14 International Conference The Experience of designing and application of cad systems in microelectronics, Svalyava, Ukraine, pp. 185-192, 2017.
- [8] A. Skorik, A. Zverev, I. Tkachenko, and R. Varenik, "Mathematical modeling of the launcher complex (anti-aircraft missile system «BukM1»)," collection of scientific works of Kharkiv University Air Forc, vol 3 (43), Kharkiv: Publishing house of the Kharkiv University Air Force named after Ivan Kozhedub, pp. 71-75, 2015.
- [9] V. V. Vyshinsky, A. N. Zamyatin, and G. G. Sudakov. "Theoretical and experimental study of the evolution of the vortex trail behind an aircraft flying in the boundary layer of the atmosphere," Air fleet technique, vol 3. Moscow: Publishing house of the Central Aerohydrodynamic Institute. Professor N.E. Zhukovsky, pp. 25-38, 2008.

Cloud Datacenter Workload Prediction Using Complex-Valued Neural Networks

Igor Aizenberg
Department of Computer Science
Manhattan College
New York, USA
igor.aizenberg@manhattan.edu

Kashifuddin Qazi
Department of Computer Science
Manhattan College
New York, USA
kqazi01@manhattan.edu

Abstract—Cloud computing infrastructures and datacenters depend on intelligent management of underlying CPU, memory, network, and storage resources. A variety of techniques such as load balancing, load consolidation, and remote memory allocation are used to maintain a fine balance between conflicting goals of high performance, and low costs and energy consumption. To meet these goals, successful prediction of the workloads is an important problem. By accurately predicting the resource utilization of host machines, datacenter owners can better manage the available resources. This paper presents a host resource usage prediction approach, based on a Multilayer Neural Network with Multi-Valued Neurons (MLMVN). An enhancement is further implemented to MLMVN to make it suitable for cloud datacenter applications. The approach is evaluated on real world load traces from Google’s cluster data, as well as two grid based load traces. The algorithm is compared against some current state-of-the-art host-load prediction algorithms to show its accuracy, as well as performance gains.

Index Terms—cloud datacenter, workload prediction, complex-valued neural networks

I. INTRODUCTION

Cloud computing provides on-demand computing resources such as CPU, memory, networking, and storage for a variety of big data applications. The infrastructure for implementing clouds is a pool of resources hosted on large clusters of physical machines in datacenters. The sharing of this pool of resources within host machines provides benefits of scale, and results in low costs to use, high performance, energy savings, and elasticity. A variety of dynamic, online techniques such as load balancing, load consolidation, remote memory allocation, memory ballooning, etc. have been implemented to achieve the goals. However, using these techniques without considering the ever-changing resource requirements of the hosts can lead to severe issues. For example, consolidating hosts with large resource requirements can result in resource starvation, which in turn will cause degradation of performance. Similarly, remote memory may allow one host to access large amounts of memory, but may degrade performance on another host that is offering the remote memory. Therefore, it is imperative to accurately predict future host resource usage (host load) in advance, to intelligently make these management decisions. In general, longer (hours), but finer-grained predictions allow better management than shorter (minutes), coarse predictions.

Another concern while predicting cloud workloads, is the computational speed of the solution used. The time spent predicting is overhead performed in real time, and needs to be minimized, while maintaining a high accuracy.

Multi-layered neural networks with multi-valued neurons (MLMVN) are complex-valued neural networks with a derivative-free backpropagation learning algorithm. They have been shown to outperform competing machine learning techniques in a number of problems unrelated to cloud computing ([1], [2], [3]). MLMVNs converge quickly, and can predict multiple steps into the future without a large propagation of previous errors. Additionally, as shown by [4], complex-valued neural networks could be converted to algorithms suitable for quantum computing. The speed-up due to utilizing quantum computers in the prediction phase of datacenters will allow even finer grained prediction across thousands of machines. A preliminary version of this approach has been briefly discussed in [5]. All of this motivates the application of MLMVN to the problem of host load prediction for cloud datacenters.

The main contributions of this paper include a novel cloud host load prediction mechanism based on a modified MLMVN. Second, the paper describes an enhancement to MLMVN that mitigates some of the issues associated with it, making it suitable for host load prediction in a cloud computing environment. The overall approach is rigorously evaluated, and compared against a number of other host load prediction methods. Through all the results, the approach shows state-of-the-art short-term and long-term prediction performance, in terms of both accuracy and computational speed.

The rest of the paper is divided as follows. An overview of current host load prediction techniques is given in Sect. 2. The proposed method is described in Sect. 3. Experiment results and comparisons are presented in Sect. 4. The paper is concluded in Sect. 5.

II. RELATED WORK

According to [6], research in host load prediction has been performed using Bayesian networks, artificial neural networks, decision trees, support vector machines, arima models, and cubic smoothing splines. Further, a number of other mathematical tools have been employed to solve the problem including

Fourier transforms [7], chaos theory [8], and fractal methods [9]. For grid-based systems, [10] utilized the Markov model for single-step predictions while [11] proposed a Kalman filter and autoregressive (AR) based multi-step method. The work in [12] used a feedforward artificial neural network (ANN) and achieved higher accuracy than other methods.

A comparison between Google cluster and grid system host loads shows that Google loads exhibit higher variance and noise, with lower seasonality [13]. Hence, it is more challenging to predict host loads in a cloud than in a grid. It has been shown that classical time series prediction methods such as arima models underperform when used in cloud environments [14].

The authors of [14] proposed a cloud load prediction method based on Bayes model. However, this approach only predicted the mean host load value, and did not aim to make fine-grained predictions. For fine-grained predictions, [15] proposed combining phase space reconstruction with an evolutionary algorithm. This method showed limited capability in multi-step ahead predictions [16]. Moreover, the prediction performance of this method is closely related to the parameters chosen, as the evolutionary algorithm is a stochastic global search method which may get stuck in local optimas [17].

A literature review reveals two methods that currently offer the best accuracy results for fine-grained host load prediction.

The first method used an autoencoder as the pre-recurrent feature layer of an echo state network (ESN) to perform host load predictions multiple steps in the future [17]. However, this method has some limitations. ESN uses a manually chosen leaking rate to control the degree of delays, which reduces its generalization ability when applied to different load traces. Moreover, the random initialization of a large reservoir, could degrade the performance of ESN when applied to noisy loads.

Currently, the approach described by [16] yields the best prediction accuracy for host load time-series prediction. They used the long short-term memory (LSTM) model in a recurrent neural network (RNN). By learning long-term dependencies, this approach demonstrated high accuracy for a large number of timesteps into the future even for noisy cloud host loads. However, RNNs suffer from computationally long training times due to the backpropagation algorithm being applied to recurrent layers. Even with optimizations, feed-forward neural networks are generally faster than RNNs.

As opposed to these approaches, the proposed prediction mechanism utilizes MLMVN which is a feed-forward neural network, utilizing complex-valued neurons. This offers higher functionality, better generalization capability and simplicity of learning. Additionally, MLMVN learning is derivative-free, and avoids falling into local optimas [2]. The method is used to achieve both short-term and long-term predictions.

III. BACKGROUND

This section discusses the core details of MLMVN (input/output structure, neuron structure, activation function, error-correction rule, and backpropagation algorithm) to reproduce the results of this paper. Interested readers are encouraged to refer to [2] for a more thorough understanding of MLMVN.

A. Multi-Valued Neurons and MLMVN

The main distinction of MLMVN as compared to the classical feedforward neural network, is that its building blocks are Multi-Valued Neurons (MVN) with complex-valued weights. Using complex-valued inputs/outputs, weights and activation functions, it is possible to increase the functionality of a single neuron and a neural network, to improve their performance, and to reduce the training time ([3], [4]).

MVN was initially introduced as a discrete MVN in [2]. A continuous MVN was then introduced in [18]. This paper employs a continuous MVN. It implements a mapping between n inputs and a single output. All real-valued inputs (x_r) need to be in the range [0.0, 1.0] and are initially transformed to complex-valued inputs (x). The complex-valued outputs (y) are transformed back to real values (y_r) at the end.

$$x = e^{ix_r}$$

$$y_r = \arctan2(y_{imag}, y_{real})$$

While MVNs inputs and output are complex numbers located on the unit circle, its weights are arbitrary complex numbers. An input/output mapping of a continuous MVN is described by a function of n variables

$$f(x_1, \dots, x_n) = P(w_0 + w_1x_1 + \dots + w_nx_n)$$

where $x_1, \dots, x_n (x_j \in E_k, j = 1, \dots, n)$ are neuron inputs and w_0, w_1, \dots, w_n are the weights. P is the activation function given by

$$P(z) = e^{iArg(z)} = z/|z|$$

where $z = w_0 + w_1x_1 + \dots + w_nx_n$ is the weighted sum. $Arg(z)$ is the main value of the argument of the complex number z . Thus a continuous MVN output is a projection of its complex-valued weighted sum onto the unit circle.

The MVN learning for hidden neurons is based on the error-correction learning rule as described in [2], [18].

Using the described MVNs, the MLMVN is constructed. The MLMVN backpropagation learning algorithm is derivative-free and it is based on the generalization of the error-correction learning algorithm for a single MVN. This algorithm was proposed in [2] where it is described in detail. The batch version of this algorithm, used for this paper was proposed in [19].

B. Time Series Forecasting

Previously MLMVN has been considered for time series prediction of oil well production [1]. Time series forecasting can be formulated as a classification problem, with n past values of the time series (host loads in this case) as inputs and the $n + 1$ st value as the output. Presumably, there exists some functional dependence among the series members, according to which the $n + 1$ st member's value is a function of a certain number of preceding n members' values.

$$x_{n+1} = f(x_0, \dots, x_n)$$

TABLE I
PARAMETERS FOR MLMVN AND LSTM-RNN

Network Parameters	MLMVN	LSTM-RNN
No. of Inputs	24	24
No. of neurons in Hidden layer 1	32	128
No. of neurons in Hidden layer 2	64	-
Learning rate	1/(no.(weights))	0.05
Iterations	150	90

Suppose the historical data of host load usage is the time series x_0, x_1, \dots, x_r . A training set is formed from this time series by using overlapping subsets of $n + 1$ values. MLMVN is trained on the training set, using n values as input and the $n+1$ st value as the desired output in each subset. The network weights learned through training can be thought of as an approximation of the function f . Therefore, using this approximate \hat{f} , future values of host loads can be predicted as follows:

$$\hat{x}_{r+1} = \hat{f}(x_{r-n+1}, \dots, x_r)$$

$$\hat{x}_{r+2} = \hat{f}(x_{r-n+2}, \dots, x_r, \hat{x}_{r+1})$$

where \hat{x} is the predicted value of x . Multi-step ahead prediction is achieved by performing repeated one-step ahead prediction. At each step the predicted value is used as part of the next step's input. Thus, as many points as required in the future (called prediction window) can be predicted.

C. Minimizing Dependency on Initial Weights

A potential drawback of using MLMVN, is its high dependence on the initial random weights chosen. This results in the quality of network training varying heavily across different runs for the same dataset. To mitigate this problem, a batch learning algorithm based on complex QR decomposition was introduced for MLMVN in [20]. In [19], a linear least squares (LLS) based batch algorithm was proposed for a complex-valued neural network with a *single* hidden layer. The algorithm proposed adding adjustment factors to all the weights after each iteration of training. The algorithm works as follows. In each iteration of training 1) Calculate the errors for all training samples using current weights. 2) Create an overdetermined system of linear algebraic equations for 'adjustment factors'. This system is given by the following

$$\Delta w_0^h + \Delta w_1^h x_1^j + \dots + \Delta w_n^h x_n^j = \delta_j^h$$

Where, for the h th neuron, and the j th training sample, Δw_i^h is the weight adjustment factor for the neuron's i th weight, x_i^j is the i th input, and δ_j^h is the calculated error for the neuron. 3) Solve this system of equations for adjustment factors using LLS. 4) Adjust weights of all neurons in a layer simultaneously by adding the adjustment factors to the corresponding weights.

This resulted in faster learning, ability to maintain big learning sets, and improved generalization capability. For this paper, the algorithm is generalized and implemented for MLMVN with multiple hidden layers. For the rest of the paper, MLMVN refers to the modified MLMVN.

IV. EXPERIMENTAL EVALUATION

A. Experimental Setup

To evaluate the accuracy of the MLMVN based host load prediction algorithm, experiments were performed on two types of datasets. The cloud-based, noisy data from Google clusters, and the more regular grid data from sahara and themis datasets. The MLMVN model was built in Matlab. A network with two hidden layers was chosen based on observations from 100 host loads outside of the ones used for the experimental evaluations. The choice of two hidden layers is intuitive for time series predictions, since using only a single layer acts as a powerful low pass filter, and averages the output, ignoring local changes of the series to be predicted. The two layers consisted of 32 and 64 neurons respectively. Similar to previous methods, the number of inputs (preceding values considered for prediction) was chosen to be 24.

For all the datasets, MLMVN is compared against LSTM based RNN, and ESN. The two frameworks were built in Python using the TensorFlow library, to the specifications in [16] and [17]. The MLMVN neural network parameters are listed in Table I. For comparison, LSTM-RNN model's optimal parameters are also shown. To compare the accuracy of the various methods, the Mean Square Error is calculated as

$$MSE = 1/N \sum_{i=1}^N (y_i - p_i)^2$$

where N , y_i , and p_i are the length of the prediction window, actual values, and predicted values respectively.

B. Google Cluster Data

The Google cluster dataset [21] traces approximately 670000 jobs, and 40 million tasks across 12000 host machines during 29 days. The trace includes a variety of parameters, including CPU and memory usage of the tasks, as well as the location of the tasks on host machines. For this evaluation, the CPU usage of individual host machines for the 29 days were utilized to compare all methods. These values were obtained by aggregating the CPU usages of all the tasks residing on a host (located under the task_usage directory) during a time sample. The cluster data provides information every 5 minutes. Thus, the 29 day load trace for each host consists of 8352 data points. Since the Google trace provides CPU usage as a fraction of utilization in the range [0.0, 1.0], no further scaling is performed for these evaluations. For reference, the host load of one machine is shown in Fig. 1. The lack of any obvious pattern should be clear from this figure.

Similar to the other methods being compared, the first 26 days were used as training/ validation sets, while the last 3 days (day 27 to day 29) were used for testing. The results that follow indicate predictions for the testing set.

To demonstrate the efficacy of MLMVN based load prediction, predictions were performed on 4000 host machines from the Google cluster. These multi-step predictions were

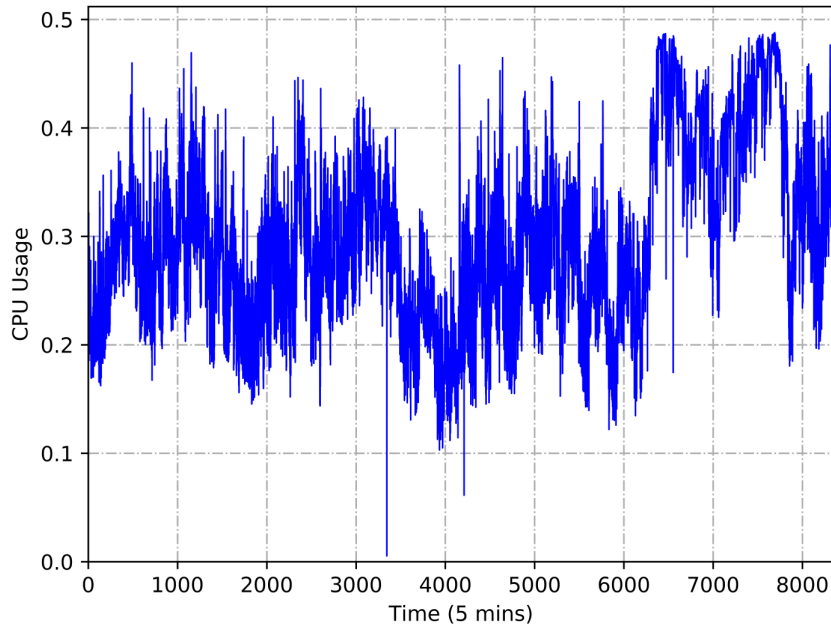


Fig. 1. Load trace for sample Google host - 29 days

done for prediction windows of 30 minutes (6 steps), 1 hour (12 steps), 1.5 hours (18 steps), 2 hours (24 steps), 2.5 hours (30 steps), and 3 hours (36 steps). Fig. 2 shows the MSE achieved by MLMVN, LSTM-RNN, and ESN for each of these prediction window sizes. It can be observed that MLMVN outperforms the other two prediction methods, and maintains its better accuracy even as the prediction window size is increased. Specifically for predictions 30 minutes into the future, MLMVN shows an MSE of 0.0032.

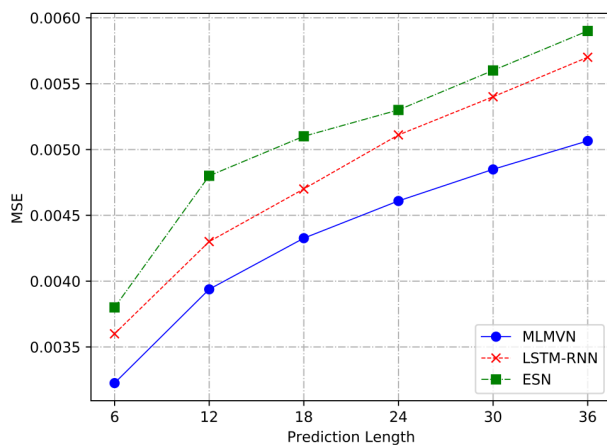


Fig. 2. Average MSE comparison for different prediction windows

For additional comparisons on the Google cluster data, the Cumulative Distribution Functions (CDFs) of MSEs for two

prediction windows (30 minutes and 1 hour) are presented in Fig. 3. For reference, results of real-valued feed-forward neural networks (ANN) and Hybrid-Autoregression (AR) built according to [12] and [11] are also shown. It can be deduced, that MLMVN outperforms all other methods, for all prediction window sizes. Methods such as the hybrid AR have a large variance in their accuracy, as opposed to MLMVN which shows consistent accuracy for most of the Google cluster hosts. For 30 min predictions, about 42% of predictions made by the hybrid AR show an MSE greater than 0.005, compared to approximately 18% for MLMVN. The benefits are even more apparent in the 60 min predictions.

C. Grid Data

In order to evaluate MLMVN based predictions on a different type of load trace, experiments were also performed on the grid-based loads provided by [22]. Specifically, two traces were chosen, namely the sahara and themis traces. These data traces include four days worth of data, sampled at one second each, for a total of 345600 data points each. The sahara dataset belongs to a compute server, while the themis dataset belongs to a desktop. The load is the number of processes that are running or are ready to run (the length of the scheduler's ready queue). For the original trace, the kernel sampled the length of the ready queue at a fixed rate, and averaged a window of previous samples to produce a load average.

For evaluation in this paper, the load traces were scaled to a range of [0.1, 0.9]. Similar to other methods being compared, each load trace was normalized using

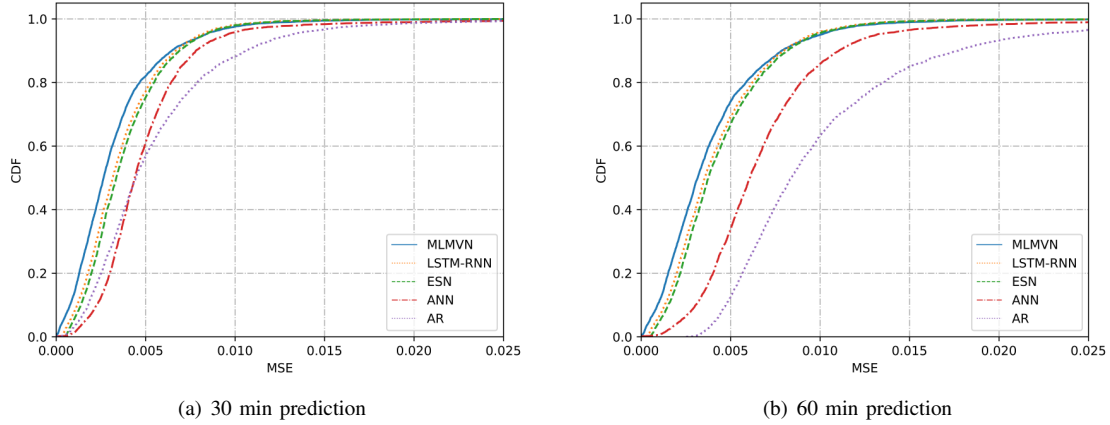
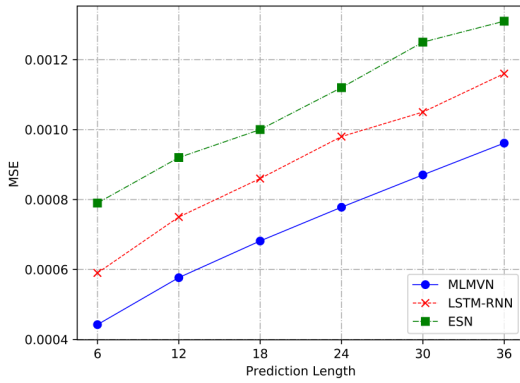
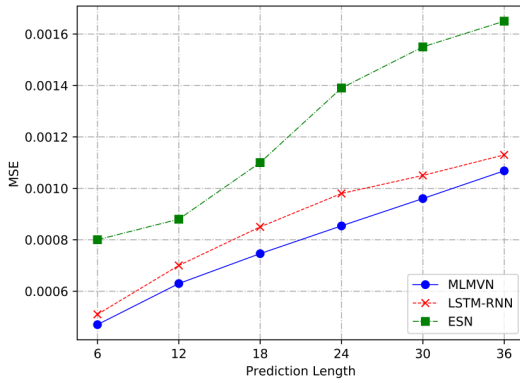


Fig. 3. CDFs of prediction MSEs for different methods



(a) sahara



(b) themis

Fig. 5. Average MSE of prediction in two grid load traces.

$$x_i = 0.1 + \frac{x_i - x_{min}}{x_{max} - x_{min}} (0.8)$$

where x_{max} and x_{min} are the maximum and minimum

value of the load trace, respectively. For reference, a sample of the sahara trace is shown in Fig. 4. It can be observed, that compared to the Google cluster data trace, this trace is more regular, and has a visible pattern to it.

To compare accurately with the other state of the art methods, each load trace was split 80% of its length into a training set and the rest was the testing set. The prediction results are shown in Fig. 5. The results for LSTM-RNN and ESN are as reported by [16]. According to the results, MLMVN shows higher accuracy as compared to both LSTM-RNN and ESN, for all the prediction lengths.

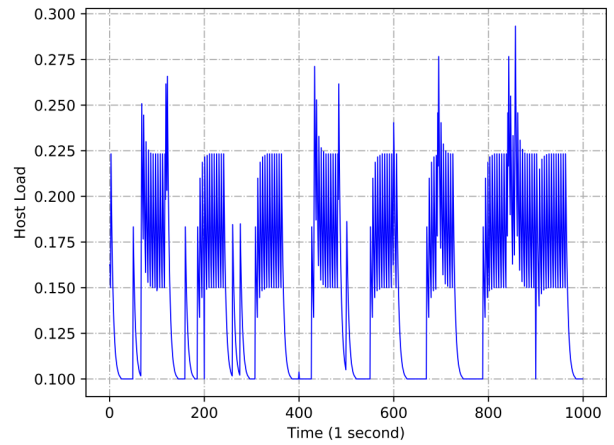


Fig. 4. Load trace sample from grid-based sahara data

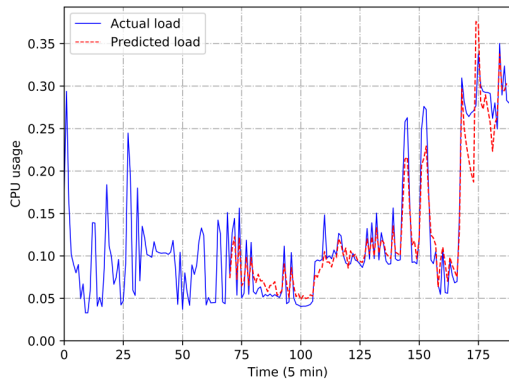
The accuracy gains of MLMVN vs LSTM-RNN are least prominent in the themis dataset, compared to other datasets. The reason is that of all the loads evaluated, the themis dataset appears to be most regular, and predictable. Thus, most sophisticated algorithms achieve satisfactory performance. In spite of this, there is a visible improvement in accuracy with MLMVN across all prediction windows.

TABLE II
TIME TO TRAIN VS ACCURACY

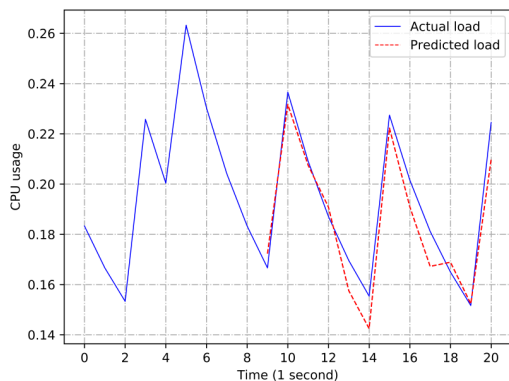
Method	Time to train (s)	Avg. 6 step MSE
MLMVN	38	0.0032
LSTM-RNN	105	0.0036
ANN	35	0.0052

D. Prediction Samples

In order to identify the difference in quality of MLMVN predictions between the cloud and grid loads, Fig. 6 presents 30 minute predictions in a sample from the Google cluster data, as well as the sahara data set. It can be observed that in the more regular sahara dataset, the predictions follow the actual loads closely. On the other hand, as expected, in the noisy Google cluster sample, the predictions follow the actual loads, albeit not as closely as the sahara sample.



(a) Google cluster sample



(b) sahara sample

Fig. 6. Actual loads vs Predicted loads

E. Time to Train Analysis

As previously noted, RNNs tend to train slower than similar Feed Forward Neural Networks. The implementations used for evaluation in this paper were executed on Ubuntu 16.04

machines with 8GB of RAM. The avg. training time for a single host in the Google cluster data for MLMVN was 38s versus LSTM-RNN's 105s. Table II showcases the results for 6 step predictions. MLMVN offers the most beneficial balance of time to train and accuracy. It should be noted that the MLMVN implementation is on Matlab, which should have a general disadvantage in computational speeds versus Python. Moreover, LSTM-RNN needs to train separately for each prediction window size needed (6 steps, 12 steps, etc). MLMVN uses iterative prediction, thus after training only once, as many future steps as required can be predicted.

F. MLMVN vs Vanilla-MLMVN

To compare the difference between the MLMVN (with LLS algorithm) used in these evaluations and vanilla-MLMVN (without LLS algorithm), the networks were trained on the same host load 10 times. The standard deviation of MSEs for MLMVN was 0.000067 as opposed to vanilla-MLMVN's 0.007. The best MSE in each case was 0.001701. Thus, while both methods achieve similar MSEs, the vanilla-MLMVN needs to be trained multiple times to obtain best results. In contrast, MLMVN trains consistently, making it suitable for online, real-time predictions in cloud computing environments.

V. CONCLUSION

This paper proposed an approach to predict host resource usage in cloud environments using complex-valued neural networks (MLMVN). The real-time prediction of host loads could be utilized for a variety of datacenter management concerns such as load balancing, load consolidation, remote memory allocation, etc. Through extensive experimental analysis, it was demonstrated that the proposed prediction solution produces state-of-the-art accuracy for real world Google cluster and grid load traces. Additionally, an analysis of the computational time revealed its superiority compared to other high-accuracy approaches that utilize recurrent neural networks. The complex-valued network also has the potential to be transformed into a quantum computing algorithm, which could offer even greater speed benefits in the future.

REFERENCES

- [1] I. Aizenberg, L. Sheremetov, L. Villa-Vargas, and J. Martinez-Muñoz, "Multilayer neural network with multi-valued neurons in time series forecasting of oil production," *Neurocomputing*, vol. 175, pp. 980–989, 2016.
- [2] I. Aizenberg and C. Moraga, "Multilayer feedforward neural network based on multi-valued neurons (MLMVN) and a backpropagation learning algorithm," *Soft Computing*, vol. 11, no. 2, pp. 169–183, 2007.
- [3] N. N. Aizenberg and I. N. Aizenberg, "CNN based on multi-valued neuron as a model of associative memory for grey scale images," in *Cellular Neural Networks and their Applications, 1992. CNNA-92 Proceedings., Second International Workshop on.* IEEE, 1992, pp. 36–41.
- [4] A. Hirose, *Complex-valued neural networks.* Springer Science & Business Media, 2012, vol. 400.
- [5] K. Qazi and I. Aizenberg, "Towards quantum computing algorithms for datacenter workload predictions," in *Cloud Computing (CLOUD), 2018 IEEE International Conference on.* IEEE, 2018, p. In Press.
- [6] N. Herbst, A. Amin, A. Andrzejak, L. Grunské, S. Kounev, O. J. Mengshoel, and P. Sundararajan, "Online workload forecasting," in *Self-Aware Computing Systems.* Springer, 2017, pp. 529–553.

- [7] Z. Gong, X. Gu, and J. Wilkes, "Press: Predictive elastic resource scaling for cloud systems," in *Network and Service Management (CNSM), 2010 International Conference on*. IEEE, 2010, pp. 9–16.
- [8] K. Qazi, Y. Li, and A. Sohn, "Workload prediction of virtual machines for harnessing data center resources," in *Cloud Computing (CLOUD), 2014 IEEE 7th International Conference on*. IEEE, 2014, pp. 522–529.
- [9] M. Ghorbani, Y. Wang, Y. Xue, M. Pedram, and P. Bogdan, "Prediction and control of bursty cloud workloads: a fractal framework," in *Proceedings of the 2014 International Conference on Hardware/Software Codesign and System Synthesis*. ACM, 2014, p. 12.
- [10] S. Akioka and Y. Muraoka, "Extended forecast of cpu and network load on computational grid," in *Cluster Computing and the Grid, 2004. CCGrid 2004. IEEE International Symposium on*. IEEE, 2004, pp. 765–772.
- [11] Y. Wu, Y. Yuan, G. Yang, and W. Zheng, "Load prediction using hybrid model for computational grid," in *Grid Computing, 2007 8th IEEE/ACM International Conference on*. IEEE, 2007, pp. 235–242.
- [12] T. V. T. Duy, Y. Sato, and Y. Inoguchi, "Improving accuracy of host load predictions on computational grids by artificial neural networks," *International Journal of Parallel, Emergent and Distributed Systems*, vol. 26, no. 4, pp. 275–290, 2011.
- [13] S. Di, D. Kondo, and W. Cirne, "Characterization and comparison of cloud versus grid workloads," in *Cluster Computing (CLUSTER), 2012 IEEE International Conference on*. IEEE, 2012, pp. 230–238.
- [14] —, "Host load prediction in a google compute cloud with a bayesian model," in *High Performance Computing, Networking, Storage and Analysis (SC), 2012 International Conference for*. IEEE, 2012, pp. 1–11.
- [15] Q. Yang, C. Peng, H. Zhao, Y. Yu, Y. Zhou, Z. Wang, and S. Du, "A new method based on psr and ea-gmdh for host load prediction in cloud computing system," *The Journal of Supercomputing*, vol. 68, no. 3, pp. 1402–1417, 2014.
- [16] B. Song, Y. Yu, Y. Zhou, Z. Wang, and S. Du, "Host load prediction with long short-term memory in cloud computing," *The Journal of Supercomputing*, pp. 1–15, 2017.
- [17] Q. Yang, Y. Zhou, Y. Yu, J. Yuan, X. Xing, and S. Du, "Multi-step-ahead host load prediction using autoencoder and echo state networks in cloud computing," *The Journal of Supercomputing*, vol. 71, no. 8, pp. 3037–3053, 2015.
- [18] I. Aizenberg, C. Moraga, and D. Paliy, "A feedforward neural network based on multi-valued neurons," in *Computational Intelligence, Theory and Applications*. Springer, 2005, pp. 599–612.
- [19] E. Aizenberg and I. Aizenberg, "Batch LLS-based learning algorithm for MLMVN with soft margins," in *Proceedings of the 2014 IEEE Symposium Series of Computational Intelligence (SSCI-2014)*. IEEE, 2014, pp. 48–55.
- [20] I. Aizenberg, A. Luchetta, and S. Manetti, "A modified learning algorithm for the multilayer neural network with multi-valued neurons based on the complex QR decomposition," *Soft Computing*, vol. 16, no. 4, pp. 563–575, 2012.
- [21] J. Wilkes, "More Google cluster data," Google research blog, Nov. 2011, posted at <http://googleresearch.blogspot.com/2011/11/more-google-cluster-data.html>.
- [22] P. A. Dinda, "The statistical properties of host load," *Scientific Programming*, vol. 7, no. 3-4, pp. 211–229, 1999.

Computer System of Building of the Semantic Model of the Document

O.S. Volkovsky

*Department of Computer Science and Information Technologies,
Oles Honchar Dnipro National University
Dnipro, Ukraine
didivave@mail.ru.*

Y. R. Kovylin

*Department of Computer Science and Information Technologies,
Oles Honchar Dnipro National University
Dnipro, Ukraine
kovilin.yegor@gmail.com.*

Abstract – The analysis of the existing approaches to the development of applied models of natural language was performed. The algorithms of building of semantic representation of the text were developed. The structure of the computer system of building of the semantic program model of the document was examined. The application of the developed system for automatic determination of coherence of the text in the natural language was described.

Key words — *semantic network, inquiry/response system, automated text processing*

I. INTRODUCTION

The development of applied program systems of automatic texts processing means the selection of one or another way of description and implementation of the model of natural language available to ECM. Due to the fact that the language is rather non-formalized system with instability and non-uniformity of its own rules then the main problem is a difficulty of description of semantic characteristics of the text at the level of algorithmic representation. As long as the natural language is not just a set of words based on some grammar constituents (the top-priority focus of automated texts processing tasks is the obtainment of the very sense bearing text) this, consequently, makes many developers to consider the semantic relations not only between certain words but also between sentences and even between the documents. In most cases the semantics and understanding of the text by the machine mean the following: we cannot speak about the semantics in the event of entering of the text into ECM memory and printing of it with the use of printing device; however, we can speak about the semantic understanding of the text by the computer if the text is processed in one or another way and, as a result of such processing, the reader obtains a new text that is understandable and adequate for him or her (for example, the translation into other language). From this perspective we can speak with confidence about the importance of creation of automated approach to the program simulation of the natural language. The present paper deals with the structure of the system of automatic building of computer model of obtainment of semantic relations in the text that is available for the further applied program implementation of systems of automated text processing.

II. THE ANALYSIS OF EXISTING COMPUTER MODELS OF TEXTS

There are several classes of systems of text processing in present-day computer linguistics; they differ against each other both in complexity of data processing and complexity

of the intelligent component. Conditionally these systems can be divided into three types according to language models forming their basis: Chomsky generative grammar, semantic network and tools of neuronal networks. Due to the fact that the necessity of applied implementation in this case ranks above the theoretic element, let's consider the certain program systems of every specified class. As a first step let's consider the program of generation of test assignments for extra mural study of students based on the paradigms of Chomsky formal grammar [1]. Generative constituent grammar is based on the axiom of the existence of the phenomenon of linguistic competence consisting in ability of human to master and to understand natural human language. According to this the generative grammar establishes a goal to simulate this ability within generation of correct sentences using certain finite set of rules, alphabet and initial symbol of the sentence – the immediate constituents. Theoretically, plenty of immediate constituents is unrestricted and infinite; in practice the very language, subject field, working text corpus and possibilities of ECM essentially reduce the size of the plenty of immediate constituents.

The technology of semantic networks that is the next step in the development of texts processing industry came into widespread acceptance in the field of automatic texts processing. The semantic network is a graph the points of which comprise the semantic units, and the arcs of which describe the notional relations between them. Typically, the semantic units mean the single word or the sentence or even the entire document. The practical application of the semantic network to the task of text processing is well represented in the paper [2] – the system of automatic counseling. Developers focus on generation of base of knowledge of certain subject field for the provision of dialogue with the user according to corresponding questions. There is provided to use the semantic network for storage of extracted knowledge on the basis of training corpus represented by the sets of pre-readied boilerplate phrases-answers.

Currently the most automated tools applicable to the task of automatic generation of texts are the applied artificial intelligence methods – implementation of automated texts processing with the use of neuronal networks. The artificial neuronal networks are in general use today for the purpose of handling of various applied problems, including the tasks of automatic language processing. For the purpose of assessment of quality of neuronal networks in the task of automatic generation of texts let's consider the paper [7] where the recurrent network is used for drafting the description of products in some on-line store. As we can see,

the results are rather mixed. The main advantage of this approach is a complete automation of the process of text generation, high degree of system adaptivity and low costs for its setting and implementation. However, some problems of notional rubbish production are evident. The reason is that in spite of the false availability of the intelligent processing, the system does not see the meaning of described and generated text, basing exclusively on pre-given patterns – teachers.

In the event of building of program system of automatic texts processing with intelligent semantic components the notional relations between the elements of the text are particularly important characteristics of the text, so the semantic networks are the best choice for description of model of natural language on the basis of which the assessment and the relationship extraction is carried out. Therefore, it opens a question of building of alternative ways of necessity of compilation of model base of knowledge on the basis of which the semantic network will be formed. The approach used in the paper [2] forms the closed system, the results of work of which do not exceed the limits of the base of knowledge added to it, whereas the most important parameters of the task of automatic texts processing is the adaptivity and the generality of applied use of development.

The creation of the semantic net of a text is not a new task. At this time, there are several approaches to the computer processing of the semantic nets for both Slavic and English languages. The basis of all these approaches, which form the basic relations between elements in the text is the ontology production model [3], an example of which is illustrated in formula (1):

$$Q_i = P_i(A, B) \quad (1)$$

where i is the product's Q name, with which the product stands out from the all working set (as a name can be used a running number from the set of products, which is stored in the system's memory), $P_i(A, B)$ – a predicate of relations, which describes how the element of language A is affected by B .

For example, the word "burn" can be described as (fire, action). The practical application of this technology is described detailed in the work of [3], on the basis of which it is created a semantic meta-description of the test document for the future semantic search. The meta description is defined as the triplets, which contain the sentences of the original text. The key feature within the frame of our work is that the basic system data is formed on the basis of the previously manually marked body of the Russian language.

The increasing scope of the semantic net led to the creation of the net formation approach as a net model, described in [4].

In this case, the predicative relation is described by formula (2):

$$Q = P(I, C_i...C_n) \quad (2)$$

where I – an information units set, $C_i...C_n$ - a set of the links types between information units. Such nets are often used as the documents search models in the body, as it is well suited

for the links description of the set of texts against each other. The applied application of the net models within the task of the document's semantic models formation is still in question.

The further development of the semantic nets technology received in work of [5]. The suggested semantic Q-net has a pyramidal structure and, therefore, all text parts, reflecting the essential units of the subject area or integrated complex objects, for detection of which the special relations were introduced, will always be reflected in this net by the corresponding vertices. Each network pyramid defines a certain text fragment of one of four types. Moreover, Q-nets have the properties of homogeneity and hierarchy, allowing the formation of relationships between semantic objects. It is expected in future that by representing with the help of one Q-net the texts selection of this subject area and using the mechanisms for formation of the generalized objects class definitions and relations in the pyramidal nets, it will be possible to automate the process of the ontology construction of this subject area.

An interesting practical development with the use of semantic nets is the forming system of a semantic net from the weakly structured text sources, described in the work of [6]. The authors of the work offer an approach for the automatic recovery of the article's sections structure of the open dictionary Wiktionary. The peculiarity of this approach is the development of a certain rules system, on the basis of which a semantic program model of the article is created.

III. NOVELTY OF RESEARCH AND COMPARISON OF EXISTING APPROACHES

Most of the applied developments of the computer systems with the use of the semantic nets suppose the use as the starting knowledge basis some block of texts, which contain a previous linguistic annotation. In such a way, it was described in the work of [3] a system, which was initially based on the articles of the Russian national corpus, which is not only closed for the public use, but also contains the markings solely based on Russian-language materials. An alternative for the automated text processing of the other flexional rich languages, as for example Ukrainian, doesn't exist at this moment. The further improvements of the semantic nets, as in the works of [4] and [5] touched upon a question of modification of the net structure itself, and not of the automation methods for the formation and processing of the original system data and it did not find the applied application within the frame of our task. The alternative approach for the net formation is the use of some rules system, as was described in the work of [6]. Such approach allows avoiding of a previous necessity of the linguistic text annotation. However, the use of such method for the natural language is limited, as due to the lack of enough formalization, high flexion, a large number of exceptions and the properties of language variability, it is not possible at the moment to create and effectively to process such a set of rules at the applied software level.

The main task of the developed within the frames of this work approach to the construction of a semantic net is overcoming of the necessity of the previous receiving of any linguistic knowledge. The described technique allows to build semantic relations between the elements of the document and to put them in line with the numerical semantic weight, forming, in such a way, a program model

of the document. Herewith, it is used in the knowledge base system neither linguistic annotation of the document, as in the work of [3-5], nor the system of language rules, as in the work of [6], that allows the receiving of the program semantic models of the documents with the high level of adaptability and independence from language.

IV. THE SYSTEM OF BUILDING OF SEMANTIC MODEL OF THE DOCUMENT

The algorithm of system's work is represented in the Fig. 1. The first stage after document download and syntactic analysis performance (detection of sentences and words) is a determination of the part of speech for every found word. For this purpose the training sample collection from 15 thousand of words and parts of speech corresponding to them was included into the system. The training of naive Bayes classifier is carried out for every element on the basis of three types of flexions (two and three last letters of the word and flexion obtained through Porter's stemming algorithm), afterwards the establishment of part of speech for every selected word in the text is carried out on the basis of trained model. The final dictionaries of auxiliary parts of speech were included into system for the purpose of accuracy increase.

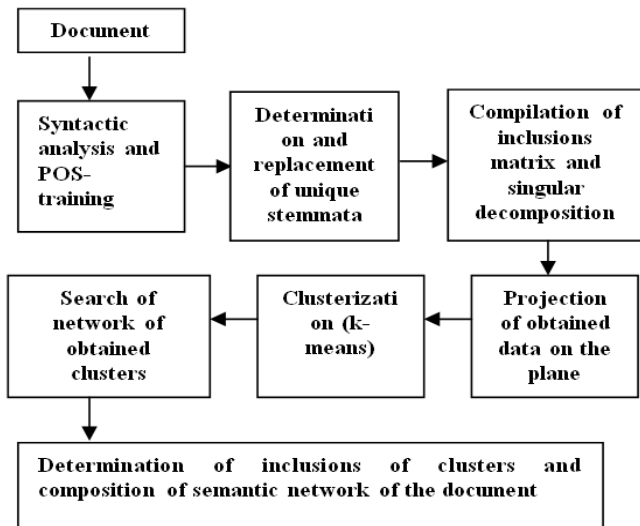


Fig. 1. General structure of work of system of building of semantic network of the text

The determination of unique stemmata is carried out after this stage: the clipping of prefix and flexion is conducted for every pair of words by Porter algorithm – if the length of the maximum common part exceeds or is equal to Levenshtein distance for this pair of words then the analyzed word is replaced by obtained stemma. As a result the text has the following form (Table I) – the system determined the part of speech and number of inclusions of stemma into text for every stemma. All words marked as auxiliary parts of speech are removed from the text at this stage.

The matrix $N*M$ the values of which are determined by quantity of inclusions of stemmata into the sentence is composed for every stemma (with a total quantity of N) and every sentence (with a total quantity of M). The operation of singular value decomposition and projection of obtained data on the plane is conducted on the obtained matrix.

TABLE I. THE RESULT OF SYNTAX PROCESSING

Text before being processed	Text after processing
Today there are various books, video courses on software engineering and other ways to learn to create programs and applications quickly and relatively inexpensively.	2[[day{ADV}]]3[[there a{V}]]2[[vari{A}]]2[[book{S}]]3[[videocours{S}]]22[[program{S}]]4[[ot he{S}]]7[[way{S}]]2[[quickl{ADV}]]2[[inexpensive{A}]]3[[to learn{V}]]8[[to crea{V}]]22[[program{S}]]2[[application{S}]]

Due to the fact that the singular value decomposition is stable, we can omit those values of left and right matrix that correspond to low singular values, and to keep only the first two that represent the vectors of coordinates of two-dimensional plane for stemmata and sentences. The obtained projection is represented in the Fig. 2.

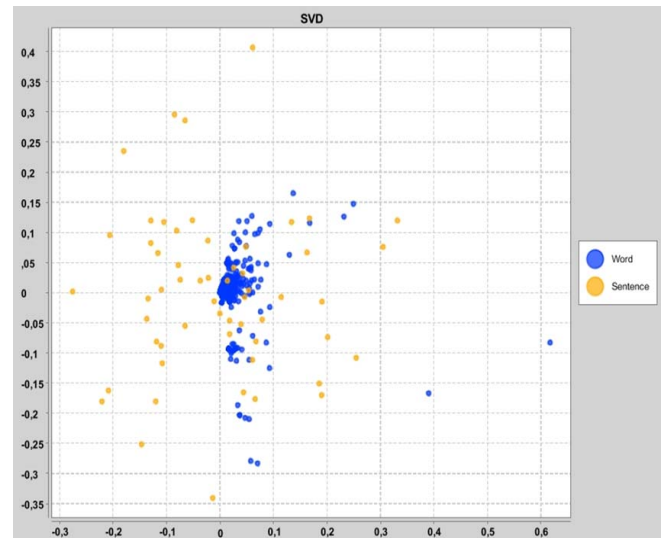


Fig. 2. Projection of singular decomposition: Word – coordinates-stemmata, Sentence – coordinates-sentences of text network

The next step is a clusterization of points-coordinates for stemmata and sentences by *k-means* algorithm. Quantity of clusters for stemmata and sentences cl is determined according to the formula (3):

$$cl(W, W_U) = \frac{count(W)}{count(W_U)} \quad (3)$$

where W means the words, W_U means stemmata. Centroids of clusters – stemmata are the coordinates of stemmata with the maximum number of inclusions into the text determined according to the formula (4):

$$Cst(W_U) = \max(W_0...W_{cl}) \quad (4)$$

where $W_0...W_{cl}$ are the weights of stemmata. Centroids of clusters-sentences are the coordinates of sentences with the maximum total weight of stemmata determined according to the formula (5):

$$Cs(W_S) = \max\left(\sum_{i=0}^{SN} W_i\right) \quad (5)$$

where W_S is a sentence, W_i is a weight of stemma in the sentence, SN is a quantity of stemmata in the sentence. The result of such operations can be seen in the Fig. 3 (for stemmata) and in the Fig. 4 (for sentences).

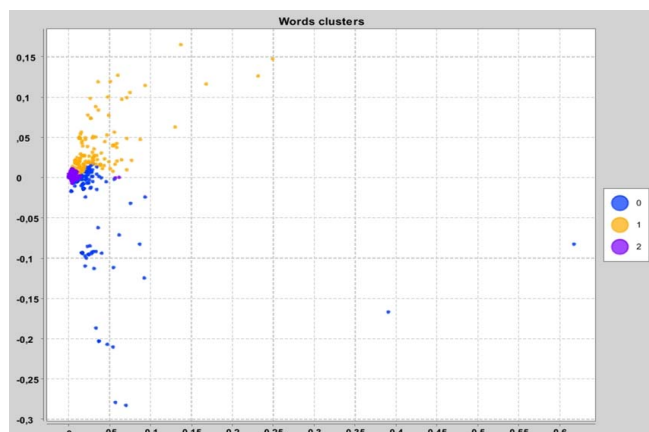


Fig. 3. Projection of clusterization for stemmata. 0, 1, 2 – numbers of clusters

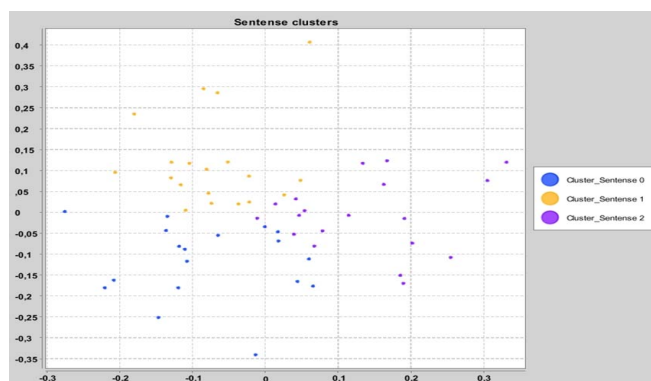


Fig. 4. Projection of clusterization for sentences. Cluster_Sentence 0, Cluster_Sentence 1, Cluster_Sentence 2 – numbers of clusters

The formation of semantic network of the document becomes the final stage. The outline of convex figure is built by Jarvis algorithm on the basis of coordinates of points of every cluster-stemma. The weight – number of stemmata included into it – is determined for every cluster – stemma from whence the semantic graph of relations of clusters in the descending order of their weight is built. Hit of points forming every cluster – sentence is verified for every figure of clusters – stemmata obtained by Jarvis algorithm. If such points are found then the cluster of sentence is connected in the network to the cluster-stemma where the weight of relation is a number of points having hit into the outline of cluster-stemma. The result of system's work is represented in the Fig. 5.

V. RESULTS OF EXPERIMENTS

Following on from the obtained results the formation of mathematical model of the document in the systems of automated texts processing becomes possible. The system was verified with the use of texts created as a result of automatic generation on the basis of patterns for the purpose of verification of the adequacy of obtained model. Such texts are statically correct, but they have weak notional relations between their parts. The result of processing of such text is

represented in the Fig. 6 (projection of clusters – stemmata and sentences) and in the Fig. 7 (resulting semantic network).

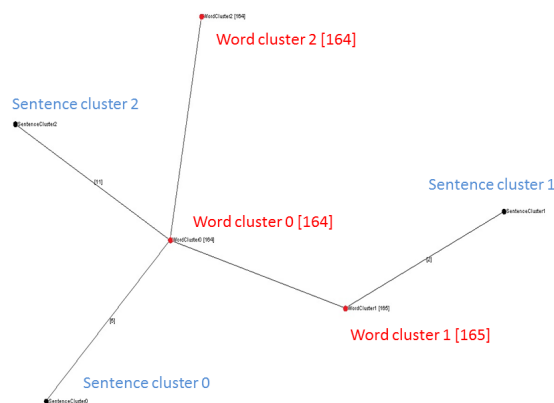


Fig. 5. The result of system's work – semantic network of the document. WordCluster corresponds to clusters – stemmata, SentenceCluster corresponds to clusters – sentences.

In spite of the fact that the size of the automatically generated text was coincident with the example provided before, its semantic network has the apparent differences. This observation enables to make an assumption not only about the adequacy of the semantic model, but also about the possibilities of its application beyond the task of building of inquiry/response systems. Such data as the quantity of clusters – sentences and clusters – stemmata, quantity of relations and their weight, weight of clusters – words, included into the semantic network, can be used for training of models, included, for example, to the systems of automatic determination of plagiarism or text coherence.

Program model of the text obtained in such way can be used for the purpose of building of systems with complex intelligent component of semantic analysis of the text in the natural language. The example of such application is a system of definition of text coherence, described in the paper [8]. Taken data are transferred to the ingress of asynchronous neuronal network that makes a decision on text coherence on the basis of data from model training corpus. It should be considered that described semantic characteristics depend on text size, sob taken data require prenormalization. For this purpose the corpus from eighty texts on the topic of information technologies, astronomy and incoherent texts obtained due to services of frequency autogeneration was compiled; each text was characterized by two values – standard text size W_N obtained according to the formula (6):

$$W_N = \frac{W_i - W_{\min}}{W_{\max} - W_{\min}} \quad (6)$$

where W_i is a total quantity of words, W_{\min} and W_{\max} is a minimum and maximum quantity of words in the training corpus and normalized semantic value SN , obtained according to the formula (7):

$$S_N = \frac{W_U}{W} \cdot \frac{CW_C}{CW} \quad (7)$$

where W_U is a total quantity of stemmata, W is a total quantity of words, CWC is a quantity of clusters – stemmata, related to clusters – sentences, CW is a total quantity of clusters – stemmata. Data obtained in such way compile the training sample collection for neuronal network.

Sample collection from 20 texts, both non-coherent (auto-generated) and real scientific texts on topics of astronomy, information technologies and economics was drawn for the purpose of system testing. In addition to it, the sample collection included also the text assembled from various coherent parts of texts of one topic, but not related semantically in general. The results of texts processing are presented in the Table II, where «n» corresponds to auto-generated text, «z» means connected text, «s» is a text assembled from various parts, 1 – prognosis points to text coherence, 0 – prognosis points to texts incoherence.

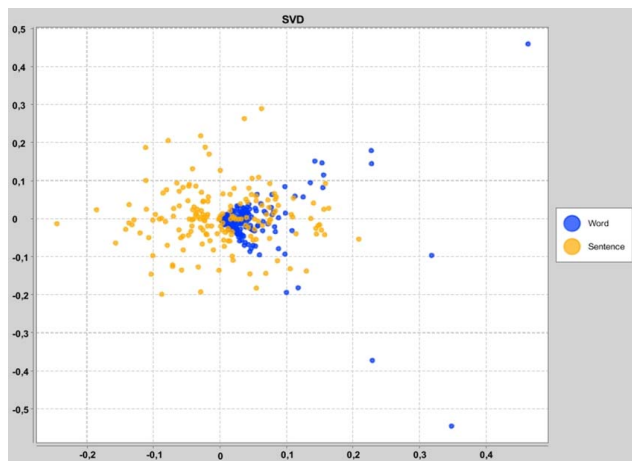


Fig. 6. Projection of singular decomposition for the text with weak semantic relations: Word – coordinates – stemmata, Sentence – coordinates – sentences.

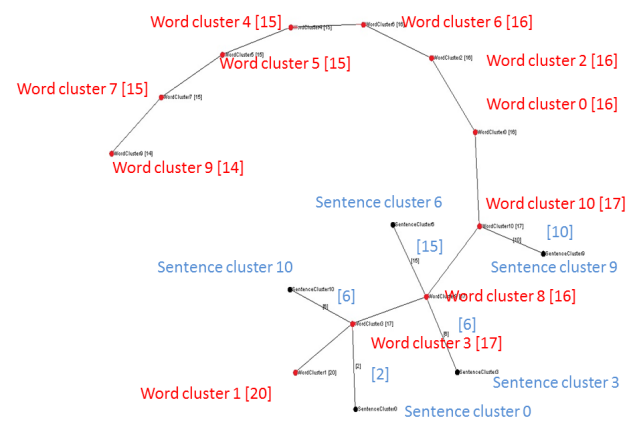


Fig. 7. The result of work of system for the text with weak semantic relations – semantic network of the document. WordCluster corresponds to clusters – stemmata, SentenceCluster corresponds to clusters – sentences.

TABLE II. PROVISION OF TEXT COHESIVENESS.

n	n	n	n	n	n	n	n	n	n
0	0	0	0	0	0	0	0	0	0
z	z	z	z	z	z	z	z	z	s
1	1	1	1	1	1	1	1	1	0

VI. CONCLUSION

There was formed and realized, in the course of the work, at the software level the approach to the applied creation of the semantic nets of the natural language text. The main advantage of the described model is a high degree of adaptability, expressed in the absence of the need for preliminary linguistic annotation of the working set of texts or adding in the system basis some set of linguistic rules. The made researches show that in spite of the preliminary absence of any deep semantic knowledge, the system is resistant to the frequency actions from the side of automatically generated text with low semantic indexes and it is oriented precisely to the processing of semantic relations in the text. The created approach provided the basis for the system of automatic determination of the text semantic connectivity, the work results of which confirm the applied applicability of the developed method for the semantic net construction to the tasks of program processing of the complex semantic text structures regardless of the natural language.

REFERENCES

- [1] A. N. Shevtsov, S. I. Sorokin, and Y. O. Mamadkulov, "System of synthesis of educational tests on the basis of formal grammars," journal "Software programs and systems", no. 2(102), p.181-185, 2013.
- [2] N. I. Gurin, and Y. A. Zhuk, "Semantic network of an electronic workbook for dialog with virtual teacher," international scientific and technical Internet conference "Information Processing Technologies in Education, Science and Production", Belorussian State Technological University, Minsk, 2015.
- [3] M. Yu. Gubin, V. V. Razin, and A. F. Tuzovsky, "Application of semantic networks and frequency characteristics of texts on natural languages for the creation of semantic metapopsis," Problems of Informatics, p.59-64, 2011.
- [4] I. I. Yusupova, and M. M. Gayanova, "Semantic Networks and Producing Models for the Analysis of University Educational Programs," Newsletter of Ufa state aviation technical university, pp. 120-126, 2006.
- [5] N. G. Zagoruiko, A. M. Naletov, and I. M. Grebenkin, "On the way to automatic construction ontology," Materials of international conference "Dialog", 2013.
- [6] Pismak A.E., Kharitonova A.E. (2016) // The method of automatic formation of a semantic network from weakly structured sources // . Nauch.-tekhnich. vestn. ITMO [Scientific and Technical Journal of Information Technologies, Mechanics and Optics]. 2016, vol. 16, no. 2, p. 324-330.
- [7] Tarasov D. S. (2015) - Natural language generation, paraphrasing and summarization of user reviews with recurrent neural networks [Text] /Tarasov D. S./ "Computer linguistics and Intellectual Technologies", No14(vol. 1), 2015, p.607-614//Materials of international conference "Dialog", 2015.
- [8] O.S. Volkovsky, Y.R. Kovylin. Computer system of automatic determination of the text coherence [Text] // System Technologies; Regional Interuniversity Collection of Scientific Papers. -Release 1 (112) 2017. - Dnipro, 2017.
- [9] O.S. Volkovsky, Y.R. Kovylin. Analysis of the modern approaches to the task of automatic text generation in the natural language [Text] // System Technologies; Regional Interuniversity Collection of Scientific Papers. - Release 1 (100) 2016. - Dnepropetrovsk, 2016.
- [10] N.N. Leontyeva. (2006) Automated comprehension of texts: Systems, models, resources. [Text]/N.N. Leontyeva//Moscow – 2006.

Fuzzy Clustering of Distorted Observations Based On Optimal Expansion Using Partial Distances

Alina Shafronenko
Informatics Department
Kharkiv National University of Radio Electronics
Kharkiv, Ukraine
alina.shafronenko@nure.ua

Yevgeniy Bodyanskiy
Control Systems Research Laboratory
Kharkiv National University of Radio Electronics
Kharkiv, Ukraine;
yevgeniy.bodyanskiy@nure.ua

Artem Dolotov
Control Systems Research Laboratory
Kharkiv National University of Radio Electronics
Kharkiv, Ukraine
artem.dolotov@gmail.com

Galina Setlak
Rzeszow University of Technology
Rzeszow, Poland;
gsetlak@prz.edu.pl

Abstract—The neural system that solves a problem of fuzzy clustering of distorted observations based on optimal expansion strategy using partial distance is proposed in this article. To solve this problem we propose the learning algorithm based on hybrid of rule “Winner-Takes-More” using modified self-organizing neuro-fuzzy Kohonen network. This modified system is characterized by basic characteristics, such as: high speed, simple numerical realization, processing of distorted information in online mode.

Keywords— *Kohonen self-organizing network, fuzzy clustering; incomplete observations with gaps, partial distance, optimal expansion*

I. INTRODUCTION

The problem of data sets clustering often occurs in many practical tasks, and for its solution has been successfully used artificial neural networks [1] and methods of fuzzy systems [2]. It is usually assumed that original array is specified a priori and processing is realized in batch mode. Currently, due to the widespread using of Dynamic Data Mining [3], which is associated with the processing of observations arriving sequentially, sometimes at a high frequency, the known methods of distorted data clustering become incompetent.

More effective in situation when the data are fed to the processing in on-line mode clustering is using of self-organizing Kohonen network [4], the use of which implies that the original vector data contain all the components. Here we attempt to hybridize the self-organizing maps and methods of fuzzy clustering of distorted observations with missing values that is based on the optimal expansion using the partial distances and nearest prototypes [2].

II. PROBLEM STATEMENT

Let's present the distorted data arriving for processing in the form of the table "object-property" as shown on Table I.

The Table 1 contains information about N feature vectors of order n $X = \{x_1, x_2, \dots, x_N\} \subset R^n$, $x_k \in X, k = 1, 2, \dots, N$ that arrive for processing in online mode. Result of observations clustering is the partition of

initial data into m classes ($1 < m < N$) with some level of membership $U_q(k)$, where here k -th is a feature vector to the q -th cluster ($1 \leq q \leq m$).

TABLE I. THE “OBJECT-PROPERTY” TABLE WITH DISTORED OBSERVATION

	l	...	p	...	j	...	n
l	x_{ll}	...	x_{lp}	...	x_{lj}	...	x_{ln}
...
i	x_{il}	...	x_{ip}	...	x_{ij}	...	x_{in}
...
k	x_{kl}	...	x_{kp}	...	x_{kj}	...	x_{kn}
...
N	x_{Nl}	...	x_{Np}	...	x_{Nj}	...	x_{Nn}

Incoming data in a first stage are standardized and centered so that all observations belong to the hypercube $[-1, 1]^n$. If there is an unknown number of missing values in the vector images \tilde{x}_k , that form the array \tilde{X} , let's introduce the sub-arrays:

$$X_F = \{\tilde{x}_k \in \tilde{X} \mid \tilde{x}_k - \text{if vector containing all components}\}$$

$$X_P = \{\tilde{x}_{ki}, 1 \leq i \leq n, 1 \leq k \leq N \mid \text{if all values } \tilde{x}_k, \text{ available in } \tilde{X}\}$$

$$X_G = \{\tilde{x}_{ki} = ?, 1 \leq i \leq n, 1 \leq k \leq N \mid \text{if all values } \tilde{x}_k, \text{ absent in } \tilde{X}\}$$

III. OVERVIEW OF PARTIAL DISTANCES

The choice of the distance between objects is the focal point of the investigation, and the final variant of the partitioning of objects into classes depends on it for a given partitioning algorithm. The simplest way to calculate the distances between objects in a multidimensional space is to calculate the Euclidean distances, but the Euclidean metrics (and its square) is calculated from the source, rather than from the standardized data. In this case it is suggested to use the partial distance described by formula 1

$$D_P^2(\tilde{x}_k, w_q) = \frac{n}{\delta_{k\Sigma}} \sum_{i=1}^n (\tilde{x}_{ki} - w_{qi})^2 \delta_{ki} \quad (1)$$

where w_{qi} -ith component of q -th prototype (centroid) of the corresponding cluster ($q = 1, 2, \dots, m$),

$$\delta_{ki} = \begin{cases} 0 & | \tilde{x}_{ki} \in X_G, \\ 1 & | \tilde{x}_{ki} \in X_F, \end{cases} \quad \delta_{k\Sigma} = \sum_{i=1}^n \delta_{ki} .$$

Easy to see that for $\tilde{x}_k \in X_F$ the partial distance (1) becomes an usual Euclidean metric. In the opposite case, the distance between \tilde{x}_k and prototype w_q is estimated on the basis of the components available in the \tilde{x}_k as shown in Fig.1.

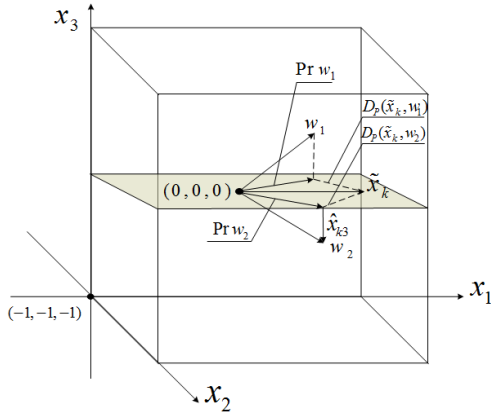


Fig. 1. The strategy of partial distances

Here, the three-dimensional vector \tilde{x}_k lacks one component \tilde{x}_{k3} so that the distance is measured on the plane x_1, x_2 , and instead of the prototypes w_1 and w_2 we use their projections onto this plane $\text{Pr } w_1$ и $\text{Pr } w_2$.

The competition process, which underlies the training of Kohonen map, is organized on the basis of estimating the partial distances, i.e. when a distorted (incomplete) vector of observations arrives \tilde{x}_{k+1} first estimated distance between this vector and the centroids $w_1(k), w_2(k), \dots, w_m(k)$ and then looking for the neuron-winner $w_q(k)$ such that

$$D_P^2(\tilde{x}_{k+1}, w_q(k)) = \arg \min_q \{D_P^2(\tilde{x}_{k+1}, w_1(k)), \dots, D_P^2(\tilde{x}_{k+1}, w_m(k))\} .$$

Further, the missing component $\tilde{x}_{k+1,i}$ is replaced by the corresponding component of the centroid of the winner neuron: $\hat{x}_{k+1,i} = w_{qi}(k)$. In Fig. 1, as missing component \tilde{x}_{k3} its estimate $\hat{x}_{k3} = w_{23}$ is used. Further, the corresponding centroid is specified on the basis of the standard rule of self-learning of WTA ("The Winner Takes All") in the form

$$w_q(k+1) = w_q(k) + \eta(k+1)(\tilde{x}_{k+1} - w_q(k)) \quad (2)$$

where $0 < \eta(k+1) < 1$ - parameter of the learning step.

To improve the quality of Kohonen networks learning, it's possible by adjusting at each step not only the neuron-winner, but a whole group of neurons according to WTM-rule ("Winner Takes More") in the form

$$w_l(k+1) = w_l(k) + \eta(k+1)\varphi(q, l)(\tilde{x}_{k+1} - w_l(k)) \quad (3)$$

$$\forall l = 1, 2, \dots, m$$

where $\varphi(q, l)$ - a neighborhood function that depends on the distance between the centroid of the winner neuron w_q and arbitrary neuron w_l . As neighborhood functions, as a rule, kernel (bell-shaped) constructions with an extremum in w_q , i.e. $\varphi(q, q) = 1$ are used.

It is interesting to note that the use of Cauchian as neighborhood function

$$\varphi_l(k+1) = U_l^\beta(k+1) = \left(\frac{(D_P^2(\tilde{x}_{k+1}, w_l(k)))^{\frac{1}{1-\beta}}}{\sum_{r=1}^m (D_P^2(\tilde{x}_{k+1}, w_r(k)))^{\frac{1}{1-\beta}}} \right)^\beta \quad (4)$$

(here $\beta > 1$ - the fuzzyfier), which is associated not with the winner, but with each of the prototype centroids, leads to the fact that relations (3), (4) are transformed into an adaptive algorithm for probabilistic fuzzy clustering of data with gaps [5-8] essentially FCM - clustering method for incomplete data [2]. Thus, within the WTM-rule of Kohonen self-learning network (3) it is possible to solve on-line problems of both crisp and fuzzy clustering using the standard architecture of the self-organizing network.

IV. OPTIMAL EXPANSION STRATEGY USING PARTITIONAL DISTANCE

Today, there are many situations when data are fed to processing sequentially as it occurs during training Kohonen self-organizing maps [4] or their modifications [9]. In this connection, an adaptive neuro-fuzzy Kohonen network is proposed, that is designed to solve the problem of clustering distorted data based on the strategy of partial distances. At the same time, in situations where the amount of distorted data is too large, the strategy of partial distances is ineffective. Thus, it may be necessary to solve the clustering problem simultaneously with restoring the gaps in the "object-property" table. In this situation, an approach based on the so-called optimal expansion strategy can be more effective. The optimal expansion strategy using the partition distances is that the elements of the submatrix X_G are considered as additional variables, which are estimated by minimizing the goal function E . Thus, in parallel with clustering, an evaluation of the gaps is made. The proposed fuzzy c-means algorithm, based on the optimal expansion strategy based on partition distances, consists of a sequence of steps [10]:

1 Step: Define the initial parameters for the algorithm:
 $\beta > 0$; $1 < m < N$; $\varepsilon > 0$; $w_q^{(0)}$; $1 \leq q \leq m$;
 $\tau = 0, 1, 2, \dots, Q$; $X_G^{(0)} = \{-1 \leq \hat{x}_{ki}^{(0)} \leq 1\}$.

2 Step: Calculation of membership levels:

$$U_q^{(\tau+1)}(k) = \arg \min_{U_q(k)} E(U_q(k), w_q^{(\tau)}, X_G^{(\tau)}) = \frac{(D_p^2(\hat{x}_k^{(\tau)}, w_q^{(\tau)}))^{\frac{1}{1-\beta}}}{\sum_{l=1}^m (D_p^2(\hat{x}_k^{(\tau)}, w_l^{(\tau)}))^{\frac{1}{1-\beta}}}$$

3 Step: Calculation the cluster's centroids:

$$w_q^{(\tau+1)} = \arg \min_{w_q} E(U_q^{(\tau+1)}(k), w_q, X_G^{(\tau)}) = \frac{\sum_{k=1}^N (U_q^{(\tau+1)}(k))^{\beta} \hat{x}_k^{(\tau)}}{\sum_{k=1}^N (U_q^{(\tau+1)}(k))^{\beta}}$$

4 Step: Stop, if $\|w_q^{(\tau+1)} - w_q^{(\tau)}\| < \varepsilon \forall 1 \leq q \leq m$ or $\tau = Q$

else go to step 5.

5 Step: Estimation of gaps:

$$\hat{x}_{ki}^{(\tau+1)} = \frac{\sum_{q=1}^m (U_q^{(\tau+1)}(k))^{\beta} w_{qi}^{(\tau+1)}}{\sum_{q=1}^m (U_q^{(\tau+1)}(k))^{\beta}}$$

The processing of information using the optimal extension algorithm is organized in the form of sequences of recalculations:

$$\begin{aligned} w_q^{(0)} &\rightarrow U_q^{(1)} \rightarrow \hat{x}_{ki}^{(1)} \rightarrow w_q^{(1)} \rightarrow U_q^{(2)} \rightarrow \dots \\ &\rightarrow w_q^{(\tau)} \rightarrow U_q^{(\tau+1)} \rightarrow \hat{x}_{ki}^{(\tau+1)} \rightarrow w_q^{(\tau+1)} \rightarrow \dots \rightarrow w_q^{(Q)}. \end{aligned}$$

Thus, the clustering algorithm can be rewritten in online mode

$$\left\{ \begin{aligned} U_q^{(\tau+1)}(k+1) &= \frac{\left(\|\hat{x}_{k+1}^{(\tau)} - w_q(k)\|^2\right)^{\frac{1}{1-\beta}}}{\sum_{l=1}^m \left(\|\hat{x}_{k+1}^{(\tau)} - w_l(k)\|^2\right)^{\frac{1}{1-\beta}}}, \\ \hat{x}_{k+1,i}^{(\tau+1)} &= \frac{\sum_{q=1}^m (U_q^{(\tau+1)}(k+1))^{\beta} w_{qi}(k)}{\sum_{q=1}^m (U_q^{(\tau+1)}(k+1))^{\beta}}, \\ w_q(k+1) &= w_q(k) + \eta(k+1)(U_q^{(Q)}(k+1))^{\beta} * \\ &\quad * (\hat{x}_{k+1}^{(Q)} - w_q(k)). \end{aligned} \right. \quad (5)$$

The centroids of clusters can be recalculated in accelerated time:

$$\left\{ \begin{aligned} U_q^{(\tau+1)}(k+1) &= \frac{\left(\|\hat{x}_{k+1}^{(\tau)} - w_q^{(\tau)}(k+1)\|^2\right)^{\frac{1}{1-\beta}}}{\sum_{l=1}^m \left(\|\hat{x}_{k+1}^{(\tau)} - w_l^{(\tau)}(k)\|^2\right)^{\frac{1}{1-\beta}}}, \\ w_q^{(0)}(k+1) &= w_q^{(Q)}(k), \\ w_q^{(\tau+1)}(k+1) &= w_q^{(\tau)}(k+1) + \eta(k+1) * \\ &\quad * (U_q^{(\tau+1)}(k+1))^{\beta} (\hat{x}_{k+1}^{(\tau)} - w_q^{(\tau)}(k+1)), \\ \hat{x}_{k+1,i}^{(\tau+1)} &= \frac{\sum_{q=1}^m (U_q^{(\tau+1)}(k+1))^{\beta} w_{qi}^{(\tau+1)}(k+1)}{\sum_{q=1}^m (U_q^{(\tau+1)}(k+1))^{\beta}}. \end{aligned} \right. \quad (6)$$

V. EXPERIMENTAL RESEARCH

In experimental studies, the three main algorithms for clustering: FCM, Gustafson - Kessel and the proposed clustering algorithm based on the optimal expansion using partial distances were compared by the main clustering parameters: Classification Entropy (CE), Partition Coefficient (PC), Separation Index (S), Partition Index (SC), Dunn's Index (DI), Xie and Beni's Index (XB). We operated on data provided by the UCI repository data: Wine data set and Iris dataset. Each of the data sets has a certain number of observations with its attributes. For example, the Wine data set contains the results of a chemical analysis of three types of wines from different regions of Italy.

Table II and Table III show the results of clustering algorithms with different amounts of data distorted by gaps.

TABLE II. RESULTS OF EXPERIMENTS WITH 10 GAPS

Algorithms	Iris UCI repository						Wine UCI repository						
	PC	CE	SC	S	XB	DI	PC	CE	SC	S	XB	DI	PC
Optimal expansion strategy using partial distance	9.1249e-07	-4.6617e-04	0.3733	48.7067	48.8056	0.4010	1.1600e-13	-4.5675e-04	7.3872e-05	2.7115e-08	2.7180e-08	0.0218	9.11249e-07
FCM	0.7617	0.4383	0.0143	1.4946e-04	3.8569	0.0275	0.7908	0.3806	7.3348e-04	6.8417e-06	5.7110	0.0117	0.7617
Gustafson-Kessel	0.9462	0.1145	0.4789	0.0032	3.4618	0.3398	0.5507	0.6393	8.5933	0.0483	1.0750	0.1015	0.9462

As you can see from the obtained results of the algorithms, the proposed clustering method for many parameters of the data clustering quality is not inferior to the well-known algorithms and demonstrates quite good results of clustering data in online mode.

TABLE III. RESULTS OF EXPERIMENTS WITH 50 GAPS

Algorithms	Iris UCI repository							Wine UCI repository						
	PC	CE	SC	S	XB	DI	PC	CE	SC	S	XB	DI	PC	
Optimal expansion strategy using partitionial distance	9.1249e-07	-4.6617e-04	0.3775	48.7067	48.8301	0.3365	7.5181e-12	-5.4992e-04	1.1834e+04	4.1981e-06	4.2024e-06	0.0240	9.1249e-07	
FCM	0.7399	0.4632	0.0174	1.8345e-04	4.4887	0.0355	0.7892	0.3838	7.6110e-04	7.1760e-06	8.8618	0.0237	0.7399	
Gustafson-Kessel	0.9422	0.1177	0.5219	0.0035	3.4413	0.3341	0.5824	0.6010	4.7678	0.0288	1.1703	0.1030	0.9422	

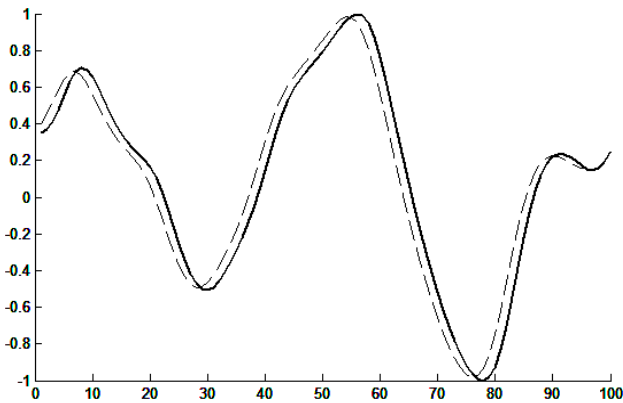


Fig. 2. Graph of estimates of gaps (dashed line) and real data (solid line)

Analyzing the results, we have plotted the recovered and original data. On the Fig.2 and Fig.3 Figures 1 and 2 show some of the results obtained, which demonstrates the work of the proposed algorithm. As can be seen from the graphs, the proposed algorithm well solves the problem.

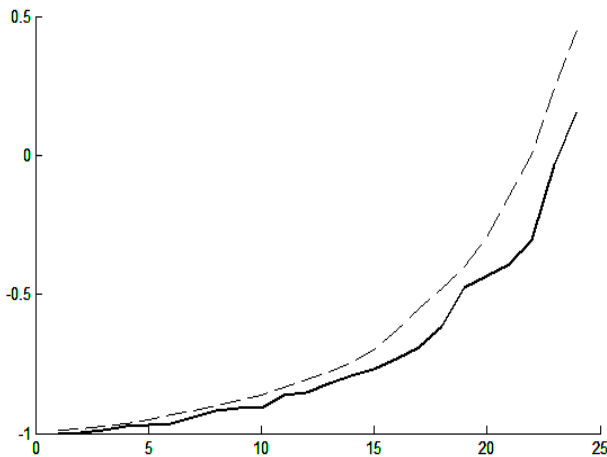


Fig. 3. Graph of estimates of gaps (dashed line) and real data (solid line)

VI. CONCLUSION

The neural systems that solves a problem of fuzzy clustering of distorted observations based on optimal expansion using partial distances is proposed in this article. To solve this problem we have proposed the learning algorithm based on hybrid of rule “Winner-Takes-More” using modified self-organizing neuro-fuzzy Kohonen network. This modified system is characterized by basic characteristic, such as: high rate, simple numerical realization, processing of distorted information in online mode.

REFERENCES

- [1] T Marwala, “Computational Intelligence for Missing Data Imputation, Estimation, and Management: Knowledge Optimization Techniques,” Hershey-New York: Information Science Reference, 2009.
- [2] J. C. Bezdek, “Pattern Recognition with Fuzzy Objective Function Algorithms,” Plenum Press, New York, 1981.
- [3] E. Lughofer, “Evolving Fuzzy Systems. Methodologies, Advanced Concepts and Applications,” Berlin-Hagenberg, 2011.
- [4] T. Kohonen, “Self-Organizing Maps,” Berlin: Springer-Verlag, 1995.
- [5] A. Y. Shafronenko, V. V. Volkova, and Ye. Bodyanskiy, “Adaptive clustering data with gaps,” Radioelectronics, informatics, control, no. 2, pp. 115-119, 2011. (in Russian)
- [6] Ye. Bodyanskiy, A. Shafronenko, and V. Volkova, “Adaptive clustering of incomplete data using neuro-fuzzy Kohonen network. Artificial Intelligence Methods and Techniques for Business and Engineering Applications,” ITHEA, Rzeszow, Poland; Sofia, Bulgaria. pp. 287-296, 2012.
- [7] Ye. Bodyanskiy, A. Shafronenko, and V. Volkova, “Adaptive fuzzy probabilistic clustering of incomplete data,” Int. J. “Information, models and analyses”, vol.2, no. 2, pp. 112-117, 2013.
- [8] Ye. Bodyanskiy, A. Shafronenko, and V. Volkova, “Neuro fuzzy Kohonen network for incomplete data clustering using optimal completion strategy,” Proceedings 20th East West Fuzzy Colloquium 2013, Zittau, pp. 214-223, 25-27 September 2013.
- [9] V. Kolodyazhniy, Ye. Bodyanskiy and Ye. Gorshkov, “New recursive learning algorithms for fuzzy Kohonen clustering network,” Proc. 17th Int. Workshop on Nonlinear Dynamics of Electronic Systems, Rapperswil, Switzerland, pp. 58-61., June 21-24, 2009.
- [10] R. J. Hathaway, and J. C Bezdek, “Fuzzy c-means clustering of incomplete data,”. IEEE Trans. on Systems, Man, and Cybernetics, vol. 31, no. 5, pp. 735-744, 2001.

On the Equivalence between AR Family Time Series Models and Fuzzy Models in Signal Processing

Anna Walaszek-Babiszewska
Department of Computer Science
Opole University of Technology
Opole, Poland
a.walaszekbabiszewska@gmail.com

Marek Rydel
Department of Computer Science
Opole University of Technology
Opole, Poland
m.rydel@po.opole.pl

Nataliia Kashpruk
Department of Computer Science
Opole University of Technology
Opole, Poland
n.kashpruk@gmail.com

Abstract— In the paper an advanced analysis of the relationships between statistical *Autoregressive* (AR) type models and fuzzy models have been presented. The examined family of AR type models includes *Autoregressive models of order p*, AR(p), *Threshold AR* (TAR) as well as *Smooth Transition Autoregressive* (STAR) models. On the other hand, fuzzy models representing different approach, characteristic for *Computational Intelligence* technics, have been tested for time series analysis and forecasting. The data have been taken from financial market. The research can enrich knowledge which is useful for experts using both approaches to modelling.

I. INTRODUCTION

Statistical models worked out in the area of mathematical statistics played a breakthrough role in time-discrete signal processing and were widely employed in a number of fields of science. These models were mainly developed for application in econometrics and control theory. Some clarification in the theory of time series, primarily on account of its application in control theory, was introduced by the work written by Box and Jenkins concerning linear models of time series: *Autoregressive* (AR), *Moving Average* (MA), their combinations *Autoregressive Moving Average* (ARMA) and *Autoregressive Integrated Moving Average* (ARIMA) [1]. Works in the field of econometrics resulted in another types of stochastic models including modelling of nonlinear time series, e.g. *Threshold Autoregressive* (TAR), *Smooth Transition Autoregressive* (STAR), *Self-Exciting Threshold Auto-Regressive* (SETAR), *Auto-Regressive Conditional Heteroscedasticity* (ARCH), as well as a number of other models. Together with the development of methods and techniques of artificial intelligence, fuzzy models and neuro-fuzzy models started to be applied for analysis and forecast of time series. During a few decades of existence of *Computational Intelligence*, the literature on the subject presents still not enough comparisons of both approaches to building models, identification and effects in terms of convenience of application, accuracy, computational volume, etc. The works of J.L. Aznarte and J.M. Benitez constitute an exception. In their paper we find the following proposition:

„The STAR (Smooth Transition Autoregressive) model is functionally equivalent to an Additive TSK Fuzzy Rule-Based (FRB) model with only one term in the rule antecedents.” [2].

The aim of the article is theoretical analysis of building models from autoregressive family (AR, TAR and STAR

models) as well as fuzzy models for indicating opportunities of gaining and using knowledge useful for constructing these models. The main criteria of the comparison analysis include:

- type of the state-space domain granulation,
- mathematical form of the models from AR family and fuzzy Mamdani’s as well as Takagi-Sugeno-Kang’s (TSK) models,
- statistical metrics, as Mean Squared Error (MSE) and autocorrelation function of residues.

Theoretical analysis will be supported by computational examples.

II. AUTOREGRESSIVE FAMILY OF MODELS

A. Autoregressive Models of Time Series

Autoregressive model constitute a scheme of a time-discrete stochastic process $\{X_n\}$, $n=1,2,\dots$ which assumes that future values of the process stand for a linear combination of its p past values

$$x_n = a_0 + a_1x_{n-1} + \dots + a_px_{n-p} + \varepsilon_n \quad (1)$$

where $p \geq 1$, $\{\varepsilon_n\}$ is a *white noise process* of the finite variance, $\sigma_\varepsilon^2 < \infty$, and a covariance function $\text{cov}(\varepsilon_i, \varepsilon_j) = 0$, $i \neq j$. Such model defined by (1) is known as the autoregressive model of order p , $AR(p)$.

When applied this model to forecast future values of the process, \hat{x}_n , parameters a_0, a_1, \dots, a_p can be determined by the least squares method and past values of the process $\{x_1, x_2, \dots, x_K\}$:

$$\sum_n [x_n - \hat{x}_n]^2 = \min_{\hat{a}} \sum_{n=p+1, \dots, K} [x_n - (a_0 + \dots + a_px_{n-p})]^2 \quad (2)$$

The assumption of stationarity of the process $\{X_n\}$ secures that $AR(p)$ model with parameters $\hat{a}_0, \hat{a}_1, \dots, \hat{a}_p$ estimated on the base of large sample $\{x_{p+1}, x_{p+2}, \dots, x_K\}$,

$K \gg p$ suites sufficiently for whole data population.

B. Threshold Autoregressive and Smooth Transition Autoregressive Models

In a practice the realizations of stochastic processes rather do not meet assumptions of stationarity, especially the assumption concerning the constant value of the mean process value. This is the reason that autoregressive models are not very good approximation for whole series.

The piecewise approach to modelling, by using e.g. the *threshold autoregressive model* (TAR) usually improves such approximation. In different subspaces of \mathcal{X} , process $\{X_n\}$ is described by local autoregressive models x_n^r . Transition from a local model to another one is described by the so-called transition function, λ_r , which assumes the value of 1 for each r -th subspace, whereas in the remaining ones it is equal to zero. The global model may be written as :

$$x_n = \sum_r \lambda_r x_n^r, \quad (3)$$

where x_n^r is an autoregressive model of the form (1) defined in the r -th range of the selected variable. The boundaries of the intervals stand for the thresholds of the model.

Smooth transition autoregressive (STAR) model contains continuous transition functions $\lambda_r(b^r, x_{n-1}^r)$ which define the location and shape of the transition between local autoregressive models x_n^r . Transition function $\lambda_r(b^r, x_{n-1}^r)$ takes its values from unit interval [0,1] and usually is a nonlinear function of independent variables.

III. FUZZY MODELS IN TIME SERIES ANALYSIS

A. Fuzzy Rule-Based Linguistic Model

The basis of constructing fuzzy models of systems is the input-output space division, X^p , into sub-areas where behaviour of the modelled system can be described by one conditional statement. It is the so-called information granulation process. The operation is analogous to space division in constructing TAR and STAR models.

Let us analyse the dependence $x_n = \varphi(x_{n-1})$, defined in space \mathcal{X}^2 , which is modelled as the set $\{R_i\}$, $i=1,2,\dots,I$ of conditional linguistic rules of the form:

$$R_i: \text{If } (x_{n-1} \text{ is } A_i) \text{ Then } (x_n \text{ is } A_i). \quad (4)$$

The input-output space, $\mathcal{X} \times \mathcal{X}$, is divided by fuzzy sets, $A_i \times A_i$, $i=1,2,\dots,I$. The antecedent of the rule defines fuzzy condition and the consequent part of the rule defines fuzzy conclusion. Fuzzy sets are most often defined by piece-wise linear membership functions (5), (6) or Gaussian membership function (7):

$$\mu_{A_i}(x; a_i, m_i, c_i) = \begin{cases} \frac{x - a_i}{m_i - a_i}, & a_i \leq x \leq m_i \\ \frac{c_i - x}{c_i - m_i}, & m_i < x \leq c_i \\ 0, & (x < a_i) \cup (x > c_i) \end{cases} \quad (5)$$

The form of membership fuzzy sets function is responsible for transformation of input information, that is to say for fuzzyfication process.

$$\mu_{A_i}(x; a_i, b_i, c_i, d_i) = \begin{cases} \frac{x - a_i}{b_i - a_i}, & a_i \leq x \leq b_i \\ 1, & b_i < x \leq c_i \\ \frac{d_i - x}{d_i - c_i}, & c_i < x \leq d_i \\ 0, & (x < a_i) \cup (x > d_i) \end{cases} \quad (6)$$

$$\mu_{A_i}(x) = \exp\left(-\frac{(x - m_i)^2}{2\sigma_i^2}\right). \quad (7)$$

In the fuzzyfication process the numerical value of input x_{n-1}^* is transformed into the activation level of a rule

$$\tau_i = \mu_{A_i}(x_{n-1}^*), \quad (8)$$

according to linear relation, (5) and (6), accordingly to non-linear relation (7) or in relation changing numerical values x_{n-1}^* to a constant value equal to 1 if $x_{n-1}^* \in (b_i, c_i]$, according to (6).

When the fuzzy reasoning procedure runs in compliance with Mamdani-Assilan formula, the fuzzy conclusion membership function is determined as follows [3]:

$$\mu_{A'}(x_n) = \max[\min[\tau_i, \mu_{A_i}(x_n)]]. \quad (9)$$

The *maximum* and *minimum* operations correspond to the logical union and intersection of fuzzy sets, whereas $\tau_i = \mu_{A_i}(x_{n-1}^*)$ is the level of activation of i -th formula for numerical value of input x_{n-1}^* .

Another operation influencing the transformed signal is defuzzyfication. In continuous space, $\mathcal{X} \subset R$, non-fuzzy value of output, x_n^* , constitutes the following weighted value, dependent on the area below the function $\mu_{A'}(x_n)$ of fuzzy conclusion A' :

$$x_n^* = \int_{\mathcal{X}} x_n \mu_{A'}(x_n) dx_n / \int_{\mathcal{X}} \mu_{A'}(x_n) dx_n \quad (10)$$

To sum up, while applying Mamdani's linguistic model for time series modelling the result of reasoning in the form of numerical output value $x_n^* \in \mathcal{X}$, for a given value of premise (input variable), $x_{n-1}^* \in \mathcal{X}$, generally constitutes a non-linear dependence $x_n^* = \varphi(x_{n-1}^*)$ mainly due to reasoning and defuzzification procedures. As it appears in fuzzyfication process it is possible to preserve linearity of transformation.

B. Simplified Method of Fuzzy Reasoning

Applying the Simplified Method of Fuzzy Reasoning, we obtain non-fuzzy output x_n^* as the weighted average of centroids, m_i , of the output variable fuzzy sets [4]:

$$x_n^* = \frac{\sum_i \tau_i \cdot m_i}{\sum_i \tau_i} \quad (11)$$

Generally, as regards model with one input variable, the levels of activation of rules meet the condition $\sum_i \tau_i = 1$ and then the output value

$$x_n^* = \sum_i \tau_i \cdot m_i \quad (12)$$

may be the linear dependence if $\tau_i(x_{n-1}^*)$ is the linear function, that is to say when fuzzy sets of membership function (5) have been chosen. Moreover, for $\tau_i = 1$, when only one formula is active, the output value in the local model is a constant value, $x_{n,i} = m_i$.

C. Takagi-Sugeno-Kang's Fuzzy Model

Using Takagi-Sugeno-Kang's (TSK) fuzzy model [5], for modelling dependencies $x_n = \varphi(x_{n-1})$ we build set $\{R_i\}$, $i=1,2,\dots,I$ of rules of the form:

$$R_i: \text{If } (x_{n-1} \text{ is } A_i) \text{ Then } x_n^i = \alpha_i x_{n-1}. \quad (13)$$

These rules differ from the form (4) in that, there is a non-fuzzy function of the input numerical values in formula successor (13), in this case this being a linear function. Usually, parameters α_i , $i=1,2,\dots,I$ are known. A single rule provides a local linear model. The global model is obtained as the weighted sum of active rule outputs

$$x_n^* = \frac{\sum_i \tau_i \cdot x_n^i}{\sum_i \tau_i}, \quad (14)$$

where τ_i is the activation level of i -th rule, (8). Assuming that $\sum_i \tau_i = 1$, for the input x_{n-1}^* , relationship (14) leads to the form:

$$x_n^* = \sum_i \tau_i \cdot \alpha_i x_{n-1}^* \quad (15)$$

Taking into account that $\tau_i(x_{n-1}^*)$ is a linear or exponential function, formula (15) does not provide a linear dependence $x_n = \varphi(x_{n-1})$ any more. While for $\tau_i = 1$ with only one rule being active, the output value is identical with the equation in the i -th rule consequent's part

$$x_n^* = x_n^i = \alpha_i x_{n-1}^*. \quad (16)$$

Hence, the fuzzy sets in a part of premise rule (13) of the TSK model cannot be entirely arbitrary. They are usually sets of trapezoid membership functions where the linear part (increasing or decreasing) corresponds to that part of space \mathcal{X} which belongs to two fuzzy sets simultaneously. Reasoning provides smoothing, according to (14), of two linear models described by two rules.

IV. EXEMPLARY CALCULATIONS

In the research a real time series of WIG20, Polish market indicator, $\{x_n\}$, $n=1,2,\dots,150$, $x \in \mathcal{X} = [1630, 2080]$, was used to demonstrate features of tested models. Preliminary test of the series was composed of: calculations of a mean value, variance and autocorrelation function of the series. According to that, space \mathcal{X} has been divided into subspaces by the threshold $x_f=1855$ for building TAR model (see Fig.1 and Fig. 2). Autocorrelation function of the series proved, that $\{x_n\}$ is not a realization of the white noise process but constitutes the realization of the long memory stochastic process. Therefore searching for time series models is justified.

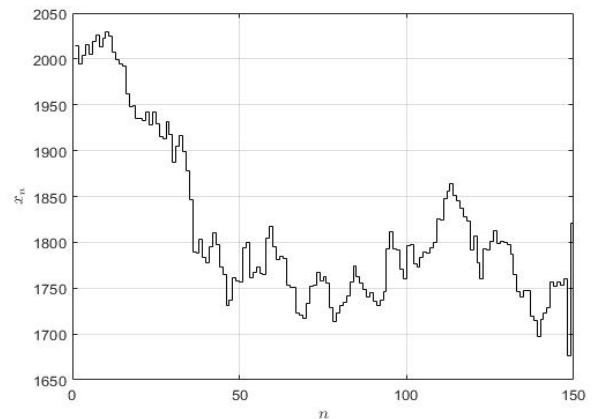


Fig. 1. Time series x_n of WIG 20

It is assumed to search for models containing only one lagged value of the series, $\hat{x}_n = \hat{\varphi}(x_{n-1})$, which means a

one-step prognosis. The models AR(1), TAR and TSK of rules numbers 3 and 5 have been established. The mean square errors MSE of models and MSE related to variance of row data D_x^2 , have been computing. Moreover, autocorrelation functions of residues for all models have been calculated and presented in Fig. 3.

The models have the following form:

$$\text{AR}(1) \quad \hat{x}_n = 0.9991x_{n-1}$$

$$\text{TAR}(1) \quad \hat{x}_n = \begin{cases} 0.9979x_{n-1}, & x_{n-1} \geq 1855 \\ 0.9997x_{n-1} & x_{n-1} < 1855 \end{cases}$$

TSK (3)

$$R_1: \text{If } (x_{n-1} \text{ is } A_1) \text{ Then } x_n^1 = 0.6601 \cdot x_{n-1} + 621.6$$

$$R_2: \text{If } (x_{n-1} \text{ is } A_2) \text{ Then } x_n^2 = 1.296 \cdot x_{n-1} - 550.1$$

$$R_3: \text{If } (x_{n-1} \text{ is } A_3) \text{ Then } x_n^3 = 1.389 \cdot x_{n-1} - 794.4$$

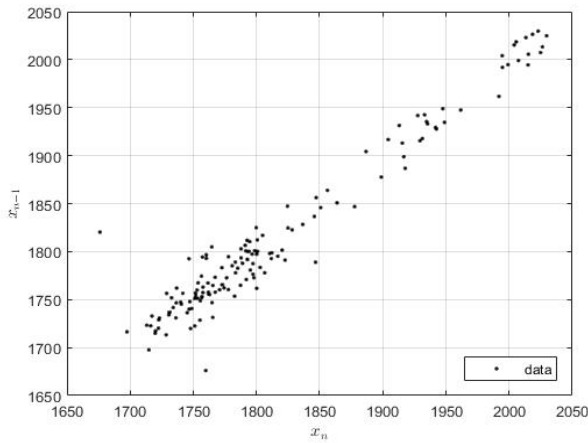


Fig. 2. Diagram $x_n = \phi(x_{n-1})$ of data

where membership functions of particular fuzzy sets A_1, A_2, A_3 are triangular with following parameters:

$$\mu_{A_1}(x; a_1, m_1, c_1) = \mu_{A_1}(x; 1630, 1675, 1855),$$

$$\mu_{A_2}(x; a_2, m_2, c_2) = \mu_{A_2}(x; 1675, 1855, 2035),$$

$$\mu_{A_3}(x; a_3, m_3, c_3) = \mu_{A_1}(x; 1855, 2035, 2080).$$

TABLE I. PARAMETERS OF RESIDUES OF PARTICULAR MODELS

Model	MSE	$MSE / D_x^2, [\%]$
AR(1)	451.49	5.78
TAR(1)	288.18	3.67
TSK(3)	371.43	4.73
TSK(5)	320.51	4.08

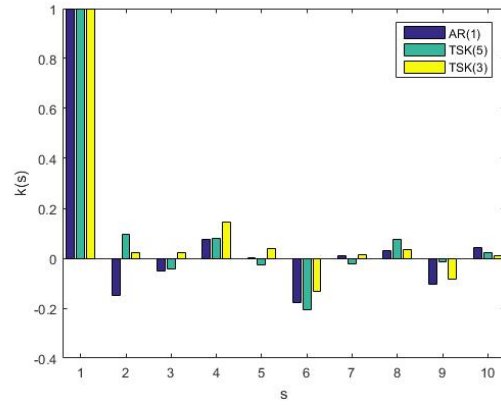


Fig. 3. The autocorrelation function of residues of tested models

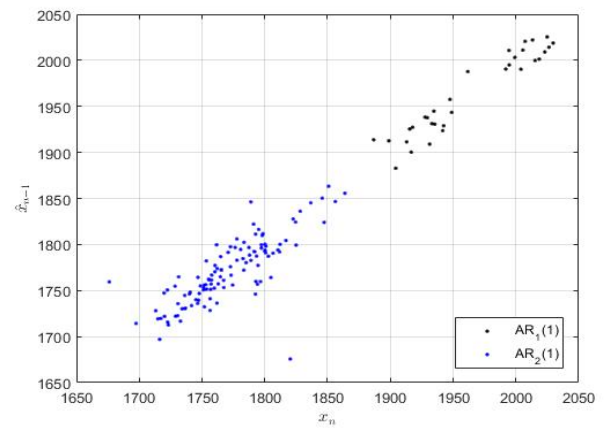


Fig. 4. Diagram $\hat{x}_n = \phi(x_n)$ of TAR model and data

Models AR(1) and TAR contain one parameters each, close to 1, however in the consequent parts of TSK models two-parameters linear relationships, $\hat{x}_n = ax_{n-1} + b$, are included. The values of MSE shown in Table 1 and diagram presented in Fig.4 prove that TAR model consisting of two AR(1) models gives the best mapping of input time series. The autocorrelation function of residues point out that there is a correlation on a level lower than 0.2. Hence it can be assumed that all the models describe changeability of input series satisfactorily.

V. CONCLUSIONS

Taking into account theoretical analysis of the models conducted in paragraphs I. – III. the following differences in the procedures of creating AR linear models and fuzzy rule-based models may be pointed out :

- AR linear models are created as *Least Squares* approximation of the entire data set.
- In fuzzy rule-based modelling each rule constitutes a local model built on a part of data set and aggregation procedure plays the role of fuzzy transition.
- Linguistic fuzzy models with Mamdani-type reasoning and Simplified Method of Fuzzy Reasoning

give mainly nonlinear dependencies $x_n = \varphi(x_{n-1})$ due to the specificity of reasoning and defuzzyfication procedures.

- Fuzzy rule-based TSK models are closest to TAR and STAR models; local linear models represented by particular rules are aggregated as weighted sum where weight coefficients are not constant but depend on an input variable.
- In order to obtain a linear mapping of local TSK, $x_n^i = \varphi(x_{n-1}^i)$, it is advisable to apply trapezoid fuzzy sets in the input variable space.

The conducted exemplary calculations for a given time series show more comparability of accuracy of the tested models although they differ in structure and even the equation form. The choice of model and method of its obtaining is a matter of the user's choice.

REFERENCES

- [1] G. Box and G. Jenkins, *Time Series Analysis: Forecasting and Control*, San Francisco: Holden-Day, 1970.
- [2] J. L. Aznarte and J. M. Benitez, "The links between statistical and fuzzy models for time series analysis and forecasting," in: *Time Series Analysis, Modeling and Applications; A Computational Intelligence Perspective*, W. Pedrycz and Shyi-Ming Chen, Eds. *Intelligence Systems Reference Library*, vol. 47, Springer, pp. 1-30, 2013.
- [3] E. H. Mamdani and S. Assilan, "An experiment in linguistic synthesis with a fuzzy logic controller," *International Journal of Man-Machine Studies*, 20(2), pp. 1-13, 1970.
- [4] R. R. Yager and D. P. Filev, *Essentials of Fuzzy Modeling and Control*, John Wiley & Sons, Inc. 1994.
- [5] T. Takagi and M. Sugeno, "Fuzzy identification of systems and its applications to modeling and control," *IEEE Transactions on Systems, Man and Cybernetics*, 15, pp.116-132, 1985.

Information Technology of Gene Expression Profiles Processing for Purpose of Gene Regulatory Networks Reconstruction

S. Babichev
Kherson National Technical University
Kherson, Ukraine
Jan Evangelista Purkyně University
Usti nad Labem, Czech Republic
sergii.babichev@ujep.cz

V. Lytvynenko
Kherson National Technical University
Kherson, Ukraine
immun56@gmail.com

J. Škvor2
Jan Evangelista Purkyně University
Usti nad Labem, Czech Republic
jskvor@physics.ujep.cz

M. Korobchynskyi
Military-Diplomatic Academy named Eugene Bereznyak
Kiev, Ukraine
maks_kor@ukr.net

M. Voronenko
Kherson National Technical University
Kherson, Ukraine

Abstract—The paper presents the information technology of gene expression profiles processing in order to reconstruct gene regulatory networks. The information technology is presented as a structural block-chart, which contains all stages of studied data processing. DNA microchips of patients, which were studied on different types of diseases, were used as experimental data. The relative criteria of validation for all reconstructed networks were calculated during simulation process. The obtained results show high efficiency of the proposed technology. High values of the validation criteria indicate a high level of the obtained gene networks objectivity.

Keywords—objective clustering, reduction, biclustering gene expression profiles, gene regulatory network, reconstruction, validation

I. INTRODUCTION

Actuality of the problem is determined by the modern state of works in the field of gene expression profiles processing for the purpose of gene regulatory networks reconstruction. Gene regulatory network is a set of genes, which interact with each other to control the specific cell functions [1]. Qualitatively reconstructed gene regulatory network promotes to better understanding of the gene interaction mechanism in order to create new methods to early diagnostics and treatment of complex genetic diseases. The gene expression profiles, which are obtained by DNA microarray experiments or by RNA sequences technology, are the basis for the reconstruction of gene regulatory networks [2, 3]. High dimension of feature space is one of the distinctive peculiarities of the studied data. The reconstruction of gene networks based on the whole dataset of gene expression profiles is very complicated task/due to the following aspects: it requests large computer resources; complexity of the obtained networks complicates the obtained results interpretation. Therefore, it is necessary at early stage of gene regulatory network reconstruction to process the gene expression profiles with the use of current computational and information technologies of complex data processing. This process includes data filtering in the case of DNA microchip experiment performing, non-informative genes reducing, data clustering and biclustering in order to select mutually correlated genes and samples.

The issues concerning creation of the hybrid clustering methods were investigated in [4,5]. The authors propose neural network, which allows them to increase the quality of information processing. In [6,7] the authors proposed system which solves a clustering task of non-stationary data streams under uncertainty conditions when data come in the form of a sequential stream in an online mode. However, it should be noted that authors' researches are primarily focused on low-dimensional data processing. High-dimensional data processing are not considered in these works. In the papers [8, 9] the author presented a novel approach to solving the problem of task allocation among the agents which takes into account the restrictions on agents' communications and self-diagnosis strategies for multinodular systems. In [10–12] the authors considered the issues of handling uncertainties in the problems of modeling and forecasting dynamic systems within the framework of the dynamic planning methodology. However, the proposed methods do not allow us to increase the efficiency of the genes and conditions grouping. Bicluster analysis is actual to solve this problem nowadays [13,14]. The rows and columns are grouping in accordance with their mutual correlation during biclustering process. One of the main disadvantages of this technology is a high percent of information lost due to the high dimension of the initial data array. However, it should be noted that the effective technology of gene expression profiles processing does not exist nowadays. This fact can be explained by high dimension of features space that requests the use of complex data processing modern techniques.

The Aim of the paper is the development of information technology of gene expression profiles processing in order to reconstruct gene regulatory networks.

II. INFORMATION TECHNOLOGY OF GENE EXPRESSION PROFILES PROCESSING

The structural block-chart of the information technology of gene expression profiles processing in order to reconstruct gene expression networks is presented in Fig. 1.

The implementation of this technology involves the following stages:

Stage I. Formation of gene expression profiles array in the case of DNA microchip experiments.

The matrix of light intensities is obtained during DNA microchip experiment performing. Firstly, it is necessary to transform the light intensities to the expression of the corresponding genes. Implementation of this stage involves the following steps: background correction, normalization, PM correction and summarization. The methods, which can be used at each step, are shown in Fig. 2.

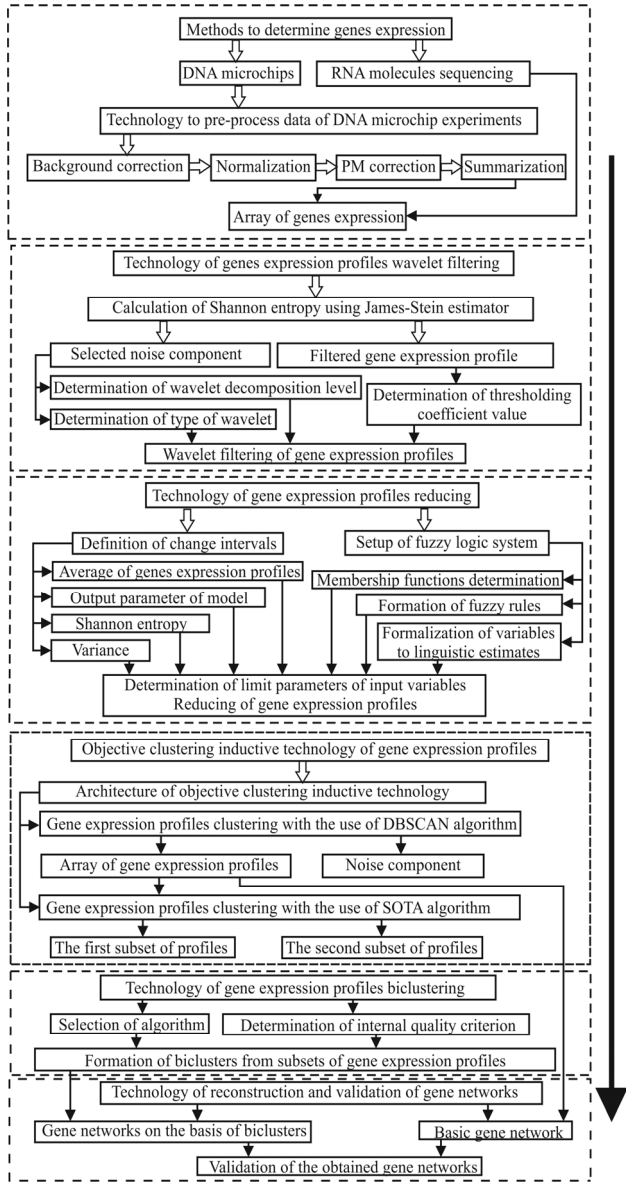


Fig. 1. Information technology of gene expression profiles processing

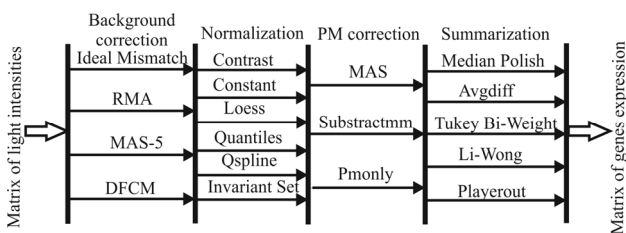


Fig. 2. Methods to evaluate the expression of genes

Estimation of data processing quality was performed with the use of Shannon entropy criterion, which was calculated based on James-Stein shrinkage estimator [15]. This method is based on the complex use of two different models: a high-dimensional model with low bias and high variance, and a lower dimensional model with larger bias but smaller variance. The probability of values distribution in a cell is calculated as follows:

$$p_i^{Shrink} = \lambda p_i + (1 - \lambda) p_i^{ML} \quad (1)$$

where p_i^{ML} is the probability of data values distribution in i -th cell, which is calculated by the maximum likelihood method, $p_i = 1/n_i$ is the estimation of probability in i -th cell, n_i is the quantity of features in this cell. Obviously, that p_i^{ML} corresponds to the high-dimensional model with low bias and high variance and p_i is the estimation with higher bias and lower variance of the features distribution. Intensity parameter λ in the proposed model is calculated as follows:

$$\lambda = \frac{1 - \sum_{i=1}^k (p_i^{ML})^2}{(n-1) \sum_{i=1}^k (p_i - p_i^{ML})^2}, \quad (2)$$

where n is the features quantity in the studied vector. Shannon entropy value in this case is estimated with the use of standard formula taking into account the method of probability calculation in the appropriate cell:

$$H^{Shrink} = - \sum_{i=1}^k p_i^{Shrink} \log_2 p_i^{Shrink} \quad (3)$$

It is obvious, that in the case of gene expression profiles informativity evaluation, the minimum value of Shannon entropy criterion corresponds to higher quality of the investigated data processing. Choice of the optimal combination of the methods was performed based on the minimum value of Shannon entropy during the enumeration of all combinations of these methods.

Stage II. Wavelet filtering of gene expression profiles.

The necessity of this stage is determined by the existence of background noise, which can be appeared during scanning of information from DNA microchip. The proposed technology of wavelet filter optimal parameters determination involves concurrent evaluation of Shannon entropy for both the filtered data and allocated noise component. Structural flowchart of this process is presented in Fig. 3.

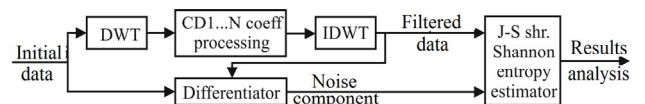


Fig. 3. Structural flowchart of wavelet-filtering process

Implementation of this process involves the following steps:

1. Choice of the mother wavelet from the list of the available wavelets.
2. Determination of the wavelet decomposition optimal level for the studied profiles based on the maximum

value of Shannon entropy, which is calculated for the allocated noise component. At this step the choice of the wavelet type from the family of the mother wavelet and the value of the thresholding coefficient are setup randomly from the range of the available values.

3. Determination of the wavelet type from the family of the appropriate mother wavelet based on the maximum value of Shannon entropy, which is calculated for the allocated noise component.

4. Determination of the thresholding coefficient optimal value based on the minimum value of Shannon entropy, which is calculated for the filtered data.

The algorithm works in such a way that if the value of Shannon entropy increases at the first step of thresholding coefficient change, the filtering process is stopped. In this case the studied data do not need filtering process.

Stage III. Gene expression profiles reducing.

The aim of this stage is division of the studied gene expression profiles into informative and non-informative in terms of complex use of statistical criteria and Shannon entropy. It is assumed that if variance or average absolute value of gene expression profiles is less than the corresponding boundary values, or if Shannon entropy of the corresponding gene expression profiles is greater than the boundary value, then these profiles are non-informative and they can be removed without significant loss of useful information. The fuzzy logic system was used to determine the boundary values of the appropriate parameters. Structural block-diagram of the algorithm of gene expression profiles reducing within the framework of fuzzy logic system is presented in Fig. 4. Practical implementation of this technology involves the following steps:

1. Calculation of the variance, the average absolute value and Shannon entropy for the expression profiles of the studied genes. Formation of data in the form of corresponding vectors: $var = \{var_1, var_2, \dots, var_m\}$, $abs = \{abs_1, abs_2, \dots, abs_m\}$, $entr = \{entr_1, entr_2, \dots, entr_m\}$.
2. Statistical analysis of the obtained vectors, determining the range of the appropriate parameters change.
3. Formation of the basic term-set for input variables (variance, average, Shannon entropy), and the output parameter, which determines the level of informativity of gene expression profiles QL (Quality).
4. Formation of the fuzzy rules, which are agreed/ with the input variables and the output parameter.
5. Determination of the boundary value of the output parameter QL_{lim} , which allows the gene expression profiles to divide into informative and non-informative. Determination of the step of the input variables changing within a given range.
6. Calculation of the output parameter QL for each combination of the input variables values corresponding to the appropriate gene expression profile. The result is formed as a vector: $QL = \{QL_1, QL_2, \dots, QL_m\}$.
7. Analysis of the results. Determination the values of the input variables that correspond to the boundary value of the output parameter.

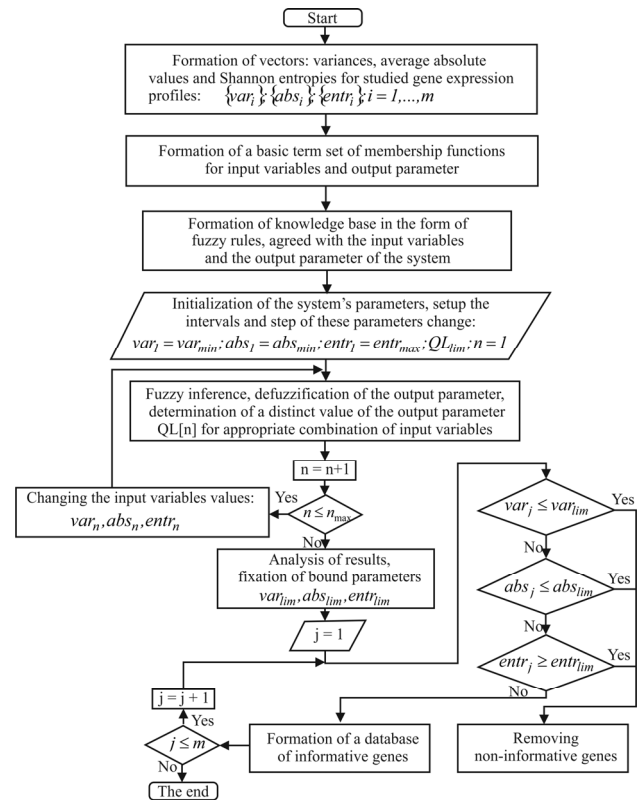


Fig. 4. Structural block-diagram of the algorithm of gene expression profiles reducing

8. A stepwise comparison of the variance, the average absolute value and Shannon entropy values of the gene expression profiles with the boundary values of the appropriate criteria. If the following conditions are true:

$$var \leq var_{lim} ; abs \leq abs_{lim} ; entr \leq entr_{lim}$$

then this gene is allocated from the data as non-informative. Otherwise, the gene profile is recognized as informative for the further analysis.

Stage IV. Step-by-step gene expression profiles clustering within the framework of the objective clustering inductive technology.

The studies concerning development of the objective clustering inductive technology and its practical implementation based on DBSCAN and SOTA clustering algorithms are described in the papers [16-18]. The proposed methods allow us to determine objectively the parameters of the appropriate clustering algorithm operation with the use of the internal, external and complex balance clustering quality criteria. The implementation of the objective clustering inductive technology involves division of the initial dataset into two equal power subsets (containing the same quantity of pairwise similar objects). Then, the clustering process is carried out on both subsets concurrently with calculation of the internal and external clustering quality criteria at each step of the algorithm operation. At final step the complex balance criterion is calculated based on the internal and external criteria. The maximum value of the balance criterion corresponds to the optimal parameters of the appropriate clustering algorithm operation.

The use of DBSCAN clustering algorithm allows us to allocate the genes, which are identified as noise. These genes are removed from the studied data. At the second step of the clustering process the gene expression profiles are divided into two clusters with the use of SOTA clustering algorithm. These subsets are used for the following bicluster analysis.

Stage V. Bicluster analysis of the obtained subsets of the gene expression profiles.

Allocation of small groups of mutually correlated genes and samples from DNA microarray is carried out during the biclustering process. Implementation of this stage allows us to reconstruct the gene network, which will be able to reflect objectively the influence of the appropriate genes to functional possibilities of the studied biological object. Structural block-chart of biclustering technology based on “ensemble” algorithm [19] is shown in Fig. 5. Practical implementation of the technology involves the following stages:

1. The studied data preprocessing and their formation in the form of a matrix, where rows and columns are the genes and samples respectively.

2. Fixation of $simthr$ parameter value, which determines the ratio of rows and columns quantity in biclusters. Setup of interval and step of thresholding coefficient value variation.

3. Data biclustering within the range of thresholding coefficient value change. Biclusters fixation at each step and calculation of the internal biclustering quality criterion.

4. Analysis of the obtained results, fixation of the thresholding coefficient value, which corresponds to the minimum of the internal quality criterion value.

5. Setup of range and step of $simthr$ parameter change. Data biclustering within the given range. Fixation of biclusters at each step and calculation of the internal biclustering quality criterion.

6. Analysis of the obtained results, fixation of the ratio of rows and columns quantity in biclusters, which corresponds to the minimum of the internal quality criterion value.

7. Data biclustering with the use of the “ensemble” algorithm optimal parameters. Fixation of the biclusters.

Stage VI. Gene regulatory networks reconstruction and validation of the obtained models.

The technologies of gene regulatory networks reconstruction and validation are presented in [20]. The reconstruction of gene networks was performed based on correlation inference algorithm. The optimal topology of the obtained gene networks was determined on the basis of the maximum value of general Harrington desirability index, which contains as its components the topological parameters of networks. Validation of the obtained models was performed based on the comparison analysis of the interconnection between the appropriate genes in the basic network and in the networks based on the obtained biclusters. ROC-analysis was used to calculate the relative criterion, which indicates a quality of the obtained gene networks.

III. RESULTS OF THE SIMULATION AND DISCUSSION

The DNA microchips of patients, which were investigated on diseases Alzheimer [21] and Parkinson [22], were used during simulation process. The first data contained 75 microchips, 16 samples were in the second database. Each of the simple contained 54675 genes. Fig. 6 shows the results

of the simulation to determine the optimal combination of the methods of DNA microchip processing in order to evaluate the expression of the studied genes.

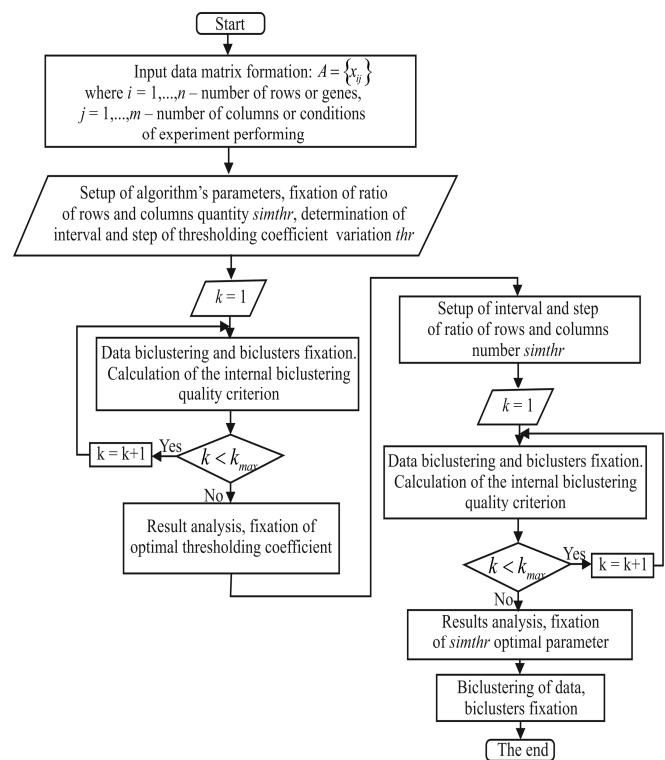


Fig. 5. Structural block-charts of biclustering technology based on “ensemble” biclustering algorithm

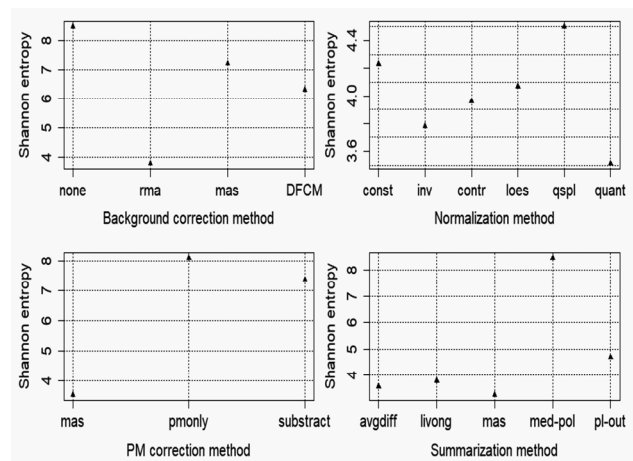


Fig. 6. Charts of distribution of Shannon entropy average values versus the used method of DNA microchip processing

Analysis of the simulation results allows us to conclude that in terms of the minimum values of Shannon Entropy the following methods are optimal for the studied DNA microchips: rma method background correction, quantile normalization, and mas methods PM correction and summarization. Fig. 7 shows the simulation results to determine the wavelet filter optimal parameters. Biorthogonal wavelet bior1.5 was used during the simulation process.

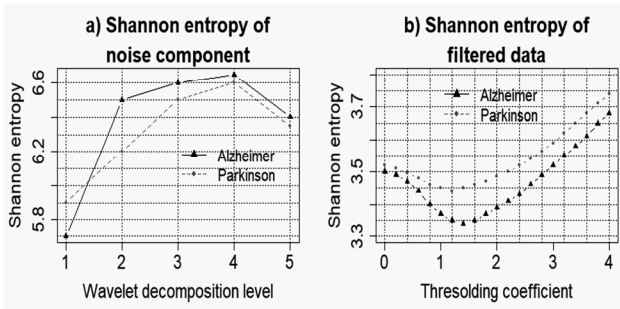


Fig. 7. Results of the simulation to determine the wavelet filter optimal parameters

The analysis of the obtained results allows concluding that in both cases the fourth wavelet decomposition level is the optimal one therefore Shannon entropy for the allocated noise component achieved the maximum values in these cases. The optimal thresholding coefficient values are 1.4 and 1.2 for gene expression profiles of patients, who were investigated on diseases Alzheimer and Parkinson respectively. In these cases, Shannon Entropies for the filtered data are minimal that indicates the maximum informativity of the studied gene expression profiles.

The next stage of data processing is the reducing of gene expression profiles based on fuzzy logic system with the use of statistical criteria and Shannon entropy. The variance, average and Shannon entropy of gene expression profiles were used as input variables. The quality of gene expression profiles was used as output parameter. The range of the output parameter change was divided into five equal intervals (very low, low, median, high, very high). The genes, which were indicated as very high by quality parameter, were allocated for the following investigation. The Gaussian and triangular membership functions were used for the input and output variables respectively. Fig. 8 and Fig. 9 show the simulation results for the determination of the input parameters boundary values.

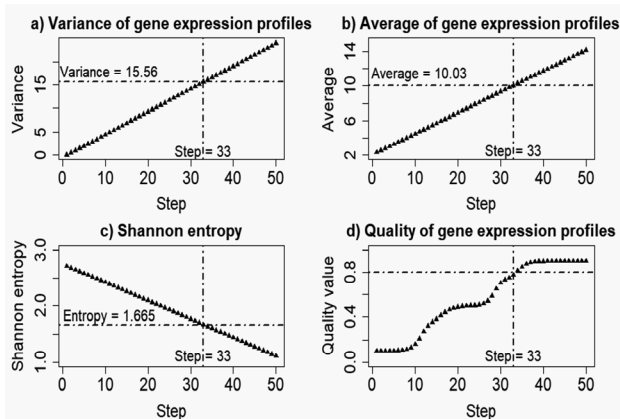


Fig. 8. Results of the simulation to determine the input parameters boundary values in the case of Alzheimer disease

As result of the simulation, the number of genes was changed from 54675 to 2037 in the case of Alzheimer disease and from 54675 to 1979 in the case of Parkinson disease. The implementation of the following cluster-bicuster technology was performed in three steps. Firstly, the gene expression profiles which were identified as noise

were removed from the data with the use of DBSCAN clustering algorithm within the framework of objective clustering inductive technology. The number of genes was changed from 2037 to 1771 in the case of Alzheimer disease and from 1979 to 1649 in the case of Parkinson disease. Then, the obtained data were divided into two clusters using SOTA clustering algorithm. 770 and 1001 genes were in the first and in the second clusters respectively in the case of Alzheimer disease. Clusters in the case of Parkinson disease contained 606 and 1043 genes respectively. Finally, five biclusters were obtained from each cluster using biclustering technology, which is presented in Fig. 5.

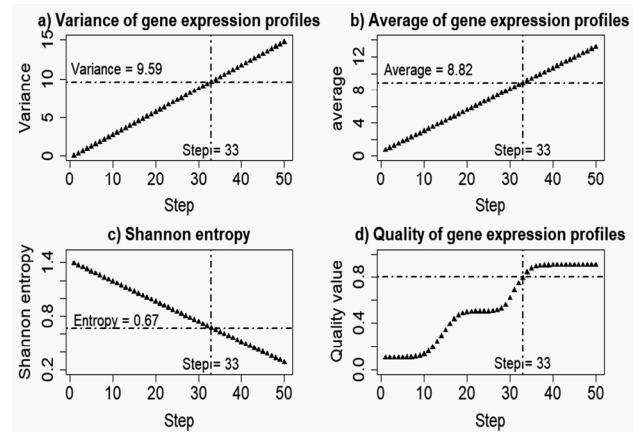


Fig. 9. Results of the simulation to determine the input parameters boundary values in the case of Parkinson disease

Gene regulatory networks reconstruction and their validation was performed with the use of the technology, which is described in detail in [20]. Fig. 10 and Fig. 11 present the results of the obtained model validation. The analysis of the obtained results allows us to conclude about the high efficiency of the proposed technology of gene expression profiles processing because the value of the relative validity criterion is high ($\gg 1$) for all obtained models. This fact indicates a high present of coincidence of appropriate genes interconnection in basic gene network and networks based on the obtained biclusters.

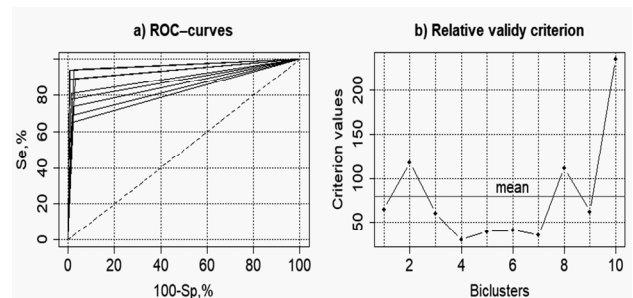


Fig. 10. Results of the gene networks validation in the case of Alzheimer disease

III. CONCLUSION

The results of the practical implementation of the information technology of gene expression profiles processing for the purpose of the gene regulatory networks reconstruction and validation are presented in the paper. The patients' DNA microchip data, which were investigated on

Alzheimer and Parkinson diseases were used during simulation process. The step-by-step procedure of the studied data processing included: determination of the optimal combination of the methods for evaluation of the genes expression array at the first step, determination of the wavelet filter optimal parameters and filtration the studied profiles at the second step, genes reducing, clustering, biclustering, and gene network reconstruction and validation at the last step. The obtained results have shown high efficiency of the proposed technology. The perspective of the authors' research is the implementation of the proposed technology for reconstruction of different types of gene regulatory networks.

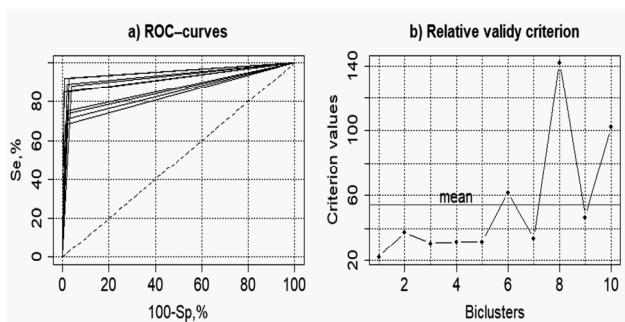


Fig. 11. Results of the gene networks validation in the case of Parkinson disease

REFERENCES

[1] D. Zak, R. Vadigepalli, E. Gonye, F. Doyle, et al. "Unconventional systems analysis problem in molecular biology: a case study in gene regulatory network modeling," *Computational and Chemical Engineering*, 29(3), pp. 547-563, 2005.

[2] M. Schena, and R. W. Davis. "Microarray biochip technology," Eaton Publishing, pp. 1-18, 2000.

[3] J. M. Heather, and B. Chain, "The sequence of sequencers: The history of sequencing DNA," *Genomics*, vol. 107, pp. 1-8, 2016.

[4] G. Setlak, Y. Bodyanskiy, I. Pliss, O. Vynokurova, D. Peleshko, and I. Kobylin, "Adaptive fuzzy clustering of multivariate short series with unevenly distributed observations based on matrix neuro-fuzzy self-organizing network," *Advances in Intelligent Systems and Computing*, 643, pp. 308-315, 2018.

[5] Y. Bodyanskiy, O. Vynokurova, V. Savvo, T. Tverdokhlib, and P. Mulesa, "Hybrid clustering-classification neural network in the medical diagnostics of the reactive arthritis," *International Journal of Intelligent Systems and Applications*, 8 (8), pp. 1-9, 2016.

[6] Y. Bodyanskiy, O. Tyshchenko, and D. S. Kopaliani, "An evolving connectionist system for data stream fuzzy clustering and its online learning," *Neurocomputing*, 262, pp. 41-56, 2017.

[7] Z. Hu, Y. Bodyanskiy, O. Tyshchenko, and O. Boiko, "A neuro-fuzzy Kohonen network for data stream possibilistic clustering and its online self-learning procedure," *Applied soft computing*, 2017.

[8] V. Mashkov, "Task allocation among agents of restricted alliance," *Eighth IASTED International Conference on Intelligent Systems and Control, ISC 2005*, pp. 13-18, 2005.

[9] V. A. Mashkov, and O. V. Barabash, "Self-testing of multimodule systems on optimal check-connection structures," *Engineering Simulation*, 13 (3), pp. 479-492, 1996.

[10] P. Bidyuk, A. Gozhyj, I. Kalinina, V. Gozhyj. Analysis of uncertainly types for model building and forecasting dynamic processes, *Advances in Intelligent Systems and Computing*, 689, pp. 66-78, 2018.

[11] A. Gozhyj, I. Kalinina, and V. Gozhyj, "Fuzzy cognitive analysis and modeling of water quality," *IEEE 9th International Conference on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications, IDAACS 2017*, 1, art. no. 8095092, pp. 289-293, 2017.

[12] P. Bidyuk, A. Gozhyj, I. Kalinina, and V. Gozhyj, "Methods for processing uncertainties in solving dynamic planning problems," *12th International Scientific and Technical Conference on Computer Sciences and Information Technologies, CSIT 2017*, 1, art. no. 8098757, pp. 151-155, 2017.

[13] A. Mukhopadhyay, U. Maulik, and S. Bandyopadhyay, "On biclustering of gene expression data," *Current Bioinformatics*, vol. 5, pp. 204-216, 2010.

[14] B. Pontes, R. Giráldez, and J. S. Aguilar-Ruiz, "Biclustering on expression data: A review," *Journal of Biomedical Informatics*, vol. 57 pp. 163-180, 2015.

[15] J. Hausser and K. Strimmer, "Entropy inference and the James-Stein estimator with application to nonlinear gene association networks," *Journal of Machine Learning Research*, vol. 10, pp. 1469-1484, 2009.

[16] S. Babichev, V. Lytvynenko, M. Korobchynskiy, and M. A. Taif, "Objective clustering inductive technology of gene expression sequences features," *Communications in Computer and Information Science*, vol. 716, pp. 359-372, 2016.

[17] S. Babichev, V. Lytvynenko, J. Skvor, and J. Fiser. "Model of the objective clustering inductive technology of gene expression profiles based on SOTA and DBSCAN clustering algorithms," *Advances in Intelligent Systems and Computing*, vol. 689, pp. 21-39, 2018.

[18] S. Babichev, M. A. Taif, V. Lytvynenko, and V. Osypenko, "Critical analysis of gene expression sequences to create the objective clustering inductive technology," *IEEE 37th International Conference on Electronics and Nanotechnology, ELNANO 2017*, pp. 244-248, 2017.

[19] S. Kaiser, *Biclustering: Methods, Software and Application*, Minchin, 2011.

[20] S. Babichev, M. Korobchynskiy, O. Lahodynskiy, O. Korchomnyi, and V. Borynskiy, "Development of a technique for the reconstruction and validation of gene network models based on gene expression profiles," *East-European journal of enterprise technologies*, vol. 1/4 (91), pp. 19-32, 2018.

[21] W. S. Liang, E. M. Reiman, J. Valla, T. Dunckley, et al. "Alzheimer's disease is associated with reduced expression of energy metabolism genes in posterior cingulate neurons," *Proc. Nat. Acad. Sci. USA*, vol. 105(11), pp. 4441-4446, 2008.

[22] B. Zheng, Z. Liao, J. J. Locascio, K. A. Lesniak, et al. "PGC-1 α , a potential therapeutic target for early intervention in Parkinson's disease," *Sci. Transl. Med.*, vol. 2(52), pp. 52-73, 2010.

A New Approach for Forming a Probabilistic Risk Assessment Model of Innovative Project Implementation Under Risk

Ali Rekik
 Computer Science Department
 Higher Institute of Computer Sciences
 Medenine, Tunisia
 alirekik1@yahoo.com

Nissen Masmoudi
 Computer Science Department
 Higher Institute of Technological Studies
 Sfax, Tunisia
 nissen.masmoudi@gmail.com

Abstract— In this paper we propose a new approach for constructing a probabilistic risk assessment model of innovative project. The method is based on the task of comparison and ranking of fuzzy numbers that has an important role in more applications related to the decision analysis. In the literature there are many approaches to compare fuzzy numbers. The majority of these approaches are based on the quantitative measurements. For that, we propose a new method to calculate the risk level that we can accord to equality $X < Y$ when X and Y are two generalized fuzzy numbers.

Keywords—Innovative project selection; fuzzy logic; Risk; Uncertainty.

I. INTRODUCTION

The risk assessment of innovative project is related to the optimization task under uncertainty and risk [5]. To invest in innovative projects, companies must have a various strategies; these strategies must reach an effective level of coherence through a variety of decisions [4, 7]. Among the various models of innovative project assessment, we can distinguish those based on linear, nonlinear, dynamic, stochastic, multicriteria decision support system [2], and fuzzy programming [9]. The fuzzy sets theory is used to represent uncertain information in multiple systems [8], such as planning support systems and decision support in the innovative project assessment [6, 7]. Risk assessment is a main element in project success and should be integrated in all innovative projects. Furthermore, exist a great link between project risk assessment and a project's success [1, 3]. To deal with the risk rates of innovative projects, decision makers must use specific methods and techniques that will allow them to assess and manage these risks effectively. In this paper we build a fuzzy probabilistic approach to assess a risk related to innovative project task, after that we provide a numerical example to describe the results of the proposed approach.

II. THE TASK OF BUILDING A PROBABILISTIC RISK ASSESSMENT MODEL OF INNOVATIVE PROJECT

As a risk assessment of innovative project implementation it is advisable to take the estimated parameter of project's profitability P_{Prof} and their rate of return value Q_{RR} .

In the case when $P_{Prof} < Q_{RR}$, the implementation of the innovative project is considered inappropriate. The values of

P_{Prof} and Q_{RR} are given in the triangular fuzzy numbers form $P_{Prof} = [P_{min}, P_0, P_{max}]$ and $Q_{RR} = [Q_{min}, Q_0, Q_{max}]$. Their membership functions are respectively represented as follow: [1-3]

$$\mu_{P_{Prof}}(x) = \begin{cases} \frac{1}{P_0 - P_{min}}x + \frac{P_{min}}{P_{min} - P_0}, & P_{min} < x < P_0; \\ \frac{1}{P_0 - P_{max}}x + \frac{P_{max}}{P_{max} - P_0}, & P_0 < x < P_{max}; \\ 0, & (x < P_{min}) \vee (x > P_{max}). \end{cases} \quad (1)$$

$$\mu_{Q_{RR}}(x) = \begin{cases} \frac{1}{Q_0 - Q_{min}}x + \frac{Q_{min}}{Q_{min} - Q_0}, & Q_{min} < x < Q_0; \\ \frac{1}{Q_0 - Q_{max}}x + \frac{Q_{max}}{Q_{max} - Q_0}, & Q_0 < x < Q_{max}; \\ 0, & (x < Q_{min}) \vee (x > Q_{max}). \end{cases} \quad (2)$$

When we build the graphs of $\mu_{P_{Prof}}(x)$ and $\mu_{Q_{RR}}(x)$ in one coordinate system, depending on the current values $[Q_1(\alpha), Q_2(\alpha)]$ and $[P_1(\alpha), P_2(\alpha)]$, we have various possible arrangements of graphs prescribed functions in relation to each other. The general scheme of reasoning used in the present method does not depend on the location of triangular numbers $\overline{P_{Prof}}$, $\overline{Q_{RR}}$, therefore, we will consider in more detail one of the variants, shown in Figure 1.

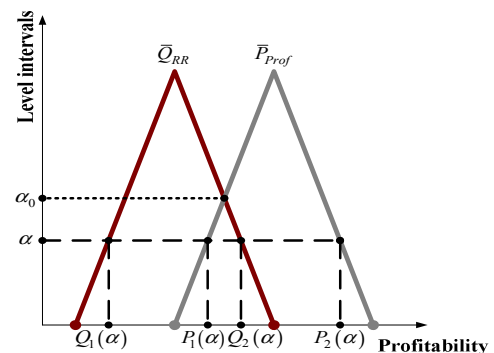


Fig. 1. The dependence between the values $\overline{P_{Prof}}$ and $\overline{Q_{RR}}$ in α -level intervals

These two membership functions intersect at the point with the ordinate α_0 . When $\alpha \geq \alpha_0$ and $P_{prof} > Q_{RR}$, the α -level sets do not intersect, and the risky zone is absent. When $\alpha < \alpha_0$ there is a risk that the value of P_{prof} , included in the intersection of the intervals $[Q_1(\alpha), Q_2(\alpha)]$ and $[P_1(\alpha), P_2(\alpha)]$, may be less than the value of Q_{RR} , that is to say the $[P_1(\alpha), Q_2(\alpha)]$ interval is the risk zone. We conclude that if $0 \leq \alpha \leq \alpha_0$ the α -level sets intersect. By shifting each selected α level in the (P, Q) -plane, we obtain the results shown in the Figure 2. The shaded area of inefficient investments is limited by the straight lines $P_{prof} = P_{prof_1}$, $P_{prof} = P_{prof_2}$, $Q_{RR} = Q_{RR_1}$, and the bisector of the quadrangle angle $P_{prof} = Q_{RR}$ as shown on Figure 2.

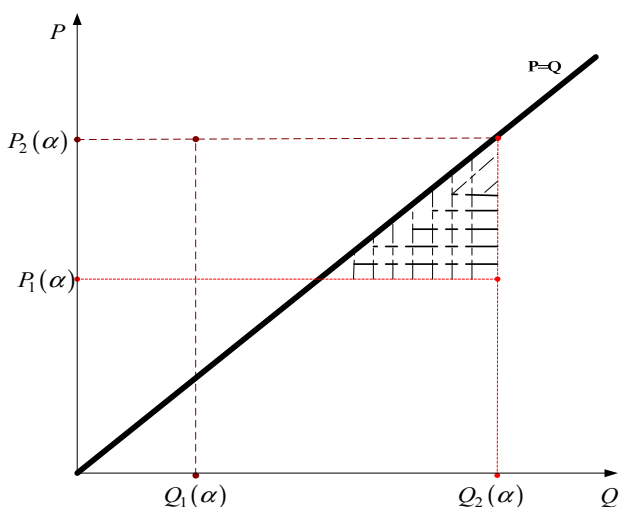


Fig. 2. The result of the transition from α -level sets to the (P, Q) -plane for one selected criterion.

In Figure 2 the shaded area indicates the risk area, and the entire rectangle is the range of possible implementations of the selected parameter. For a selected α -level, the probability of hitting the point with the current coordinates (P, Q) in the shaded area represents the probability of an insufficient level of profitability for a given pair of values.

We denote this probability by $P(\alpha)$. Then $P(\alpha)$ is determined according to the expression (3) and the graph of the function shown in Figure 3.

$$P(\alpha) = \frac{S_1(\alpha)}{S_2(\alpha)}$$

where:

$S_1(\alpha)$ is the shaded area;

$S_2(\alpha)$ is the rectangular area.

If we express the area $S_1(\alpha)$ through in explicit form, after elementary transformations we obtain the following expressions:

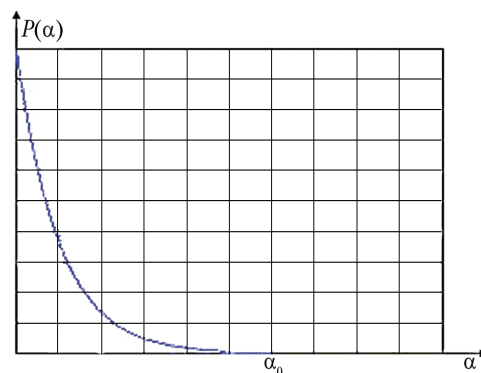


Fig. 3. The probability of hitting the point with the current coordinates (P, Q) in the shaded area of the selected α level

$$S_1(\alpha) = \begin{cases} 0, & \alpha \in I_1; \\ \frac{(P_2(\alpha) - Q_1(\alpha))^2}{2}, & \alpha \in I_2; \\ \frac{(P_1(\alpha) - Q_1(\alpha) + (P_2(\alpha) - Q_1(\alpha))) \times (P_2(\alpha) - P_1(\alpha))}{2}, & \alpha \in I_3; \\ \frac{(P_2(\alpha) - Q_2(\alpha) + (P_2(\alpha) - P_1(\alpha))) \times (Q_2(\alpha) - Q_1(\alpha))}{2}, & \alpha \in I_4; \\ \frac{(P_2(\alpha) - P_1(\alpha))(Q_2(\alpha) - Q_1(\alpha)) - \frac{(Q_2(\alpha) - P_1(\alpha))^2}{2}}{2}, & \alpha \in I_5; \\ (P_2(\alpha) - P_1(\alpha))(Q_2(\alpha) - Q_1(\alpha)), & \alpha \in I_6; \end{cases} \quad (3)$$

$$I_1 = \{\alpha \mid Q_1(\alpha) > P_2(\alpha)\} = [\alpha_4, 1];$$

$$I_2 = \{\alpha \mid P_1(\alpha) < Q_1(\alpha) < P_2(\alpha) \leq Q_2(\alpha)\} = [\alpha_3, \alpha_4];$$

$$I_3 = \{\alpha \mid Q_1(\alpha) < P_1(\alpha) < P_2(\alpha) \leq Q_2(\alpha)\} = [\alpha_2, \alpha_3];$$

$$I_4 = \{\alpha \mid P_1(\alpha) < Q_1(\alpha) < Q_2(\alpha) \leq P_2(\alpha)\} = [\alpha_1, \alpha_2];$$

$$I_5 = \{\alpha \mid Q_1(\alpha) \leq P_1(\alpha) \leq Q_2(\alpha) \leq P_2(\alpha)\} = [\alpha_0, \alpha_1];$$

$$I_6 = \{\alpha \mid Q_2(\alpha) \leq P_1(\alpha)\} = [0, \alpha_0].$$

It is understood that the area depends on the relative position of intervals $[Q_1(\alpha), Q_2(\alpha)]$ and $[P_1(\alpha), P_2(\alpha)]$.

For each point with the coordinates (P, Q) belongs to the shaded area, it represents the probability of an insufficient profitability level for a pair of values.

Since all variants (P, Q) are equally possible at the set level of belonging α , the extent of the risk of project inefficiency $P(\alpha)$ is a geometrical probability of finding a point (P, Q) in the zone of inefficient investments:

$$P(\alpha) = \frac{S_1(\alpha)}{(P_2(\alpha) - P_1(\alpha))(Q_2(\alpha) - Q_1(\alpha))} \quad (4)$$

In the proposed method, as a risk assessment we take the risk probability value $P(\alpha)$ in a point of interest. Thus, for each value of α has its own risk.

However, the risk for a specific values of α cannot describe the value of a lack of profitability in general, because $P(\alpha)$ have a local characteristic.

Therefore, it is appropriate to introduce the value of maximum risk of profitability, ie.

$$R_{\max} = \max_{0 < \alpha < 1} P(\alpha) = \frac{(Q_{\max} - P_{\min})^2}{2(Q_{\min} - Q_{\max})(P_{\min} - P_{\max})}. \quad (5)$$

Maximal risk does not depend on the values of α , but depends exclusively on the parameters $P_{\min}, P_{\max}, Q_{\min}, Q_{\max}$. We deduce that, the degree of risk is determined by formulas (3) and (4) for each α level can be represented as follow:

$$Risk = \int_0^1 p(\alpha) d\alpha.$$

$$\text{Where: } P(\alpha) = \frac{S_1(\alpha)}{(P_2(\alpha) - P_1(\alpha))(Q_2(\alpha) - Q_1(\alpha))}. \quad (5.1)$$

Since $S_1(\alpha)$ as we see in (3) is based on five conditions, and then $P(\alpha)$ also takes the following values:

$$P(\alpha) = \begin{cases} 0, & \text{for } P_{\alpha}^2 \leq Q_{\alpha}^1; \\ P_1 = \frac{(P_2(\alpha) - Q_1(\alpha))^2}{2(P_2(\alpha) - P_1(\alpha))(Q_2(\alpha) - Q_1(\alpha))}, & \text{for } P_1(\alpha) < Q_1(\alpha) < P_2(\alpha) \leq Q_2(\alpha); \\ P_2 = \frac{(P_1(\alpha) - Q_1(\alpha)) + (P_2(\alpha) - Q_1(\alpha))}{2(Q_2(\alpha) - Q_1(\alpha))}, & \text{for } Q_1(\alpha) \leq P_1(\alpha) < P_2(\alpha) \leq Q_2(\alpha); \\ P_3 = \frac{(P_2(\alpha) - Q_2(\alpha)) + (P_2(\alpha) - P_1(\alpha))}{2(P_2(\alpha) - P_1(\alpha))}, & \text{for } P_1(\alpha) \leq Q_1(\alpha) < Q_2(\alpha) \leq P_2(\alpha); \\ P_4 = 1 - \frac{(Q_2(\alpha) - P_1(\alpha))^2}{2(P_2(\alpha) - P_1(\alpha))(Q_2(\alpha) - Q_1(\alpha))}, & \text{for } Q_1(\alpha) \leq P_1(\alpha) \leq Q_2(\alpha) \leq P_2(\alpha); \\ P_5 = 1, & \text{for } Q_2(\alpha) \leq P_1(\alpha). \end{cases} \quad (6)$$

Should be noted that with a triangular fuzzy numbers P and Q , the function P cannot exist simultaneously in all intervals, and the integral will take the following form:

$$\int_0^1 P(\alpha) d\alpha = \int_{I_1} P_1(\alpha) d\alpha + \int_{I_2} P_2(\alpha) d\alpha + \int_{I_3} P_3(\alpha) d\alpha + \int_{I_4} P_4(\alpha) d\alpha + \int_{I_5} P_5(\alpha) d\alpha \quad (7)$$

In addition, we must formally express the function $P(\alpha)$ and find the $Q_1(\alpha)$, $Q_2(\alpha)$, $P_1(\alpha)$ and $P_2(\alpha)$ values depending on α as follow: $\alpha = a.Q_{\alpha}^1 + b$, and using the points $(Q_{\min}, 0)$ and $(Q_0, 1)$ of the line we can determine the coefficients a and b and therefore the value of Q_{α}^1 be as follow:

$$Q_{\alpha}^1 = \alpha(Q_0 - Q_{\min}) + Q_{\min}. \quad (8.1)$$

Similarly, we obtain the relation for Q_{α}^2 , P_{α}^1 and P_{α}^2 .

$$Q_{\alpha}^2 = Q_{\max} - \alpha(Q_{\max} - Q_0) \quad (8.2)$$

$$P_{\alpha}^1 = \alpha(P_0 - P_{\min}) + P_{\min} \quad (8.3)$$

$$P_{\alpha}^2 = P_{\max} - \alpha(P_{\max} - P_0) \quad (8.4)$$

By using the formulas (8.1, 8.2, 8.3, 8.4) producing the corresponding changes in the function (6) we can write the resulting expressions as follow:

$$\int_{\alpha_{j-1}}^{\alpha_j} P_j d\alpha, \quad j = \overline{1, 5}.$$

III. NUMERICAL EXAMPLE

We present an example of this model in the case of the project evaluation by one criterion shown in Fig. 4. Suppose that $\overline{P_{prof}} = [-800, 700, 1300]$ and $\overline{Q_{RR}} = [-220, 0, 280]$. Their membership functions and graphical presentation are represented as follow:

$$\mu_{P_{prof}}(x) = \begin{cases} \frac{x+800}{1500}, & -800 < x < 700; \\ \frac{-x+1300}{600}, & 700 < x < 1300; \\ 0, & (x < -800) \vee (x > 1300). \end{cases}$$

$$\mu_{Q_{RR}}(x) = \begin{cases} \frac{-x+280}{280}, & 0 < x < 280; \\ \frac{x+220}{220}, & -220 < x < 0; \\ 0, & (x < -220) \vee (x > 280). \end{cases}$$

For given fuzzy numbers μ_P and μ_R , the function $P(\alpha)$ exists only on three intervals. The first one is: $P_\alpha^1 < Q_\alpha^1 < Q_\alpha^2 < P_\alpha^2$ when $\alpha \in [0, \alpha_0]$, the second one is: $Q_\alpha^1 < P_\alpha^1 < Q_\alpha^2 < P_\alpha^2$ when $\alpha \in [\alpha_0, \alpha_1]$, and the last one is: $Q_\alpha^2 < P_\alpha^1$ when $\alpha \in [\alpha_1, 1]$. We must find the values of α_0 and α_1 . Equating the functions μ_P and μ_R at the corresponding intervals, we obtain the following result:

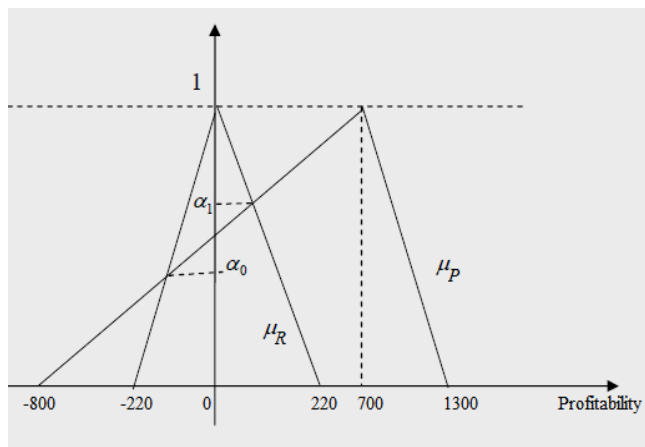


Fig. 4. Example of the values $\overline{P_{prof}}$ and $\overline{Q_{RR}}$ in α -level intervals

$$\alpha_0 = 0.45 \text{ then } P = Q = -115.$$

$$\alpha_1 = 0.65 \text{ then } P = Q = 120.$$

Based on these data, we calculate the risk degree by implementing the project as follow:

$$Risk = \int_0^{0.65} P(\alpha) d\alpha = \int_0^{0.45} P_2 d\alpha + \int_{0.45}^{0.65} P_1 d\alpha.$$

The risk degree of this project is: Risk = 0.144.

If you determine the risk of the project by an approximate method according (5.1) and (6) we obtained the following results:

TABLE I. RISK ASSESSMENT USING APPROXIMATE METHOD

Alpha	$P_1(\alpha)$	$P_2(\alpha)$	$Q_1(\alpha)$	$Q_2(\alpha)$	$P(\alpha)$
0	-800	1300	-220	280	0.401
0.1	-660	1250	-190	265	0.352
0.2	-506	1200	-162	243	0.310
0.3	-350	1121	-135	205	0.253
0.4	-205	1068	-119	178	0.174
0.5	-53	1005	-110	140	0.075
0.6	100	941	-75	115	0.018
0.7	240	880	-58	87	0.000
0.8	380	805	-37	55	0.000
0.9	538	743	-23	28	0.000

Based on these results we obtain the risk value:

$$Risk = 0.159.$$

The approximate method gives us a 10% higher risk rating than that used in our method.

IV. CONCLUSION

The application of the fuzzy set theory provides a new method for the risk assessment of innovative project. In this paper, we have developed a fuzzy approach to deal with risk by introducing an innovative project. As a result, the decision makers have now a better possibility for describing the information uncertainty in the project, by applying the fuzzy set theory. Consequently, the fuzzy sets allow the users to determine the project qualitative characteristics, and to transform them into a mathematical model. In conclusion, our proposed approach can describe the risk level in an uncertain environment.

REFERENCES

- [1] B. A. Perera, R. Rameezdeen, N. Chileshe, and M. R. Hosseini, "Enhancing the effectiveness of risk management practices in Sri Lankan road construction projects: A Delphi approach", *International Journal of Construction Management*, vol. 14, no.1, pp. 1–19, 2014.
- [2] C. Stummer, E. Kiesling, and W. J. Gutjahr, "A multicriteria decision support system for competence-driven project portfolio selection", *International Journal of Information Technology & Decision Making*, vol.8, pp. 379–401, 2009.
- [3] J. Teller, "Portfolio risk management and its contribution to project portfolio success: An investigation of organization, process, and culture", *Project Management Journal*, vol. 44, No.2, pp. 36–51, 2013.
- [4] G. P. Pisano, "Creating an R&D Strategy", *Harvard Business School Working Paper*, no. 12-095, April 24, 2012.
- [5] R. Ali, G. Mounir, V.E. Balas, and M. Nissen, "Fuzzy Evaluation Method for Project Profitability", *Advances in Intelligent Systems and Computing*. Springer International Publishing, pp. 17-27, 2017.
- [6] G. Mounir, R. Ali, T. Moncef, "Coordinated Resource Management Models in Hierarchical Systems", *International Journal of Advanced Computer Science and Applications*, Vol. 4, No. 2, pp. 105-109, 2013.
- [7] Q. Tian, J. Ma, and O. Liu, "A hybrid knowledge and model system for R&D project selection", *Expert Systems with Applications*, vol 23, pp. 265–271, 2002.
- [8] R. Ali, "Coordination mechanisms in complex hierarchical systems", *International Journal of Information and Decision Sciences.*, vol. 9, no. 3, pp. 297– 313, 2017.
- [9] W. G. Zhang, Y. J. Liu, and W. J. Xu, "A new fuzzy programming approach for multi-period portfolio optimization with return demand and risk control", *Fuzzy Sets and Systems*, vol. 246, pp. 107–126, 2014.

Managing of Change Streams in Projects of Development Distributed Information Systems

Viktor Morozov
Management Technology Department
Kyiv National Taras Shevchenko
University
Kyiv, Ukraine
knumvv@gmail.com

Olena Kalnichenko
Management Technology Department
Kyiv National Taras Shevchenko
University
Kyiv, Ukraine
kv_vl@ukr.net

Andrii Khrutba
Management Technology Department
Kyiv National Taras Shevchenko
University
Kyiv, Ukraine
hrutba.andrey@gmail.com

Grigory Steshenko
Management Technology Department
Kyiv National Taras Shevchenko
University
Kyiv, Ukraine
gmsteshenko@gmail.com

Iuliia Liubyma
Department of Business Administration
and Project Management
"KROK" University
Kyiv, Ukraine
Alary7@ukr.net

Abstract—The IT project models in the development of distributed information systems using cloud technologies are proposed for consideration. Concepts definitions, their relationship in the effects of turbulent dynamic changes are given. A process model has been proposed for a proactive approach for managing changes in IT projects.

Index Terms — cloud technologies, distributed information systems, IT projects, proactive management

I. INTRODUCTION

Cloud technology is relatively new and one of the most promising directions in development of modern information technologies. The rapid development of cloud computing, such as distributed data processing technologies, opens horizons and perspectives for creating new cloud-based service opportunities [1]. Recent trends in this area show that this information technology concept is both useful and relevant. It is considered as an effective tool to meet the contemporary tasks and challenges that emerge from trends of rapid development, globalization, the complicating of technology and the enhanced turbulence of the external environment.

Today, large corporate workloads actively transit to cloud solutions. According to estimates, in the next five years, 40% to 50% of corporate loads will be concentrated in cloud services, while now it is 15%. This indicates an increase in the demand for cloud services and a change in the information policy paradigm in enterprises [2].

Along with cloud technologies, experts also mention other important aspects of the improvement and development of information technologies, for example, in the development of technologies in the field of analytics of large volumes of data and the integration of mobile devices and technologies of social networks into the corporate environment [3].

International research and consulting company IDC combines all these directions into the concept of "third platform" [4]:

First platform: mainframes and terminals are the backbone of thousands of applications and programs, and it involves thousands of users.

Second platform: traditional personal computers, Internet and Web technologies, client-server software architecture - hundreds of thousands of applications. It covers millions of users.

Third platform: large volumes of information, mobile devices, cloud computing, social technologies. Covers billions of users.

The development of the third platform is expected to lead to significant changes in business models in the various sectors of the information technology market in the near future [2].

II. ANALYSIS OF RESEARCH AND PUBLICATIONS

The issue of cloud technologies application in the creation of modern IT was considered in the works of M.S. Kosyakov [5], A. Alpatov [6] and other domestic scientists. Their generalized purpose-oriented application is to consider the processes of functioning of IT in organizations, that allow to significantly save finances by integrating the functions of various IT systems while achieving a spatially-temporal balancing of business processes in order to obtain a positive synergistic effect.

The problems of the project-oriented approach based on IT projects are reflected in the works of many domestic scientists. A lot of works by Ukrainian scientists are devoted to the problems of using the project approach to IT creation: S. Bushuev [7], A. Biloschickij [8], Yu. Teslia [9], I. Chumachenko [10], V. Morozov [11] et al. The issues of proactive project management are considered in: [12], N. Bushueva [13], et al. In their works, the methodology of project management (in particular IT projects) is evolving. Its aimed at mutually coordinated management of the full range of management processes in the space of the organizations themselves and their development projects with the goal of effectively implementing the entire portfolio of projects.

However, the possibility of proactive management in IT projects have not been sufficiently researched. Especially in the aspects that would allow to take into account the complex dynamic influences of the turbulent environment on the processes of product creation and project management.

The **purpose** of the article is to substantiate and develop a structural model of components for the creation of distributed information systems using cloud technologies and design approach that would take into account responses to dynamic changes and turbulence. It is also building a process model for management these changes in the development of modern IT.

The main tasks of this research include the following:

1. Identification of the control elements that make up the elements' basis of the processes of creating distributed information systems.
2. Design of conceptual model of projects of creating complex IT products.
3. Investigate the impact of changes at interaction with the turbulent environment of the project during its implementation.
4. Conduct a formal description of the elements and build a mathematical model of the process of creating complex IT products.
5. To research the obtained models with the help of special information technologies for project management.

III. MODELS OF PROJECT INTERACTION

The implementation of integrated solutions for the different types of cloud environments (private, public, hybrid) and other third-platform directions involve the implementation of high-complexity IT projects.

Increased complexity is determined by the complexity of hardware, software, methodologies and tools, technologies, and the specificity of distributed projects on information systems creation.

Considering the pace and directions of information technologies development, which is determined by the trends in science, technology, society and so on, it is necessary to consider change management issues from a proactive perspective. This approach is imposed by many changes that occur in both external and internal environment of the IT project. This will allow us to focus on ways to minimize problems, errors, and incidents in the provided solutions performance and improve the quality of the IT projects.

In the context of the objective, it is reasonable to identify the main influences on the backbone elements in building of models of proactive approaches. To build these models, there should be defined the basic concepts including the "distributed information system", "cloud technology," and "distributed project".

A *Distributed Information System* (DIS) is a set of interconnected autonomous computers or processors and a set of independent processes (system executable program components) interacting through the transmission of messages for data exchange and coordination of actions [19].

Cloud Technologies – the concept of providing upon request (application) the network access to distributed computer resources (computing networks, servers, data warehousing, software, network services, etc.), allocated regardless of the time of day, and the channel of access to the computer network [5].

To define the concept of "distributed project" and "distributed project management", [8] introduces the concept of a "distributed system". The understanding of the above terms is formed through the principles of the system approach. The "distributed system" means a system for which the relation of the element positions is significant in terms of its functioning, analysis and synthesis. However, the concept of a geographically distributed computing environment is present (for example, the Internet, a banking network with subsidiaries in different countries or a corporate network). It represents a system with geographically distributed elements. The emphasis on the physical location of the elements of such a system is somewhat blurred by the focus on transparency of information systems.

Based on the above, we understand that a distributed IT project is a system/set of interrelated distributed (over time, territory, function, etc.) processes and distributed resources that function in a dynamic turbulent environment. For such projects, all the properties and patterns of regular projects remain relevant and existing management methodologies are applicable [6, 7].

Because cloud services are a combination of existing information technology solutions that are mutually integrated and have a spatially distributed infrastructure, they can be treated as DIS. Therefore, the activities related to the development of cloud-based information systems, regardless of the level of complexity and the range of the solved tasks for such system, involve the implementation of projects and are determined by characteristics of the project approach. Thus, the integration of the problems and challenges of the technological aspects of the subject area and the project management is inevitable.

Figure 1 shows the proposed platform for implementing distributed IT projects, which integrates existing technologies that synthesizes new solutions, allowing to achieve synergy when using proactive influence at the same time on product components and project elements.

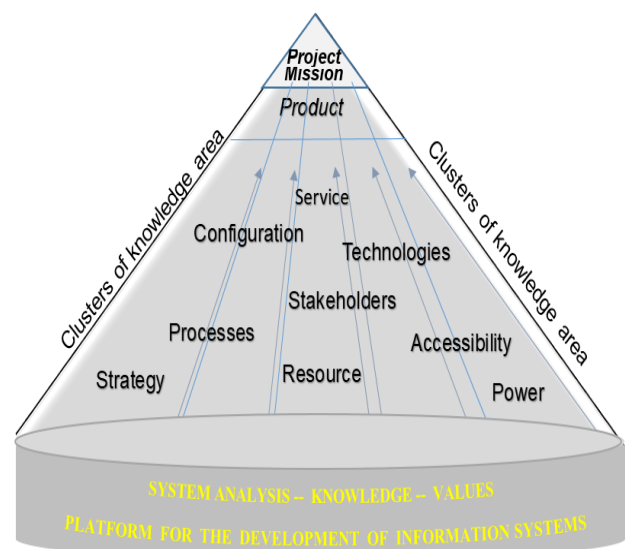


Fig. 1. "Cone" model for distributed information systems IT projects

Let us consider the components of the proposed model.

System analysis is used as a tool for investigation of the DIS as a complex adaptive system.

Convergence is now a trend in the IT sphere. It is implemented through the sharing of known and new technologies. Using convergence allows to get unexpected solutions, create new prospects and so-called "blue oceans" in IT business, creating new niches in the marketplace. Any "cloud" service is a convergent service that integrates telecommunication technologies (Internet access, network infrastructure, billing, etc.) and IT (implementation on the server of application functional, supporting service technologies of data centers, Internet protocols, etc.).

Knowledge is subjective models of action that are constantly constructed (recalled and/or generated) by a human adequately to a certain situation of life as a result of cognitive activity.

Values are the properties of products/services/phenomena demonstrated through their relevance, usefulness, or importance. The values are characterized by temporal factors and subjectivity.

As Figure 1 shows, the primary objective (mission) of such IT projects is to obtain the necessary product properties (DIS), which form the value of the product. The project mission can be achieved by identifying the directions and trajectory of the mission within the limited clusters of the knowledge areas of the project space realization.

The *strategy* means the directions and methods of the system management by running the project and the product creation system, taking into account the interaction with other available components of the "cone" model.

The product creation and IT project management processes are presented as the *processes* in the "cone" model.

A lot of *resources* in the "cone" model include existing and attracted (purchased) within the project framework material, software, labour, and information resources, which materialize in the final product and form the new desired value in the form of the IT project product (DIS).

Stakeholders [6], which have a direct influence not only on the functionality of the future product of the project, but also on the success of the whole project, significantly influence on the success of the IT project. This can include representatives of the client or customer, IT product users, vendors, developers, management commands, etc.

For IT projects, creating complex IT products is characterized by a service component that requires a strategy service, development services, transitions, operations, and continuous improvements [14].

We've already talked about using *technologies* to create an IT product. Additionally, this group also includes technologies for development, management, testing, operating, and maintaining the IT project product.

Components of IT product *configuration* [14, 15] are related to determining of the elements, their parameters and relationship to the developed information system. The same is relevant to project elements (Fig. 1) and the project environment.

Reviewing the project environment influence on its elements and success factors shows that it is the influence (or ignoring this influence) very often is the main threat of the project failure. The influence of the project's environment often leads to dynamic changes. The selective accounting of

these changes results in changes to the parameters and characteristics of virtually all elements within the "cone" model. At this, issues of proactive influence on the distributed information systems functioning remain relevant. This can be addressed by means of proactive approaches in projects on building such systems.

The product creation and IT project management processes are presented as the processes in the "cone" model.

$P = \{P^P, P^S\}$, where P^P – is the set of project management processes, $P^P = \{p_1^P, p_2^P, \dots, p_i^P\}$, here i – is the number of processes associated with IT project management; P^S – is the set of product creation processes, $P^S = \{p_1^S, p_2^S, \dots, p_j^S\}$, here j – is the number of processes associated with IT product creation.

A lot of resources in the "cone" model include existing and attracted (purchased) within the project framework material, software, labour, and information resources, which materialize in the final product and form the new desired value in the form of the IT project product (DIS).

$R = \{R^M, R^H, R^P, R^I\}$, where R^M – is the set of material resources involved in the project, $R^M = \{r_1^M, r_2^M, \dots, r_l^M\}$, here l – is the number of kinds of material resources required in the project; R^H – is the set of human resources involved in the project, $R^H = \{r_1^H, r_2^H, \dots, r_k^H\}$, here k – is the number of kinds of human resources involved in the project; R^P – is the set of program resources used in the project, $R^P = \{r_1^P, r_2^P, \dots, r_s^P\}$, here s – is the number of kinds of program resources involved in the project; R^I – is the set of information resources in the project, $R^I = \{r_1^I, r_2^I, \dots, r_z^I\}$, here z – is the number of kinds of information resources involved in the project.

Stakeholders [6], which have a direct influence not only on the functionality of the future product of the project, but also on the success of the whole project, significantly influence on the success of the IT project. This can include representatives of the client or customer, IT product users, vendors, developers, management commands, etc.

$O = \{O^L, O^D\}$, where O^L – is the set of project participants (close environment of the project), $O^L = \{o_1^L, o_2^L, \dots, o_f^L\}$, here f – is the number of participants who are part of its close environment; O^D – is the set of stakeholders (long-range environment of the project), $O^D = \{o_1^D, o_2^D, \dots, o_h^D\}$, here h – is the number of stakeholders that relate to the long-range environment of the project.

For IT projects, creating complex IT products is characterized by a service component that requires a strategy service, development services, transitions, operations, and continuous improvements [21]. $S = \{S^I, S^B, S^{US}\}$, where

S^I – is the set of services supporting the IT infrastructure, $S^I = \{s_1^I, s_2^I, \dots, s_q^I\}$, here q – the number of services to support the IT infrastructure; S^B – is the set of services supporting the business application, $S^B = \{s_1^B, s_2^B, \dots, s_w^B\}$, here w – is the number of services to support business applications; S^{US} – is the set of services that provide user support, $S^{US} = \{s_1^{US}, s_2^{US}, \dots, s_v^{US}\}$, here v – is the number of services to support users.

We've already talked about using technologies to create an IT product. Additionally, this group also includes technologies for development, management, testing, operating, and maintaining the IT project product. $Z = \{Z, Z^M, Z^I\}$, where Z^C – is the set of technologies for developing and testing a project product, $Z^C = \{z_1^C, z_2^C, \dots, z_c^C\}$, here c – is the number of technologies used to create the project product; Z^M – is the set of project management technologies, $Z^M = \{z_1^M, z_2^M, \dots, z_u^M\}$, here u – is the number of technologies used to manage the project; Z^I – a set of technologies for implementing and maintaining a project product. $Z^I = \{z_1^I, z_2^I, \dots, z_y^I\}$, here y – is the number of technologies used to implement and maintain the project product.

Components of IT product configuration [19, 22] are related to determining of the elements, their parameters and relationship to the developed information system. The same is relevant to project elements (Fig. 1) and the project environment. $K = \{K^P, K^S, K^E, K^{DP}, K^{DS}\}$, where K^P – is the set of project parameters, $K^P = \{k_1^P, k_2^P, \dots, k_r^P\}$, here r – is the number of project parameters; K^S – is the set of project product parameters, $K^S = \{k_1^S, k_2^S, \dots, k_d^S\}$, here d – is the number of project product parameters; K^E – is the set of parameters of the external environment of the project, $K^E = \{k_1^E, k_2^E, \dots, k_\gamma^E\}$, here γ – is the number of parameters of the external environment of the project; K^{DP} – is the set of requirements for the project, $K^{DP} = \{k_1^{DP}, k_2^{DP}, \dots, k_o^{DP}\}$, here o – is the number of requirements (conditions, opportunities and constraints) that the project must meet; K^{DS} – is the set of requirements to the project product, $K^{DS} = \{k_1^{DS}, k_2^{DS}, \dots, k_p^{DS}\}$, here p – is the number of requirements (conditions, opportunities and constraints) that the project product must meet.

By power we mean the set $L = \{L^S, L^P, L^E\}$, where $L^S = \{l_1^S, l_2^S, \dots, l_\mu^S\}$ – set of values of throughputs of product

components, μ – the number of throughput capabilities of all components of the product; $L^P = \{l_1^P, l_2^P, \dots, l_\varphi^P\}$ – is the set of values of throughputs of project elements, φ – is the number of throughput capabilities of all project elements (product creation management processes and project management processes); $L^E = \{l_1^E, l_2^E, \dots, l_x^E\}$ – is the set of influence values (influence degrees) of the project's external environment, x – is the number of values of the influence of the project's close and long-range environment, as well as the parameters of the project's external environment.

Accessibility (when creating a product) is determined by a set of evaluations of the level of performance of functions and requirements (reliability, supportability, serviceability, performance, security) $A = \{A^R, A^N, A^O, A^P, A^S\}$, where $A^R = \{a_1^R, a_2^R, \dots, a_\omega^R\}$ – is the set of estimates related to reliability, ω – is the number of reliability estimates; $A^N = \{a_1^N, a_2^N, \dots, a_\sigma^N\}$ – is the set of support level estimates, σ – is the number of supportability estimates; $A^O = \{a_1^O, a_2^O, \dots, a_\partial^O\}$ – is the set of estimates of serviceability, ∂ – is the number of serviceability estimates; $A^P = \{a_1^P, a_2^P, \dots, a_\delta^P\}$ – is the set of performance estimates, δ – is the number of performance estimates; $A^S = \{a_1^S, a_2^S, \dots, a_\vartheta^S\}$ – is the set of security estimates, ϑ – is the number of security estimates.

Based on this, the mathematical description of the proposed "Cone" model can be represented as follows: $M = \{X, Y, H\}$, where $X = \{G, P, R, O, S, Z, K, L, A\}$ – is the set of input parameters of the model; $Y = \{C_p, T_p, Q\}$ – is the set of output parameters, on its basis we will determine the efficiency of IT project management processes, where C_p – is the planned cost of creating project elements, T_p – is the planned duration of the project life cycle (given) [6], O – is the quality of the project, determined by the quality of the final product and the quality of the project implementation processes; H – is the set of relation-channels between the elements of the IT project management model communication channels. $H = \{h_1, h_2, \dots, h_\varepsilon\}$, ε – is the number of direct links between all model elements and $\bar{H} = \{\bar{h}_1, \bar{h}_2, \dots, \bar{h}_\varepsilon\}$, ε – is the number of feedbacks between all elements of the model.

In this case, the set of output parameters of the project model can also be presented in the form $X = \{x_{i_1} | i_1 = 1, 2, \dots, N_1\}$, where N_1 – is the number of knowledge areas of model M_1 . Then the planned cost of the project will look like:

$$C_p = \sum_{i_1=1}^{N_1} \sum_{j_1=1}^{T_p} \sum_{i_2=1}^{\varepsilon} (C_1(x_{i_1}, t_{j_1}) + C_2(h_{i_2})), \quad (1)$$

on condition $\forall (x_{i_1} \in X) \cup (q_{i_1} \in Q) \exists t_{j_1} \in T_p, T_p \geq 0$ and $C_p \leq C_b, C_b \geq 0$, where C_b – is the budget cost of the project (investments), C_1 – is the cost function of creating input parameter elements from $\{X\}$ at time $t_{j_1} \in T_p, C_2$ – is the cost function of communication channels between the elements of the model from $\{X\}$.

Reviewing the project environment influence on its elements and success factors shows that it is the influence (or ignoring this influence) very often is the main threat of the project failure. The influence of the project's environment often leads to dynamic changes. The selective accounting of these changes results in changes to the parameters and characteristics of virtually all elements within the "cone" model. At this, issues of proactive influence on the distributed information systems functioning remain relevant. This can be addressed by means of proactive approaches in projects on building such systems.

The current conditions in which complex IT projects have to be developed and implemented are characterized by downturns, intermittent funding, turnover of core personnel, changes in technologies, changes in customers' preferences, changing in market conditions, consumers, users, etc. All this requires frequent changes not only in the "cone" model "base", but also in the knowledge clusters. Therefore, the influence of environment turbulence must be introduced in the "cone" model. At the "cone+" model receiving, the influence of this environment should be studied through the subsequent changes influence on all elements and characteristics of complex IT projects. For the successful project completion, all these changes need to be managed. Otherwise, frequent changes result in chaotic abnormal inconsistencies in the system elements resulting in its failure. Fig. 2 shows the proposed change management model for IT projects.

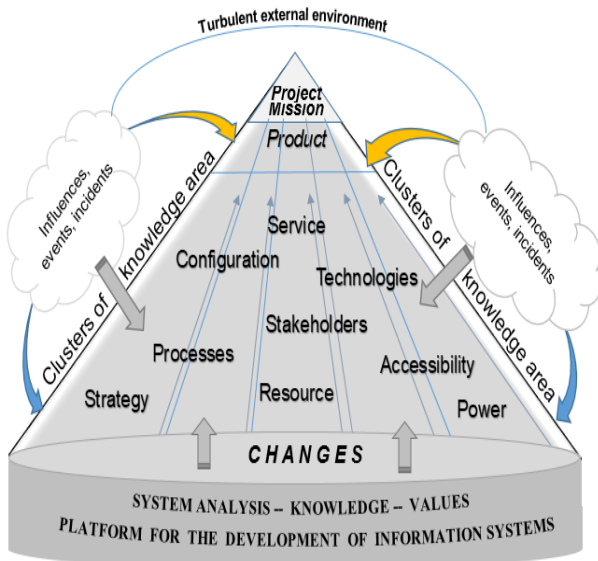


Fig. 2. "Cone+" model for investigation of the changes influence

A feature of the proposed model is a reviewing of the interaction of the product creation system and the necessary project (changes) management system under conditions of complex intersecting influences of the turbulent environment. The data of influence generates changes in individual model components resulting in changing the process of creating distributed information systems.

These changes can lead to problems and incidents at different stages of project implementation and product performance. A possible solution for avoiding problems or minimizing their influence may be early detection of specific signals of their occurrence.

The presence of the project's turbulent environment influence suggests: adding parameters of this influence into the proposed "cone" model and the reaction to it as control actions that ensure the stabilization of the model with the arising deviations.

Then the modified model of the project will have the following form: $M_2 = \{X, Y, Q, I, U, V\}$, where $I = \{I^E, I^O\}$ – is the set of influences on the project, I^E – is the set of environmental factors' influences, $I^E = \{i_1^E, i_2^E, \dots, i_e^E\}$, e – is the number of possible influences from the external environment of the project (political, economic, social, legal, environmental, technological aspects); I^O – is the set of influences of the project stakeholders, $I^O = \{i_1^O, i_2^O, \dots, i_b^O\}$, b – the number of possible influences from the long-range and close environment of the project (secondary and primary stakeholders); U – is the set of states of the IT project, $U = \{u_1, u_2, \dots, u_\beta\}$, β – is the number of possible states of the model due to the environmental impact and stakeholders of the project; V – is the set of reactions of the project to external influences, $V = \{v_1, v_2, \dots, v_a\}$, a – is the number of control actions, oriented on model's stabilization in the case of deviations of its parameters from given values.

Taking into account the influence of the external environment and stakeholders of the project, which lead to changes and deviations from the specified project parameters, it's possible to determine the actual cost of the project upon completion (C_f) and the actual completion time of the project (T_f):

$$T_f = T_p \pm (f_1(I) + f_2(U) + f_3(V)), \quad (2)$$

$$C_f = C_p \pm (C_2(I) + C_4(U) + C_5(V)) \quad (3)$$

where C_3, C_4, C_5 – are the actual costs of making changes due to the set of influences on the project, the monitoring of the set of IT project's states and the set of executable control actions, respectively; f_1, f_2, f_3 – functions for measuring the time intervals of the action of the set of influences on the project, the monitoring of the set of IT project's states and the set of executable control actions, respectively.

In this case, the target functions of the IT project management model can be represented as follows:

$$C_f - C_p = \pm \Delta C \rightarrow \min, \quad (4)$$

$$T_f - T_p = \pm \Delta T \rightarrow \min, \quad (5)$$

where $\Delta C, \Delta T$ – actual deviations in the cost and time of project execution, taking into account changes adding due to the set of influences and impacts of the environment.

IV. CONCLUSION

The proposed proactive approach for changes management in distributed information systems projects has allowed creating models displaying the key elements, which experience a constant impact of dynamic changes of turbulent environment.

For these changes management, the proposed structure of the IT project management processes based on the taken should be used. In addition, information exchange model of proactive management components should be proposed. Such processes in these systems should be further explored and elaborated with the construction of the appropriate mathematical apparatus.

The proposed change management technology has led to the construction of an algorithm that defines the response to their influence based on the identification of emerging events. Also, the changes consequences are accepted from studying the requests for changes.

REFERENCES

- [1] Cloud Terminology – Key Definitions, <https://www.getfilecloud.com/cloud-terminology-glossary/>
- [2] Cloud computing and emerging IT platforms: Vision, hype, and reality for delivering computing as the 5th utility, <http://www.sciencedirect.com/science/article/pii/S0167739X08001957>
- [3] B. Furht, A. Escalante, Handbook of Cloud Computing, <https://link.springer.com/book/10.1007/978-1-4419-6524-0>
- [4] O. Maimon, L. Rokach, Data Mining and Knowledge Discovery Handbook, 2005, <http://www.bookmetrix.com/detail/book/ae1ad394-f821-4df2-9cc4-cb8b93edf40>
- [5] M. S. Kosyakov, Introduction to distributed computing. SPb: Public Research Institute IST, 2014. (in Russian)
- [6] A. Alpatov, “Development of distributed technologies and systems,” Prospects for Science and Education, vol. 2 (14), 2015. (in Russian)
- [7] V. N. Burkov, S.D. Bushuev, Resources Management for distributed projects and programs: Monograph, Nikolayev: Publisher Torubara V.V., 2015.
- [8] A. Biloshchytskyi, A. Kuchansky, Yu. Andrashko, and S. Biloshchytska. “A method for the identification of scientists' research areas based on a cluster analysis of scientific publications,” in Eastern-European Journal of Enterprise Technologies, vol. 2, issue 89, no. 5, pp. 4-10, 2017.
- [9] Yu. Teslia, A. Khlevnyi, and I. Khlevna, “Control of informational impacts on project management,” 1th IEEE International Conference on Data Stream Mining & Processing, Lviv, Ukraine, pp. 387-391, 23-27 August, 2016.
- [10] I. Chumachenko, and V. Morozov, The Project Management: Process of Planning of Project Activities, [Textbook], K.: University of Economics and Law "Krok", 2014. (in Ukrainian).
- [11] V. Morozov, O. Kalnichenko, and I. Liubyma, “Proactive Project Management for Development of Distributed Information Systems,” 4th International Scientific and Practical Conference “Problems of Infocommunications. Science and Technology” (PIC S&T-2017), Kharkiv, Ukraine. 2017.
- [12] Proactive Project Management, <http://www.itexpert.ru/rus/ITEMS/200810062247/>
- [13] N. Bushueva, Models and Methods for Proactive Management of Organizational Development Programs: [Textbook]. K.: Scientific World, 2007. (in Russian).
- [14] Components ITIL. Edition 2011, <https://itsm365.ru/blog/chto-takoe-itilism/>
- [15] Practice Standard for Project Configuration Management. Newtown Square, Pa: Project Management Institute, 2007.

A Hybrid Neuro-Fuzzy Model for Stock Market Time-Series Prediction

Alexander Vlasenko
 Department of Artificial Intelligence
 Kharkiv National University of Radio Electronics
 Kharkiv, Ukraine
 alexander.vlasenko86@gmail.com

Nataliia Vlasenko
 Department of Informatics and Computer Engineering
 Simon Kuznets Kharkiv National University of Economics
 Kharkiv, Ukraine
 gorohovatskaja@gmail.com

Olena Vynokurova
 Kharkiv National University of Radio Electronics
 Kharkiv, Ukraine
 IT Step University
 Lviv, Ukraine
 vynokurova@gmail.com

Marta Peleshko
 Department of Monitoring and Fire Prevention
 Lviv State University of Life Safety
 Lviv, Ukraine
 marta.peleshko@gmail.com

Abstract— In this paper we propose a hybrid five-layer neuro-fuzzy model and a corresponding learning algorithm with application in stock market time-series prediction tasks. The key difference between classical ANFIS architecture and the proposed model is in the fourth layer – multidimensional Gaussian functions are used instead of polynomials in order to achieve better computational performance and representational abilities in processing highly nonlinear volatile data. The experimental results have shown the clear advantages of the described model and its learning.

Keywords— time series, neuro-fuzzy, membership function, Gaussian, prediction

I. INTRODUCTION

Time series prediction is the one of the most common complex practical problems which appears in various applied domains. Among different types of time series financial data (e.g. currency exchange, stock and derivatives market) could

be distinguished by their highly nonlinear, nonstationary, volatile and chaotic nature and high-frequency dynamics.

For decades' statistical models have dominated the field of quantitative time-series analysis. The most popular among them are Autoregressive Moving Average (ARMA), Autoregressive Integrated Moving Average (ARIMA) etc. However, such models require time-series to be stationary and this is achieved through different differencing techniques, effectiveness of which is disputed due to highly nonstationary nature of real life time-series [1].

Except classical statistical models different computational intelligence techniques have been applied to the time series forecasting problems. Among them neuro-fuzzy models have gained popularity due to their universal approximation abilities and robustness combined with good computational performance and inherited from ANNs learning capabilities. They have been successfully used in the many forecasting tasks in different domains e.g. electric networks loading [2, 3], finances [4, 5, 6, 7] and others.

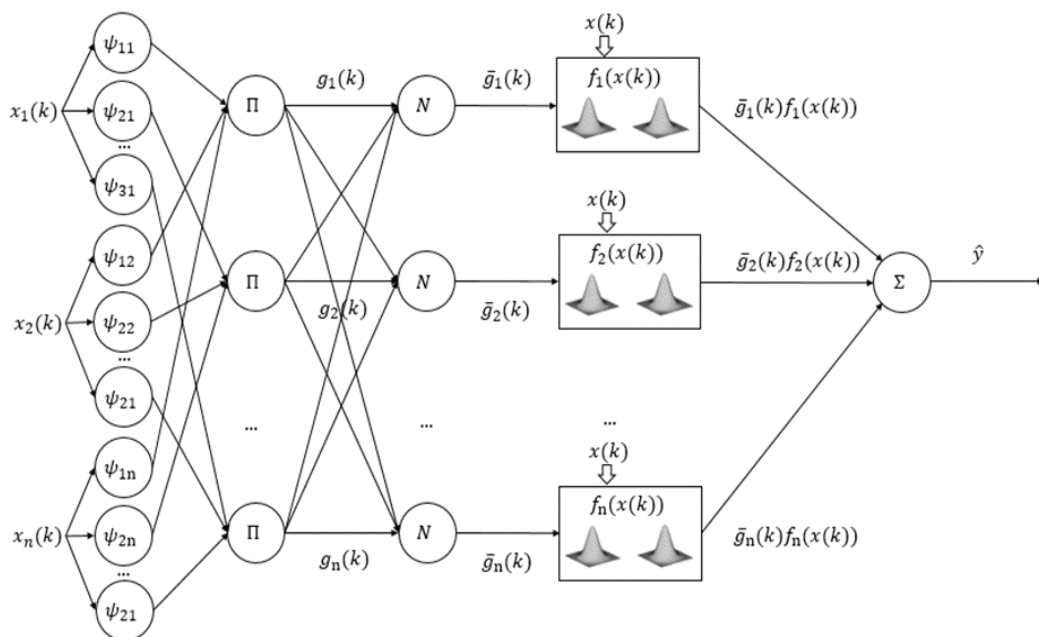


Fig. 1. The general architecture of the proposed model.

II. ARCHITECTURE AND INFERENCE

The proposed model is based on the classical ANFIS [8] model and comprises five layers. Fig.1 depicts a general architecture of the introduced model

The first layer is responsible for the fuzzification of the input variables represented by a n -dimensional vector $x(k) = (x_1(k), x_2(k), \dots, x_n(k))^T$. For the prediction task the input vector means the historical gap of the observed variable. Each input parameter is processed by the h^w membership functions, hence total amount of the first layer membership functions is $h^w \times n$. In current study we've used the Gaussian membership function, but in general case other common membership function types (e.g. triangular or generalized bell shaped) are acceptable. The Gaussian membership function has the following form:

$$\psi_{jl}(x_i(k)) = \exp\left[-\frac{(x_i(k) - c_{jl}^{\phi})^2}{2\sigma_{ji}^2(k)}\right] \quad (1)$$

Where $x(k)$ is an input vector, c_{jl}^{ϕ} – a centre of current Gaussian and σ_{ji} is a width parameter. The argument of the exponential function $-\frac{(x_i(k) - c_{jl}^{\phi})^2}{2\sigma_{ji}^2(k)}$ is a quadratic function of x_i . The main advantages of this membership function is relatively small amount of parameters and proneness to outliers.

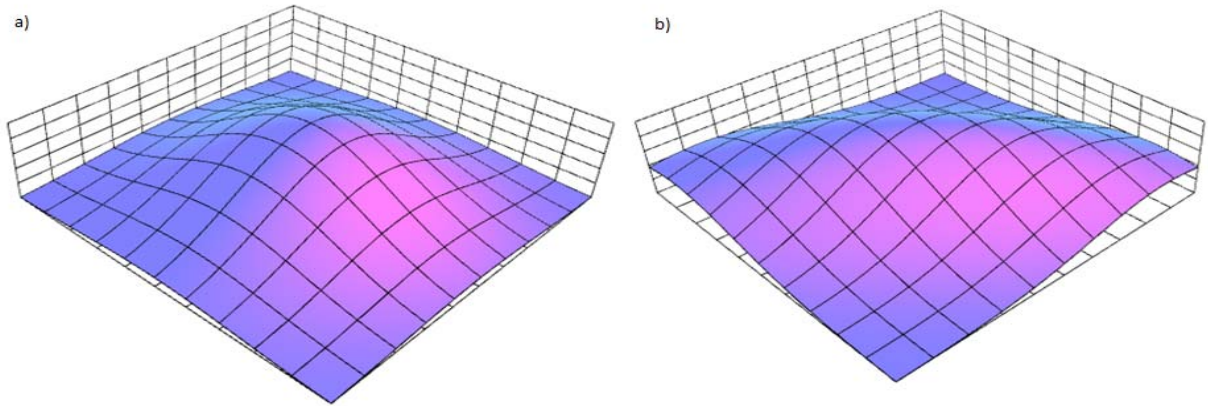


Fig. 2. The examples of multidimensional Gaussain with different covariance matrices: a) – with diagonal matrix; b – with matrix $\begin{bmatrix} 0.9 & 0.6 \\ 0.7 & 0.9 \end{bmatrix}$.

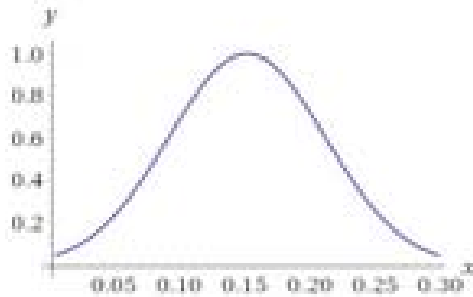


Fig. 3. The example of the first layer membership function with centre 0.15 and with width 0.06.

The second layer represents an aggregation of the antecedent premises values. It consists of h^w multiplier units which implement algebraic product fuzzy T-norm:

$$g_j(k) = \prod_{i=1}^n \psi_{jl}(x_i(k)) \quad (2)$$

The third layer is non-parametrised and responsible for normalization. It also has h^w units which output is computed by the following formula:

$$\bar{g}_j(k) = \frac{g_j(k)}{\sum_{j=1}^{h^w} g_j(k)} = \frac{\prod_{i=1}^n \psi_{jl}(x_i(k))}{\sum_{i=1}^{h^w} \prod_{i=1}^n \psi_{jl}(x_i(k))} \quad (3)$$

This is necessary for satisfying the Ruspini partitioning condition:

$$\sum_{j=1}^{h^w} \bar{g}_j(k) = 1 \quad (4)$$

The forth layer is presented by multidimensional Gaussian consequent functions $\varphi_{je}(x(k))$ and their weights p . Multidimensional Gaussians are used instead of the standard TSK/ANFIS polynomials:

$$\varphi_{je}(x(k)) = \exp\left[-\frac{(x(k) - c_{je}^{\phi})^T Q_{je}^{-1} (x(k) - c_{je}^{\phi}(k))}{2}\right] \quad (5)$$

where $x(k)$ is an input vector, c_{je}^φ – a vector which represents the centre of the current Gaussian and Q_{je} – covariance (receptive field) matrix, $Q_{je} \in S_{++}^n$. In this case a quadratic function from (1) becomes a quadratic form of the whole input vector $x(k)$.

Multidimensional Gaussian is a powerful tool to represent data which are not distributed evenly on the main axes. Hence the forth layer output is:

$$f_j(x(k)) = \sum_{e=1}^{h^\varphi} p_{je} \varphi_{je}(x(k)) \quad (6)$$

where h^φ is a number of multidimensional Gaussian functions for each unit f_j .

Fifth layer is non-parametrized and produces overall model output as a sum of its inputs:

$$\hat{y} = \sum_{j=1}^{h^\varphi} g_j f_j(x(k)) \quad (7)$$

In matrix form it could be rewritten as:

$$\hat{y} = p^T f(x(k)) \quad (8)$$

where $x(k)$ is an input vector, p -weights vector and $f(x(k))$ is a vector of normalised consequent functions values:

$$f(x(k)) = (\bar{g}_1 \varphi_{11}(x(k)) \dots \bar{g}_{h^\psi} \varphi_{h^\psi h^\varphi}(x(k))) \quad (9)$$

where h^ψ is a number of functions in the first layer and h^φ - number of multidimensional Gaussians in the fourth layer for each normalized output \bar{g}_1 .

III. LEARNING ALGORITHMS

Learning process in proposed model consists of adjusting weights vector p and tuning Multidimensional Gaussian functions parameters - both centres c_{jl}^φ and matrices Q_{jl} . First layer membership functions centres c_{jl}^φ are distributed equidistantly on initialization and they are not tuned during learning. Weights p are initialised randomly in range $[-0.1; 0.1]$ and their learning is achieved through the Kaczmarz iterative method:

$$p(k+1) = p(k) + \frac{y(k) - p^T f(x(k))}{f^T(x(k)) f(x(k))} f(x(k)) \quad (10)$$

where $p(k)$ is a weights matrix represented as a vector, $y(k)$ -reference signal, $p^T f(x(k))$ - overall model output according to (7).

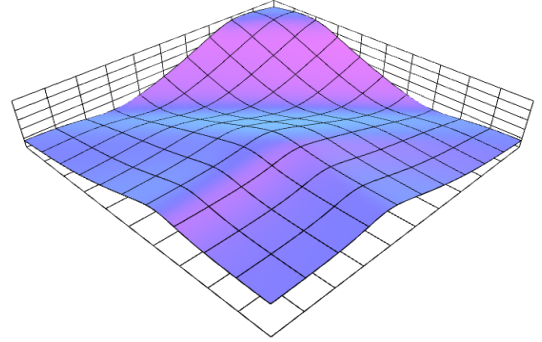


Fig. 4. The example of error surface in weights space.

Forth layer Gaussians learning performed by the first-order gradient backpropagation procedure based on the standard mean square error criterion:

$$E = \frac{1}{N} \sum_{k=1}^N (y(k) - \hat{y}(k))^2 \quad (11)$$

where $y(k)$ - reference signal value, $\hat{y}(k)$ - prognosis signal value and N is the training set length.

The Q_{jl}^{-1} matrices are initialized as identity matrices and their learning can be written in the following way:

$$\begin{cases} Q_{jl}^{-1}(k+1) = Q_{jl}^{-1}(k) + \lambda_Q \frac{\tau_{jl}^Q(k) e(k)}{\eta_Q(k)} \\ \eta_Q(k+1) = \beta_Q \eta_Q(k) + Tr(\tau_{jl}^{Q^T} \tau_{jl}^Q) \end{cases} \quad (12)$$

where λ_Q is a learning step and β_Q is a momentum hyperparameters, τ_{jl}^Q is a vector of values back propagated for each multidimensional Gaussian.

The centres c_{jl}^φ are placed equidistantly on initialization step and then tuned by the formula below.

$$\begin{cases} c_{jl}^\varphi(k+1) = c_{jl}^\varphi(k) + \lambda_c \frac{\tau_{jl}^c(k) e(k)}{\eta_c(k)} \\ \eta_c(k+1) = \beta_c \eta_c(k) + \tau_{jl}^{c^T} \tau_{jl}^c \end{cases} \quad (13)$$

where λ_c is a learning step and β_c is a momentum hyperparameters, τ_{jl}^c is a vector of back propagated values.

IV. EXPERIMENTAL RESULTS

Proper and unbiased comparison of the different forecasting models is a complex task [9]. The proposed

model and learning algorithm have shown good performance in real-life stock market datasets - daily log returns of IBM and Cisco. We have compared performance and prediction

accuracy to ANN based on bipolar sigmoid activation functions and resilient backpropagation learning algorithm. RMSE and SMAPE criteria were used.

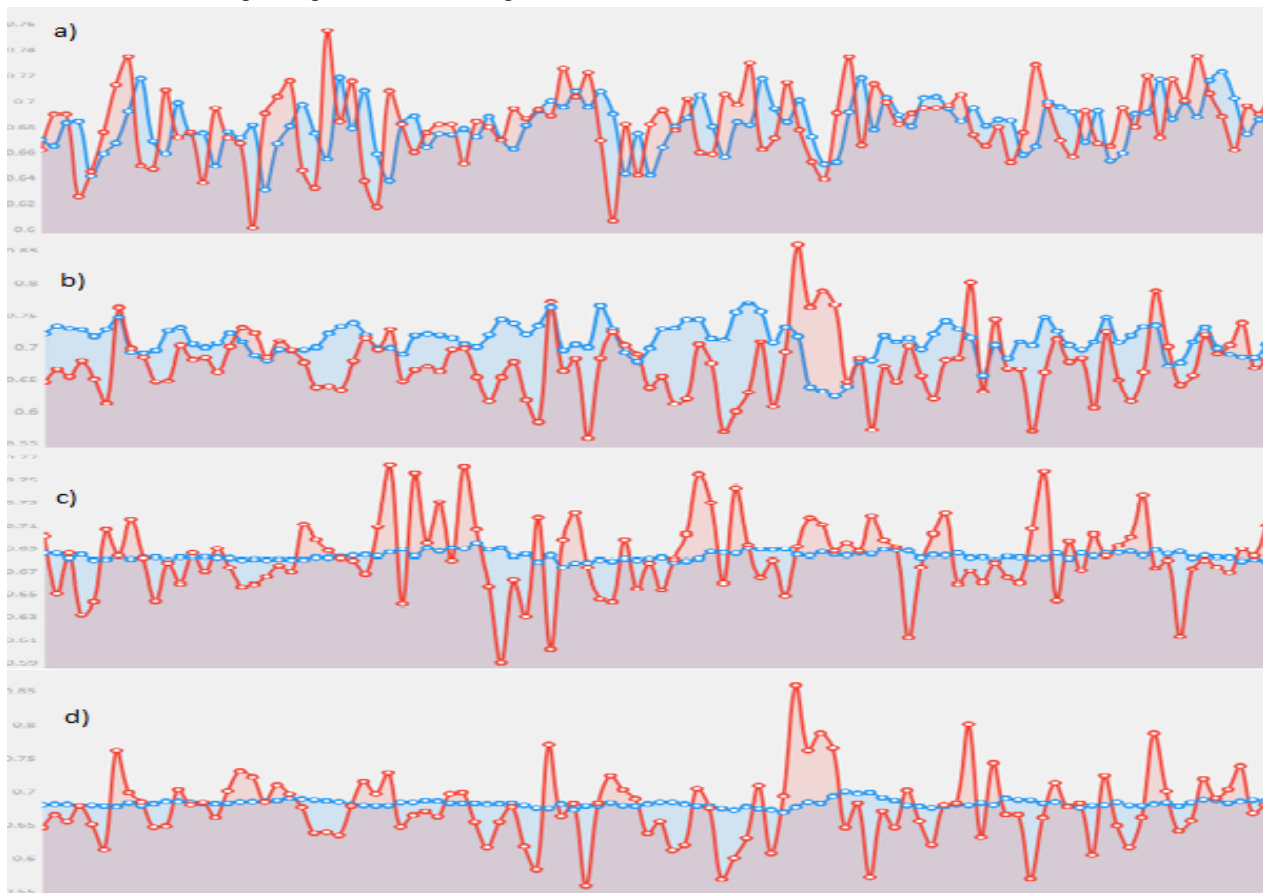


Fig. 5. Experiments on Cisco daily returns plots: a) proposed model – learning; b) proposed model – verification; c) bipolar sigmoid network – learning; d) bipolar sigmoid network – verification

The best results for the introduced model were achieved with $h^{\phi} = 2$, $\lambda_c = 0.87$, $\beta_c = 1.01$, $\lambda_Q = 0.81$, $\beta_Q = 1.02$. The bipolar sigmoid network was trained with alpha value 0.4, learning rate 0.67 and 100 epochs.

Model	Cisco daily log returns dataset results		
	Execution time, ms	RMSE, %	SMAPE, %
Proposed model	442	3.237	3.53
Bipolar Sigmoid Network	2668	3.31	3.6

V. CONCLUSION

In the paper a novel MIMO neuro-fuzzy model with multidimensional Gaussian functions in consequent layer is introduced. This solution allows to handle complex non-linear data, which is demonstrated on the stock market time-series prediction tasks.

REFERENCES

[1] J. J. F. Commandeur, and S. J. Koopman, An Introduction to State Space Time Series Analysis. Oxford University Press. 2007.

[2] Ye. Bodyanskiy, S. Popov, and T. Rybalchenko, "Multilayer neuro-fuzzy network for short term electric load forecasting," Computer Science–Theory and Applications, pp. 339-348, 2008.

[3] P. Otto, Ye. Bodyanskiy, and V. Kolodyazhnyi, "A new learning algorithm for a forecasting neuro-fuzzy network." Integrated Computer-Aided Engineering, vol. 10(4), pp. 399-409, 2003.

[4] Ye. Bodyanskiy, and S. Popov, "Neural network approach to forecasting of quasiperiodic financial time series,." European Journal of Operational Research, vol. 175(3), pp. 1357-1366, 2006.

[5] E. Hadavandi, H. Shavandi, and A. Ghanbari, "Integration of genetic fuzzy systems and artificial neural networks for stock price forecasting," Knowledge-Based Systems, vol. 23, no. 8, pp. 800-808, 2010.

[6] A. Esfahanipour, and W. Aghamiri, "Adapted neuro-fuzzy inference system on indirect approach TSK fuzzy rule base for stock market analysis," Expert Systems with Applications, vol. 37, no. 7, pp. 4742-4748, 2010.

[7] G. S. Atsalakis, and K. P. Valavanis, "Forecasting stock market short-term trends using a neuro-fuzzy based methodology," Expert Systems with Applications, vol. 36, no. 7, pp. 10696-10707, 2009.

[8] J. S. R. Jang, "ANFIS: adaptive-network-based fuzzy inference system," IEEE transactions on systems, man, and cybernetics, vol. 23, no. 3, pp. 665-685, 1993.

[9] S. F. Crone, M. Hibon, and Konstantinos Nikolopoulos. "Advances in forecasting with neural networks? Empirical evidence from the NN3 competition on time series prediction," International Journal of Forecasting, vol. 27, pp. 635-660, 2011.

Finite Generalization of the Offline Spectral Learning

Vladyslav Kotsovsky
IMST Department
Uzhhorod National University
Uzhhorod, Ukraine
kotsavlad@gmail.com

Fedir Geche
Department of Cybernetics and Applied
Mathematics
Uzhhorod National University
Uzhhorod, Ukraine
fgeche@hotmail.com

Anatoliy Batyuk
ACS Department
Lviv Polytechnic National University
Lviv, Ukraine
abatyuk@gmail.com

Abstract—We study the problem of offline learning discrete functions on polynomial threshold units over specified set of polynomial. Our approach is based on the generalization of the classical "Relaxation" method of solving linear inequalities. We give theoretical reason justifying heuristic modification improving the performance of spectral learning algorithm. We demonstrate that if the normalizing factor satisfies sufficient conditions, then the learning procedure is finite and stops after some steps, producing the weight vector of the polynomial threshold unit realizing the given threshold function. Our approach can be applied in hybrid systems of computational intelligence.

Keywords—offline learning, polynomial threshold unit, threshold function, artificial neural network.

I. INTRODUCTION

Artificial neural networks on the base of neural-like computational units have many applications and are intensively used for solving numerous important practical tasks [1]. It should be mentioned that many different models of neuron have been proposed. Polynomial threshold units (PTU) are ones of the most powerful between neural-like units with threshold activation function. They are based on separation of the n -dimensional space by the polynomial hypersurface.

Our offline learning algorithm for PTU uses the basic idea of "Relaxation" method introduced by Motzkin and Schoenberg [2]. Many different modifications of this method concerning online learning algorithms for perceptron-like devices proposed in [1, 3]. The offline modification of the algorithm with similar learning rules described in [4] for linear threshold units. Its generalization for PTU may be found in [5]. The main lack of these algorithms is the possibility of infinite learning time and convergence to the boundary point of the acceptable solution set. Hampson and Kibler proposed the modification of the choice of the amount of correction for online learning [6]. They announced that slightly larger normalizing factor in Reflection algorithm improves performance, but their reasons are rather "heuristic" and based only on empirical data.

The paper has the following organization: first the structure of PTU over given set of polynomials X and the notion of X -threshold function are given. Then basic concepts of our learning framework are described by using of the spectral technics similar to proposed in [7]. The rule of the choice of learning coefficients is discussed. In the next chapter the finiteness of the learning is proved. Finally,

we analyze the results of computer simulation and make conclusions.

II. POLYNOMIAL THRESHOLD UNITS

A computation unit with n inputs x_1, \dots, x_n and one output $y \in \{-1, 1\}$ is said to be a *polynomial threshold unit* if

$$y = \text{sgn} \left(\sum_{k=1}^m w_k \prod_{i=1}^n x_i^{j_{ki}} \right), \quad j_{ki} \in \{0, 1, \dots\}, \quad i = 1, \dots, n, \quad (1)$$

where $\mathbf{w} = (w_1, \dots, w_m)$ is the weight vector and the activation function is the sign function, given by

$$\text{sgn}(x) = \begin{cases} 1 & \text{if } x > 0, \\ 0 & \text{if } x = 0, \\ -1 & \text{if } x < 0. \end{cases}$$

It should be noted that if the weighted sum $\sum_{k=1}^m w_k \prod_{i=1}^n x_i^{j_{ki}}$ is equal to 0, then the output value of the PTU is incorrect (i.e. it does not belong to bipolar set $\{-1, 1\}$). But it is easy to prove that using "small" changes of PTU weights we can always avoid this inconvenience.

PTU and neural nets on their base are useful in pattern recognition methods providing sets separation in n -dimensional Euclidean space. One of the most important tasks is the one of separating the Boolean hypercube $\{-1, 1\}^n$ vertices to two different sets. For such task we can limit oneself in (1) with the terms of the form $w_{i_1 \dots i_k} x_{i_1} \dots x_{i_k}$. This is because $x^{2l} = 1$, $x^{2l+1} = x$ ($l \in \mathbb{N}$) in such case. From here we will restrict our attention only to PTU with inputs belonging to the bipolar set $\{-1, 1\}$.

We think of a Boolean function as a mapping $f: E_2^n \rightarrow E_2$, where $E_2 = \{-1, 1\}$. Let us define functions $\chi_j: E_2^n \rightarrow E_2$ by $\chi_j(\mathbf{a}) = a_1^{j_1} \dots a_n^{j_n}$, where $\mathbf{a} = (a_1, \dots, a_n) \in E_2^n$, $j = j_1 2^{n-1} + j_2 2^{n-2} + \dots + j_n$ ($j_i \in \{0, 1\}$, $i = 1, \dots, n$). Thus

$\chi_j(\mathbf{a}) = a_{i_1} \dots a_{i_r}$, if $j = 2^{n-l_1} + 2^{n-l_2} + \dots + 2^{n-l_r}$, $1 \leq l_1 < \dots < l_r \leq n$. The mappings χ_j , $j = 0, 1, \dots, 2^n - 1$ are well-known as characters of the group E_2^n over the real number field \mathbb{R} .

Let $X = \{\chi_{i_1}, \dots, \chi_{i_m}\}$ be an arbitrary set of characters and $f: E_2^n \rightarrow E_2$ is the given Boolean function. If there exists a such weight vector $\mathbf{w} = (w_1, \dots, w_m) \in \mathbb{R}^m$ that

$$\text{for all } \mathbf{a} = (a_1, \dots, a_n) \in E_2^n \quad f(\mathbf{a}) = \text{sgn} \left(\sum_{j=1}^m w_j \chi_{i_j}(\mathbf{a}) \right),$$

then we say that f is realizable on the PTU, or f is X-threshold function.

Furthermore, the weight vector \mathbf{w} must satisfy the following condition: for all $\mathbf{a} \in E_2^n$ $(\mathbf{w}, \chi(\mathbf{a})) \neq 0$, where $(\mathbf{w}, \chi(\mathbf{a})) = w_1 \chi_{i_1}(\mathbf{a}) + \dots + w_m \chi_{i_m}(\mathbf{a})$ is the inner product of the vector \mathbf{w} with $\chi(\mathbf{a}) = (\chi_{i_1}(\mathbf{a}), \dots, \chi_{i_m}(\mathbf{a}))$. We call such weight vectors X-acceptable. Note that the notion of X-threshold function is the generalization of the notion of threshold function [4]. We shall denote by $W_X(f)$ the set of all weight vectors of all PTU realizing the given X-threshold function f .

III. LEARNING FRAMEWORK

We assume that in our model of supervised learning the set of polynomials (characters) X is fixed and an arbitrary X-threshold function $f: E_2^n \rightarrow E_2$ is given. We will be interested in algorithm of finding some finite sequence of weight X-acceptable vectors

$$\mathbf{w}^0, \mathbf{w}^1, \dots, \mathbf{w}^r \quad (2)$$

that the function f can be realized on PTU with weight vector \mathbf{w}^r .

For function $f(x_1, \dots, x_n)$ we will define the spectral vector $\mathbf{s}_X(f) = (s_1, \dots, s_m)$ respectively to the set X in the following way:

$$s_j = \sum_{\mathbf{a} \in E_2^n} \chi_{i_j}(\mathbf{a}) f(\mathbf{a}), \quad j = \overline{1, m}.$$

Let $\mathbf{w} \in W_X(f) \subset \mathbb{R}^m$ be an arbitrary weight vector of the PTU realizing the function f over the set X . It is important for us that the set $W_X(f)$ is a cone. Suppose that the first $k+1$ members $\mathbf{w}^0, \mathbf{w}^1, \dots, \mathbf{w}^k$ of the sequence (2) are already chosen and let \mathbf{w}^k be an m -dimensional X-acceptable real vector. If the function f can be realized on PTU with the weight vector \mathbf{w}^k , then the learning succeed.

Suppose that previous assumption is wrong. This implies that f is not realizable on PTU with the weight vector \mathbf{w}^k . Let us describe how we can obtain the vector \mathbf{w}^{k+1} closer to the all $\mathbf{w} \in W_X(f)$ than the previous vector \mathbf{w}^k , i.e.

$$\|\mathbf{w}^{k+1} - \mathbf{w}\| < \|\mathbf{w}^k - \mathbf{w}\|, \quad (3)$$

where $\|\mathbf{x}\| = \sqrt{(\mathbf{x}, \mathbf{x})}$ is the Euclidean norm in the space \mathbb{R}^m . The condition (3) is well-known Fejér condition [2].

Let us use the learning rule:

$$\mathbf{w}^{k+1} = \mathbf{w}^k + \mathbf{z}^k, \quad (4)$$

where \mathbf{z}^k is the correction vector. Now we consider the question of the choice of the increment \mathbf{z}^k . Let $f^k(x_1, \dots, x_n)$ be a Boolean function realizable over the set X on PTU with weight vector \mathbf{w}^k . From previous assumption it follows that $f^k \neq f$ and $\mathbf{s}_X(f^k) \neq \mathbf{s}_X(f)$. Similarly to [4] we select the increment vector \mathbf{z}^k in the following way:

$$\mathbf{z}^k = \alpha_k (\mathbf{s}_X(f) - \mathbf{s}_X(f^k)),$$

where α_k is a some positive coefficient.

We proved in [8] that inequality (3) is held for fixed $\mathbf{w} \in W_X(f)$, if the coefficient α_k is chosen to be $t_k \alpha_k^0$, where

$$\alpha_k^0 = \frac{(\mathbf{w}^k, \mathbf{s}_X(f^k) - \mathbf{s}_X(f))}{\|\mathbf{s}_X(f) - \mathbf{s}_X(f^k)\|^2},$$

and the value of the amount of correction t_k satisfies the following inequality:

$$0 < t_k < 2 \left(1 + \frac{(\mathbf{w}^k, \mathbf{s}_X(f) - \mathbf{s}_X(f^k))}{(\mathbf{w}^k, \mathbf{s}_X(f^k) - \mathbf{s}_X(f))} \right). \quad (5)$$

According to the terminology of [4] we call the coefficient t_k a normalizing increment coefficient.

From now on, we restrict ourselves consideration of increment vectors \mathbf{z}^k of the form

$$\mathbf{z}^k = t_k \alpha_k^0 (\mathbf{s}_X(f) - \mathbf{s}_X(f^k)), \quad (6)$$

where the normalizing factor t_k satisfies (5).

Remark 1. Note that for increments (6) the inequality (3) holds for each weight vector $\mathbf{w} \in W_X(f)$ under the

condition that the normalizing coefficient t_k satisfies inequality $0 < t_k \leq 2$. Sometimes, it is more convenient to require the Fejér condition for all $\mathbf{w} \in A \subset W_X(f)$. Under this condition it can be possible to obtain the upper bound in (5) greater than 2 (for all $\mathbf{w} \in A$).

Remark 2. When we chose the increment vector \mathbf{z}^k in the form (6) it is necessary to require that for all $\mathbf{a} \in E_2^n$ $(\mathbf{w}^{k+1}, \mathcal{X}(\mathbf{a})) \neq 0$. In another case the weight vector \mathbf{w}^{k+1} is not X-acceptable and in according to our definition of PTU it is impossible to use \mathbf{w}^{k+1} for representing any Boolean function. We always can reach it by small changes of the increment vector (changing the normalizing increment coefficient t_k in such way that condition (3) holds). Let \mathbf{w}^k be an X-acceptable vector, but $\mathbf{w}^{k+1} = \mathbf{w}^k + \mathbf{z}^k$ is already an unacceptable one. Then the set $A_k = \{\mathbf{a} \mid \mathbf{a} \in E_2^n, (\mathbf{w}^{k+1}, \mathcal{X}(\mathbf{a})) = 0\}$ is nonempty. We can use the increment $\tilde{\mathbf{z}}^k = (1 - \beta_k) \mathbf{z}^k$, where β_k is an arbitrary factor satisfying following condition:

$$0 < \beta_k \leq \frac{1}{2} \min \left\{ 1, \frac{\min \left\{ (\mathbf{w}^{k+1}, \mathcal{X}(\mathbf{a})) \mid \mathbf{a} \in E_2^n \setminus A_k \right\}}{\max \left\{ (\mathbf{w}^{k+1}, \mathcal{X}(\mathbf{a})) \mid \mathbf{a} \in A_k \right\}} \right\}.$$

It is easy to verify that $\tilde{\mathbf{w}}^{k+1} = \mathbf{w}^k + \tilde{\mathbf{z}}^k$ is the X-acceptable weight vector.

IV. FINITE MODIFICATION OF LEARNING ALGORITHM

We have already mentioned that the offline modification of the rule (4) is due to Dertouzos [4] (see also [5]). In both works the coefficients t_k is from $(1, 2]$, the algorithm is not guaranteed to terminate after a finite number of steps, and some rather complicated techniques are used. The first condition guaranting the finiteness of the spectral offline learning was established in [8], where the rule was proposed of the normalizing coefficients choice, which utilization in spectral learning algorithm for X-threshold Boolean functions f ensures the finiteness of the learning procedure (i.e. after finite number of steps we certainly obtain the weight vector $\mathbf{w}^r \in W_X(f)$). If the coefficients t_k are chosen according to

$$t_k = 2 + \frac{1}{(\mathbf{w}^k, \mathbf{s}_X(f^k) - \mathbf{s}_X(f))}, \quad (7)$$

then the learning process (4), (6) is finite.

Now we can state our main results generalizing above mentioned one. For our purpose we will choose the coefficients t_k in the following way:

$$t_k = \sigma_k + \frac{\tau_k}{(\mathbf{w}^k, \mathbf{s}_X(f^k) - \mathbf{s}_X(f))}, \quad (8)$$

where $\sigma_k > 0$, $\tau_k > 0$ ($k = 0, 1, \dots$). It easy to see that the rule (8) is the generalization of (7) with two additional sequences of parameters.

Proposition 1. If a Boolean function f is X-threshold, the sequence (2) of X-acceptable weight vectors $\{\mathbf{w}^k\}$ is built on the base (4) and (6), the normalizing coefficients t_k satisfy (8) (accordingly to Remark 2), coefficients σ_k satisfy

$$0 \leq \sigma_k \leq 2 \quad (k = 0, 1, \dots), \quad (9)$$

and the sequence $\{\tau_k\}$ is bounded and does not converge to zero, then the learning process terminates after finite number of steps on some weight vector $\mathbf{w}^r \in W_X(f)$.

Proof. We need two simple lemmas established in [9].

Lemma 1. If for all $\mathbf{x} \in X$ $\|\mathbf{v} - \mathbf{x}\| < \|\mathbf{w} - \mathbf{x}\|$ then for all $\mathbf{y} \in \text{conv } X$ the inequality $\|\mathbf{v} - \mathbf{y}\| < \|\mathbf{w} - \mathbf{y}\|$ holds, where $\text{conv } X$ is the convex hull of set X in some Euclidean space.

Lemma 2. Let $\{\mathbf{x}^n\}$ be a sequence of the vectors of \mathbb{R}^n , $Z \subset \mathbb{R}^n$ and the set Z contains at least one inner point. Then if each $\mathbf{z} \in Z$ satisfies inequality $\|\mathbf{x}^{n+1} - \mathbf{z}\| < \|\mathbf{x}^n - \mathbf{z}\|$, then the sequence $\{\mathbf{x}^n\}$ is convergent.

We prove the statement of the theorem by contradiction. Suppose that the sequence $\{\mathbf{w}^k\}$ is infinite. It is easy to verify that for an arbitrary set of characters X and for an arbitrary Boolean function g the coordinates of the spectral vector $\mathbf{s}_X(g)$ are even integer numbers. Hence, for an arbitrary fixed integer vector $\mathbf{w}' \in W_X(f)$ (note that such vectors are always presented in the convex cone $W_X(f)$) the numerator $(\mathbf{w}', \mathbf{s}_X(f) - \mathbf{s}_X(f^k))$ of the fraction in the right part of (5) is an even natural number. Let A be a subset of integer weighted vectors belonging to set $W_X(f)$ ($A = W_X(f) \cap Z^m$) and $\tau = \sup\{\tau_k\}$. For all weight vectors from the set $\tau A = \{\tau \mathbf{w} \mid \mathbf{w} \in A\}$ the choice of the increment coefficients according to (8), (9) ensures (5). Thus, the Fejér condition for the set $\tau W_X(f)$ (inequality (3)) also holds for each member of the sequence $\{\mathbf{w}^k\}$. Let us also consider integer vectors $\mathbf{u}^j = 2\mathbf{w}' - \mathbf{e}^j$, $\mathbf{v}^j = 2\mathbf{w}' + \mathbf{e}^j$, ($j = 1, \dots, m$), where \mathbf{e}^j are the unit basis vector of the space \mathbb{R}^m . It easy to see that $\mathbf{u}^j \in W_X(f)$ and $\mathbf{v}^j \in W_X(f)$, $j = 1, \dots, m$. Let us see $B = \text{conv}\{\mathbf{u}^1, \dots, \mathbf{u}^m, \mathbf{v}^1, \dots, \mathbf{v}^m\}$. Now, using Lemma 1 to the polyhedron τB , we obtain that for each $\mathbf{z} \in \tau B$ and for all members of the sequence $\{\mathbf{w}^k\}$ the following inequality holds: $\|\mathbf{w}^{k+1} - \mathbf{z}\| < \|\mathbf{w}^k - \mathbf{z}\|$. The set of the interior points of the polyhedron τB is nonempty (e.g. it contains the vector $2\tau \mathbf{w}'$). Hence, following Lemma 2, the sequence $\{\mathbf{w}^k\}$ is

convergent. But from (6) it follows that for increment vectors

$$\alpha_k = \left(\sigma_k + \frac{\tau_k}{\left(\mathbf{w}^k, \mathbf{s}_X(f^k) - \mathbf{s}_X(f) \right)} \right) \frac{\left(\mathbf{w}^k, \mathbf{s}_X(f^k) - \mathbf{s}_X(f) \right)}{\left\| \mathbf{s}_X(f^k) - \mathbf{s}_X(f) \right\|^2} = \frac{\sigma_k \left(\mathbf{w}^k, \mathbf{s}_X(f^k) - \mathbf{s}_X(f) \right) + \tau_k}{\left\| \mathbf{s}_X(f^k) - \mathbf{s}_X(f) \right\|^2}.$$

The denominator of the last fraction is bounded and $\left(\mathbf{w}^k, \mathbf{s}_X(f^k) - \mathbf{s}_X(f) \right) > 0$ [8]. Then, there exists such $\tau_{\min} > 0$ that for each k_0 there exists such $k > k_0$ that numerator of the last fraction is not less than τ_{\min} . So, the condition $\mathbf{z}^k \rightarrow 0$ is violated, which is the necessary condition for the convergence of the sequence $\{\mathbf{w}^k\}$. Therefore, on some step of the learning algorithm we obtain X-acceptable weight vector $\mathbf{w}^r \in W_X(f)$.

V. EMPIRICAL RESULTS

To study the dependence of efficiency of PTU learning on the value of coefficients t_k we have implemented a simulation test based on the learning of threshold Boolean functions corresponding to 10000 randomly generated weight vectors for $n=8$ and $n=10$ features. We used $X = \{1, x_1, \dots, x_n\}$ and constant $t_k = t$ instead of (8) for the sake of simplicity. For the comparison, we learned each function using two online algorithms Reflect and Reflect1 with ShuffleCycle order of the inputs [6].

First, we learned all generated Boolean functions for each $t \in (0, 10]$ with the step 0.05 and counted the number of successful learnings and the average amount of corrections. The

initial

$$\text{appro } \bar{\varepsilon}^{(l)} = \frac{1}{(\hat{\alpha}^{(l)})^{1/2}} \sum_{s=1}^p \mu_s^{(l)}, \quad l = 1, \dots, M \text{ ximations}$$

were chosen randomly. The general tendencies are following:

1. In the case $t \in (0, 1]$ the learning failed due to violence of the practical X-acceptability of weight vectors (the absolute values of weighted sums became less than 10^{-15}).
2. In the case $t > 1$ all learning terminated successfully.
3. In the case $t > 3$ the quick growth was observed of the number of iterations.

Fig. 1 provides an illustration of growth of the correction curve in the case $t > 3.3$. In the case $t \geq 3.9$ the average number of corrections exceeds 255 (not shown in this figure).

Then we restricted ourselves on the study of the case $t \in (1, 3)$. The number of functions left the same and 200 points were chosen uniformly (with the step 0.01). The

corresponding curves of the number of iterations are shown in Fig. 2. As can be seen, the iteration number is complicated nonunimodal function which minimum point is in the neighborhood of the point $t = 2$: $t_{\min} = 2.02$, in the case $n = 8$, $t_{\min} = 2.03$ in the case $n = 10$.

Then we thoroughly studied the case $t \in [1.9, 2.1]$ with the step 0.001. The results are shown in Fig. 3. We found that $t_{\min} = 1.991$, $c_{\min} = 3.109$ in the case $n = 8$, and $t_{\min} = 1.997$, $c_{\min} = 4.462$ in the case $n = 10$, where c_{\min} is the corresponding number of iteration in spectral algorithm.

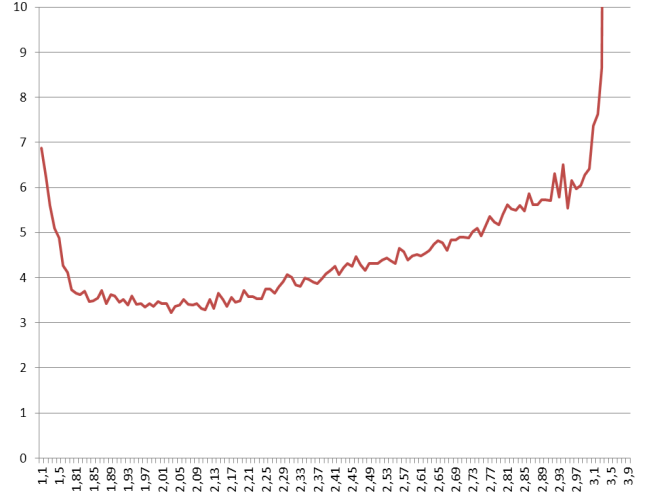


Fig. 1. Average number of iterations in (1.1, 3.9) in the case $n = 8$.

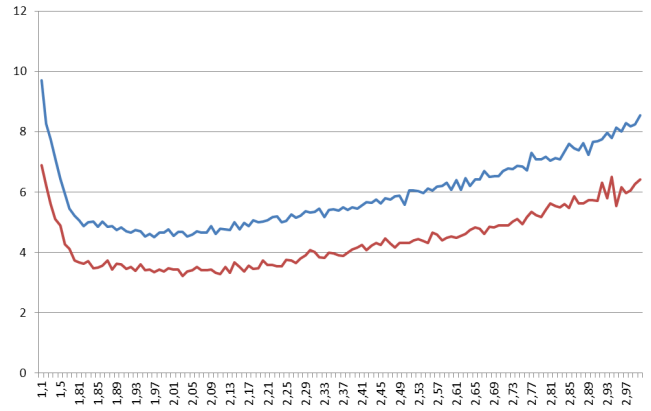


Fig. 2. Average numbers of iterations in (1, 3) in the cases $n = 8$ (lower curve) and $n = 10$ (upper curve).

For comparison, the average numbers of adjustments for Reflect1 are, respectively, 42.137 and 63.506 (the performance of Reflect is similar).

Results in Fig. 1-3 are obtained in the case of random initial approximations. The performance of the our learning algorithm can be improved by using the optimized approximation $\mathbf{w}^0 = \mathbf{s}_X(f)$. The reasons are given in [4, 5] and results are shown in Fig. 4. Comparing Fig. 3 and Fig. 4 we can observe that the correction number has halved roughly in the case $n = 8$ and has decreased by one third in the case $n = 10$.

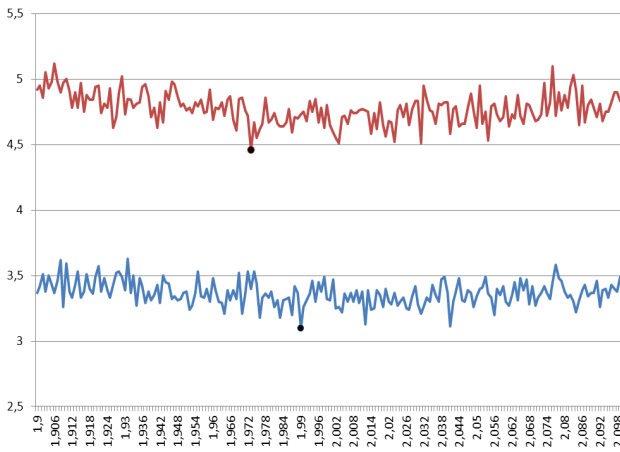


Fig. 3. Average numbers of iterations in $[1.9, 2.1]$ in the cases $n = 8$ (lower curve) and $n = 10$ (upper curve).

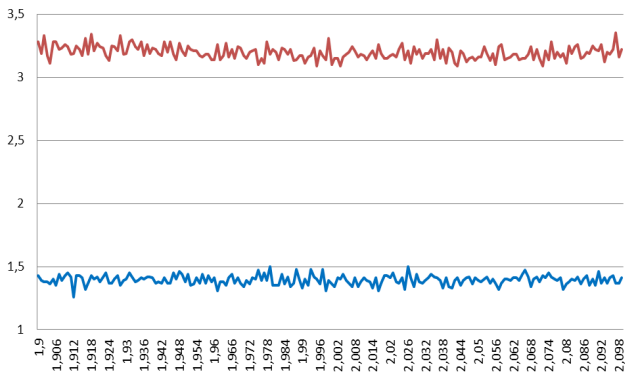


Fig. 4. Average numbers of iteration in $[1.9, 2.1]$ in the cases $n = 8$ (lower curve) and $n = 10$ (upper curve) with optimized initial approximation.

Finally, we studied the case $n = 20$ on 1000 random samples. For the spectral algorithm with optimized initial approximation $t_{\min} = 2.001$, $c_{\min} = 6.809$, and the average numbers of adjustments for Reflect1 is 252,413. The learning times are similar (Reflect is slightly faster). Note that it is possible to improve the performance of the learning algorithm by using t_k in the form (8). E.g., the simulation in the case $n = 10$ showed that we can increase performance by 12% using t_k for which $\sigma_k = 1.995$, $\tau_k = 1$.

VI. CONCLUSIONS

We proposed the new modification of offline learning method based on spectral approach. Our rule of the choice of normalizing coefficient ensures finite learning time for all X-threshold functions. In addition, in case of offline learning we confirmed the hypothesis of Hampson that it is reasonable to choose correction amount slightly larger than 2 [6]. The experimental results confirm the effectiveness of

developed method. They suggest to apply offline learning with optimized initial approximation $\mathbf{w}^0 = \mathbf{s}_x(f)$ and coefficients $\sigma_k \in [1.99, 2.01]$ ($k = 0, 1, \dots$) to obtain a good performance.

It should be mentioned that our approach can be utilized in learning PTU to recognize the subsets of an arbitrary finite set in n -dimensional Euclidean space. Our learning technics can be applied to improve performance of basic components proposed in [9]. Nonlinear classifier on the base of PTU may be also used in systems described in [10].

It seems to be also interesting to find the bounds of the number of algorithm iterations, to estimate the size of corresponding integer weights of PTU and to compare them with the ones from [11].

VII. ACKNOWLEDGMENT

We would like to thank M. Zhurbenko for many helpful discussions concerning Proposition 1.

REFERENCES

- [1] S. Haykin, *Neural Networks: A Comprehensive Foundation*, 2nd ed. Upper Saddle River, NJ: Prentice Hall, 1999.
- [2] T. Motzkin and I. Schoenberg, "The relaxation method for linear equalities," *Canadian Journal of Math.*, vol. 6, pp. 393–404, 1954.
- [3] R. Duda, P. Hart and D. Stork, *Pattern Classification*, 2nd ed. New York: Wiley-Interscience, 2001.
- [4] M. L. Dertouzos, *Threshold Logic: A Synthesis Approach*. Cambridge, MA: The MIT Press, 1965.
- [5] F. Geche. *Analysis of Discrete Functions and Logical Circuits in Neural Basis*. Uzhgorod: Vydavnytstvo V. Padyaka, 2010. (in Ukrainian)
- [6] S. Hampson and D. Kibler, "Minimum generalization via reflection: a fast linear threshold learner," *Machine Learning* vol. 37(1), pp. 51-73, 1999.
- [7] J. Bruck, "Harmonic analysis of polynomial threshold functions," *Siam Journal on Discrete Mathematics*, vol. 3 (2), pp. 168–177, 1990.
- [8] F. E. Geche, V. M. Kotsovsky and A. Ye. Batyuk, "Learning algorithms for generalized neurons over character set," *Zbirnyk naukovykh prats instytutu problem modelyuvannya v energetytsi NAN Ukrainy*, vyp. 41, pp. 124-136, 2007. (in Ukrainian)
- [9] I. Tsmots, V. Teslyuk, T. Teslyuk and I. Ihnatyev, "Basic components of neuronetworks with parallel vertical group data real-time processing," *Advances in Intelligent Systems and Computing*, vol. 689, Springer, Cham., pp. 558–576, 2018.
- [10] V. Teslyuk, V. Beregovskiy, P. Denysyuk, T. Teslyuk and A. Lozynskiy, "Development and implementation of the technical accident prevention subsystem for the smart home system," *International Journal of Intelligent Systems and Applications*, vol. 10, No.1, pp. 1–8, 2018.
- [11] F. Geche, V. Kotsovsky and A. Batyuk, "Synthesis of the integer neural elements," in *Proceedings of the International Conference on Computer Sciences and Information Technologies CSIT 2015*, Lviv, Ukraine, September 14-17 2015, pp. 121–136.

Inverse Problem for Two-Dimensional Heat Equation with an Unknown Source

Nelya Pabyrivska
Department of Mathematics
Lviv Politechnic National University
Lviv, Ukraine
nelyapab@gmail.com

Viktor Pabyrivskyy
Department of Applied Mathematics
Lviv Politechnic National University
Lviv, Ukraine
pabvic67@gmail.com

Abstract—The paper establishes existence and unique conditions for an inverse problem with an unknown source. The unknown source is a polynom for two spatial variables with unknown coefficients depending on time.

Keywords—inverse problem, Green function, Volterra integral equations, unknown source

I. INTRODUCTION

The inverse problems of identification of the coefficients for parabolic equation were considered in various formulations in numerous works of national and foreign mathematicians.

The problem of determination of the unknown free term is the simplest among coefficient inverse problems. Particularly in this direction the following results were obtained. In paper [1] a free term of the heat equation was being looked up as a product of two unknown functions $f(t)g(x)$.

In [2] the unknown source has the form $f_1(x)g_1(t) + f_2(t)g_2(x)$, where $f_1(x), f_2(t)$ have to be identified. For this problem, the local conditions of existence and uniqueness of the solution are set.

This problem was analyzed in paper [3], for the case of integrated overdetermination conditions, and contained numerous examples of its implementation.

Paper [4] also shows the numerical solution of simultaneous determination inverse problems of time-dependent coefficients of the younger members and the source of the parabolic equation.

In [5] the inverse problem for a weakly nonlinear ultraparabolic equation with three unknown functions depending on various arguments is considered. The conditions for the existence and the uniqueness of the generalized solution of this problem are established on some narrowed time interval. The components of the solution are from Sobolev spaces.

Paper [6] is dedicated to finding sources in a parabolic integral-differential equation in the form $\chi(t)\omega(x), \chi(x_n)\omega(\bar{x},t)$ and $\sum_{j=1}^n \omega_j \chi_j(x,t)$, where χ, χ_j are the specified functions.

In the paper [9] the problem for the two-dimensional parabolic equation with an unknown heterogeneous transient orthotropic thermal conductivity is studied. The author uses

initial and Dirichlet boundary conditions and fluxes as overdetermination conditions. For the numerical solving of such problems with specific initial data, the finite-difference method is used.

To ensure the unique solvability of the inverse problem solution the measurement data represented by the heat fluxes are considered.

The inverse problems of identification of an unknown source were also considered in the works [10]-[13].

II. FORMULATING THE PROBLEM

A large number of technological processes in metallurgy occur at elevated temperatures. In such cases, it is often impossible to control the temperature during experiments. This happens when the heated object is of small geometric dimensions, such as a thin wire, thin rectangular domain or when the heating takes place in an environment that is not accessible for installing temperature sensors.

In recent years, in practice, new high-speed methods of heat treatment of metals have been widely used, the temperature distribution here can be controlled using a mathematical model, using solutions for inverse problems. This raises the problem of the existence and uniqueness of the solution for such problems.

Let's have a two-dimensional rectangular domain $(0 < x < l, 0 < y < h)$. Because of $u(x, y, t)$ we denote its temperature at the coordinate point (x, y) at the time t . It is necessary to organize such a process in a way that certain points of the rectangular domain $((0,0), (0,h), (l,0), (l,h))$ have predetermined time-dependent temperature regimes.

This can be achieved due to the thermal effects of internal heat sources, that is, solving the problem of controlling the source in the heat equation. In fact, we need to investigate the inverse problem for the heat equation with an unknown free member, which depends on the two spatial variables x, y and time t .

The question of the complete determination of the source (the definition of a free member, which depends on spatial and time variables) remains open today.

A certain solution to complete definition of an unknown member of the parabolic equation is a polynomial approximation for its spatial variables with unknown coefficients that depend on time.

This task of identification of the source of the polynomial is developed in this study. The approach to a complete determination of the unknown coefficients parabolic equation was tested in [7], [8], where the conditions of existence and uniqueness of the solution of the problem of identification of the major coefficient of the linear and quadratic functions of a spatial variable with unknown coefficients depending on the time variable were found.

III. MATHEMATICAL MODEL

In the domain $\Omega = \{(x, y, t): 0 < x < l, 0 < y < h, 0 < t < T\}$ consider the parabolic equation

$$u_t = u_{xx} + u_{yy} + \sum_{i=0}^1 \sum_{j=0}^1 x^i y^j f_{ij}(t) \quad (1)$$

with unknown coefficients $f_{ij}(t), i, j = \overline{0,1}$, with the initial condition

$$u(x, y, 0) = \phi(x, y), \quad (x, y) \in [0, l] \times [0, h], \quad (2)$$

the boundary conditions

$$u_x(0, y, t) = \mu_1(y, t), \quad u_x(l, y, t) = \mu_2(y, t), \\ (y, t) \in [0, h] \times [0, T],$$

$$u_y(x, 0, t) = \mu_3(x, t), \quad u_y(x, h, t) = \mu_4(x, t),$$

$$(x, t) \in [0, l] \times [0, T] \quad (3)$$

and the over determination conditions

$$u(0, 0, t) = v_{00}(t), \quad u(0, h, t) = v_{01}(t), \\ u(l, 0, t) = v_{10}(t), \quad u(l, h, t) = v_{11}(t), \quad t \in [0, T]. \quad (4)$$

The solution of the problem (1)-(5) are the functions $(u(x, y, t), f_{ij}(t)) \in C^{2,1}(\Omega) \times (C[0, T])^4, i, j = \overline{0,1}$, that satisfy conditions (1)-(4).

IV. EXISTENCE AND UNIQUENESS OF THE SOLUTION FOR THE PROBLEM

Assuming that the output data possess the required smoothness and the conditions for their compliance, we reduce problem (1) - (5) to the system of equations for unknown functions $f_{ij}(t), i, j = \overline{0,1}$.

For this reason, let set $x = 0, l$ and $y = 0, h$, in the equation (1). In result we get the following correlations

$$v'_{00}(t) = u_{xx}(0, 0, t) + u_{yy}(0, 0, t) + f_{00}(t), \\ v'_{01}(t) = u_{xx}(0, h, t) + u_{yy}(0, h, t) + f_{00}(t) + hf_{01}(t), \\ v'_{10}(t) = u_{xx}(l, 0, t) + u_{yy}(l, 0, t) + f_{00}(t) + lf_{10}(t), \quad (5) \\ v'_{11}(t) = u_{xx}(l, h, t) + u_{yy}(l, h, t) + f_{00}(t) + \\ + hf_{01}(t) + lf_{10}(t) + lhf_{11}(t),$$

where $u(x, y, t)$ - the solution of the direct problem (1)-(3), which has the form

$$u(x, y, t) = \int_0^l d\xi \int_0^h \phi(\xi, \eta) G_2(x, y, t, \xi, \eta, 0) d\eta + \\ + \int_0^t d\tau \int_0^h \mu_2(\eta, \tau) G_2(x, y, t, l, \eta, \tau) - \\ - \mu_1(\eta, \tau) G_2(x, y, t, 0, \eta, \tau) d\eta + \\ + \int_0^t d\tau \int_0^l \mu_4(\xi, \tau) G_2(x, y, t, \xi, h, \tau) - \\ - \mu_3(\xi, \tau) G_2(x, y, t, \xi, 0, \tau) d\eta + \\ + \int_0^t \int_0^l \int_0^1 \sum_{i=0}^1 \sum_{j=0}^1 \xi^i \eta^j f_{ij}(\tau) G_2(x, y, t, \xi, \eta, \tau) d\eta d\xi d\tau,$$

$G_2(x, y, t, \xi, \eta, \tau)$ -- Green function of the second kind.

Since the determinant of the system (5)

$$\Delta = \begin{vmatrix} 1 & 0 & 0 & 0 \\ 1 & h & 0 & 0 \\ 1 & 0 & l & 0 \\ 1 & h & l & lh \end{vmatrix} = h^2 l^2$$

is not a zero, the system of equations (5) can be reduced to the form (6):

$$f_{km}(t) = F_{km}(t) + \sum_{i=0}^1 \sum_{j=0}^1 P_{ij}^{km}(t) u_{xx}(l_i, h_j, t), \quad (6) \\ k, m = \overline{0,1}$$

where $F_{km}(t), P_{ij}^{km}(t), i, j, k, m = \overline{0,1}$, - known functions, determined through input data, $l_0 = 0, l_1 = l, h_0 = 0, h_1 = h$.

The system (6) is a system of Volterra integral equations (VIEs) of the second kind.

Using the theory of VIEs of the second kind, it can be stated that the solution for the system of equations (6) exists and it is unique only if the following conditions are satisfied:

$$\phi(x, y) \in C^{2,2}[0, l] \times [0, h],$$

$$\mu_i(x, t) \in C^{2,1}[0, l] \times [0, T], i = 1, 2,$$

$$\mu_j(y, t) \in C^{2,1}[0, h] \times [0, T], i = 3, 4, v_{ij}(t) \in C^1[0, T],$$

$$i, j = \overline{0, 1}.$$

Based on the equivalence of the system of equations (6) and the problem (1) - (4), we conclude

THEOREM

The problem (1) - (4) has a unique solution if the conditions of smoothness (A) and harmonization

$$\phi'(0, y) = \mu_1(y, 0),$$

$$\phi'(l, y) = \mu_2(y, 0), \phi'(x, 0) = \mu_3(x, 0), \phi'(x, h) = \mu_4(x, 0),$$

$$\phi(0, 0) = v_{00}(0), \phi(l, 0) = v_{10}(0), \phi(0, h) = v_{01}(0),$$

$$\phi(l, h) = v_{11}(0) \text{ are satisfied.}$$

V. CONCLUSIONS

In this paper:

1) the mathematical model of the heating process in the thin rectangular plate with an available heating source with the unknown physical characteristics is constructed;

2) the obtained inverse problem for the heat equation is reduced to the system of the integral Volterra equations of the second kind;

3) basing on the theory of Volterra integral equations of the second kind there are established the conditions for the solution existence and uniqueness.

For the numerical solving of such problems, one can use the finite difference method, which is presented iteratively in the nonlinear minimization procedure.

REFERENCES

- [1] E. Savateev, "The problem of identification of a coefficient in a parabolic equation," *Siberian Mathematical Journal*. vol.36, no. 1. pp. 177-185, 1995.

- [2] M. Ivanchov, "Inverse problem for the multidimensional heat equation with unknown source," *Matematychni Studii*. no16, pp. 93-98, 2001.
- [3] Dinh Nho Hào, Phan Xuan Thanh, D. Lesnic, and M. Ivanchov, "Determination of a source in the heat equation from integral observations," *Journal of Computational and Applied Mathematics*, no. 264, pp. 82-98, 2014.
- [4] M. S. Hussein, and D. Lesnic, "Simultaneous determination of time-dependent coefficients and heat source," *International Journal for Computational Methods in Engineering Science and Mechanics*, vol. 17, pp. 401-411, August 2016.
- [5] N. Protsakh, "Inverse problem for weakly nonlinear ultraparabolic equation with three unknown functions of different arguments on the right side," *Ukrainian Mathematical Journal*. vol. 66, no. 3, pp. 371-390, 2014.
- [6] K. Kasemets, and J. Janno, "Reconstruction of a Source Term in a Parabolic Integro-Differential Equation from Final Data," *Mathematical Modelling and Analysis*, vol. 16, no. 2, pp.199-219, 2011.
- [7] M. Hussein, D. Lesnic, and M. Ivanchov, "Identification of a Heterogeneous Orthotropic Conductivity in a Rectangular Domain," *International Journal of Novel Ideas: Mathematics, S.l.*, vol. 1, pp. 1-11, apr. 2017.
- [8] N. Pabyrivska, and V. Vlasov, "The determination of major coefficient factor in parabolic equation," *Mathematical methods and physico-mechanical fields*, vol.49, no. 3, pp.18-25, 2006.
- [9] N. Pabyrivska, and O. Varenik, "The determination of major coefficient factor in parabolic equation," *Lviv University Paper*. Ed.64., pp.181-189.
- [10] M. Ivanchov, *Inverse problem for equations of parabolic type. Mathematical Studies. Monograph Series. Lviv. VNTL Publishers.* vol.10, 2003.
- [11] A. Hasanov, "Simultaneous determination of source terms in liner parabolic problem from the final overdetermination approach," *J. Math. Anal.Appl.*, vol. 330(2) pp. 766-779, 207. Doi:10.1016/j.jmaa.2006.08.018.
- [12] A. Lorenzi and G. Mola, "Identification of unknown terms in convolution integro-differential equations in a Banach space," *J.Inverse Ill-Posed Probl.*, vol. 18(3), pp. 321-355, 2010. Doi:10.1515/jllp.2010.016.
- [13] E. Pais, "Identification of memory kernels in heat flow measuring heat flux at the ends of the bar," *Math. Model. Anal.*, vol. 15(4), pp. 473-490, 2010. Doi:10.3846/1392-6292.2010.15.473-490.

AVIA: Automatic Vulnerability Impact Assessment on the Target System

Yuliia Tatarinova
Kharkiv National University of Radio Electronics
Kharkiv, Ukraine
yullia.tatarinova@gmail.com

Abstract — Manual labor in information technologies becomes too much time-consuming and not effective with increasing complexity of computer systems. As a result, automation of all process in cyber security gains a popularity in recent years. Impact evaluation of disclosed security flaw on overall system can be a challenge under certain conditions. In this paper we designed a new method, which makes automatic vulnerability impact assessment and reduces manual vulnerability impact analysis with the help of binary analysis and public non-code text related vulnerability information.

Keywords—security assessment, vulnerability, impact, binary analysis

I. INTRODUCTION

Software development is a complex and not a perfect process which generates inherently imperfect programs. Vulnerabilities and security flaws exist in all type of software [1] and even in hardware [2]. A lot of cyber attacks increasingly use these security issues. Another kind of problem is continuously increasing the number of vulnerabilities. The number of Common Vulnerabilities and Exposures (CVE) which registered in the National Vulnerability Database (NVD) [3] was 6,447 in 2016. Also in 2017 14,646 new security issues were found. Up to 3,376 undisclosed weakness were published during the first quarter of 2018. SANS estimated that over 80% of cybersecurity attacks and incidents exploit already known vulnerabilities [4]. Free and open source software which contains a huge number of publicly known weaknesses is used by the corporate or commercial software. Many problems and questions arise when discussing the above-mentioned issues by managers, security experts and software engineers. The most typical questions are the following: is a product vulnerable against found security issue? How to verify and detect known vulnerability in software products? What kind of vulnerability to fix first if we have a hundred or even a thousand vulnerabilities? What are the criteria for vulnerability ranking and what impact could be on the final product? How to compute security flaw impact on overall computing system or product? How much time and resources does it take? How quickly make a risk assessment and impact analysis and verification in case of multiple software product versions, firmware or system configuration?

There are many approaches and areas designed to make software more secure. One of them is the Security Development Lifecycle (SDL) [25], which was proposed by Microsoft to reduce number of security related bugs.

Unfortunately, this methodology loses relevance by increasing system complexity and lines of code. Organizations and software companies deploy Vulnerability Management Systems (VMS) inside their products and IT infrastructures. VMS consist of different kinds of features that may involve scanning for software products that have potential vulnerabilities, indexing of software patches, etc. Most of VMS aims to detect flows in IT infrastructures or web applications. The largest number of them are commercial solutions. Many approaches also exist in impact analysis and vulnerability management process. The Common Vulnerability Scoring System (CVSS) [5] and OWASP Risk Ranking Methodology [6] are the most popular. In practice, impact analysis of the newest disclosed vulnerability (CVE) in a product could take a lot of time and human resources due to increasing software and firmware complexity. The human expert needs to investigate security issue (CVE) and its root cause, then determine whether vulnerable software component exists in firmware. If it is not stand alone component (shared library), then find all software products, which depends or use it. Next step is to define a potential attack surface and its complexity, investigate confidentiality, integrity and availability impact according to company firmware. If there are multiple supported products with earlier firmware versions – all of them must be checked. To perform this task, the expert must have detailed knowledge of the product and its environment. Automated tools such as static and dynamic analysis come in help with this problem. But no one of them could fully automate this intense manual labor. All of these steps human expert in many cases needs to make manually. It is a tedious and sometimes error-prone task because of the complexity of a target product and its components. Even security expert could not see all possible attack surfaces and entry points to exploit a vulnerability. That's why it is important to reduce manual security analysis and automate all aforementioned impact analysis steps. In this paper, we present proof of concept system that uses CVE dataset and product binary analysis to compute vulnerability impact on firmware, especially. Also we introduced a new method for computing vulnerability impact on target system. Our system was designed for Unix-based system built for arm-based chips and device architectures.

In this paper, we demonstrate that complicated security flaw impact analysis on target system could be fully automated by the end-to-end system, using non-code text related vulnerability information, particularly CVE description, patch information, and inter-product binary

dependencies. The paper is structured as follows. Section II describes the related prior work. Section III presents our approach and system design. Section IV introduces the common evaluation of our method. Finally, section V gives a conclusion.

II. RELATED WORK

Related prior work could be classified into three categories: existing solutions for fully automated vulnerability detection and management systems on end-point software, automated public vulnerability analysis, and program analysis (in terms of security flaw evaluation).

A. Prior work in vulnerability detection and management systems

Once vulnerability gets reported, the process of developing exploits start for it. The most popular and easy accessible of the existing fully automated solutions for security issue verification are a large number of both commercial and open source tools for web applications scanning, which searching only for web security related issues or insecure server configuration [7]. But such kind of vulnerabilities and targets are beyond the scope of the present paper.

Another powerful tool which is developed to automate routine and complex tasks is Metasploit [8]. Metasploit it is a free open source penetration testing framework. Metasploit v4 has more than 700 exploits and more than 250 payloads. But it specializes on penetration testing and not always applicable for such types of devices like the internet of things (IoT), gaming consoles or handheld devices. In most cases for CVE verification, it doesn't contain a fully automated solution and security expert need to develop exploit or port existing and integrate it into the framework [9].

Good host-based tool which helps to make base auditing is Lynis, which has open source and enterprise versions [10]. The main disadvantage, that it doesn't verify any specific CVEs of widely used software products. It just shows base system misconfigurations. To get the desired functionality security expert has to develop custom plugins and test cases.

Automation of cybersecurity has advanced with DARPA's Cyber Grand Challenge [11]. It was created to develop fully automation systems to discover, prove and patch software flaws. The winning system "Mayhem" [12] could automatically find exploitable bugs in binary. Every bug reported with working shell-spawning exploit. The same problem is also resolved by Automatic Exploit Generation (AEG) tool [13]. But these works don't resolve a question of distinction between already existing security flaws and new finding. The main difference of our investigation that system looks for existing security issues and tries to compute their impact on overall device

B. Prior work related to automatic public vulnerability analysis

CVE is a method used to assign identifiers to publicly known vulnerabilities found in IT products and to provide

information (e.g., affected products) about the vulnerabilities [14]. MITRE [15] corporation and NVD provides CVEs in XML format which contain the identifiers with additional information (e.g., short description, CVSS score, vulnerable configuration, additional references to patch, exploit etc.) CPE is a method that specifies naming scheme for applications, hardware devices and operating systems [15]. Both CVE and CPE play important role in VMS processing. L. Sanguino [15] discusses issues of incomplete matching information between these two components. In [16] authors studies CVE reports using description texts to find prevalent vulnerability types and new semi-automatically trends. Analyzing CVE data in [17] authors take into account only product version and reporting time. W. You and P. Zong introduce SemFuzz [18], an end-to-end technique leveraging vulnerability-related text to quick automatic generation of proof of concept (PoC) exploits using natural language processing (NLP) and semantics-based fuzzing process guided by such information. This work focuses on vulnerability type and doesn't pay attention on product dependencies.

Our approach takes into consideration not only CVE and CPE identifiers, but makes analysis of description and references information. We use these data to get issue root cause and provide future impact analysis.

C. Prior work in automated binary analysis and vulnerability detection

Vulnerability Deep Pecker (VulDeePecker) [19] is a deep learning-based vulnerability detection system which uses static analysis technique. Its experimental results quite encouraging. But sometimes security experts don't have any source code or wide range of vulnerability types. So it makes needless not only for VulDeePecker but also other source code-based static analysis tools [20].

Approach proposed in [21], a binary analysis framework, which implements such analysis techniques as static vulnerability discovery using control flow graph (CFG), dynamic concrete execution and introduced open source binary analysis framework - Angr.

Compared with aforementioned approaches, our technique automatically analyzes software vulnerability impact on overall Linux-based system and depended products, provides a clear understanding of the threat of vulnerability in open source products to the end system.

III. DESIGN OVERVIEW OF AUTOMATIC VULNERABILITY IMPACT ASSESSMENT ON END-POINT PRODUCT SYSTEM

Our objective is to design impact weaknesses evaluation system that can automatically estimate disclosed security flaw impact (W) on the end-point product (P). We consider product P as a whole complex Unix-like operation system with huge amount of different software products, components and running services. The principles on which operates the majority of security tools and frameworks are aimed at becoming investigate and make penetration test outside the system, gradually getting inside. Unfortunately, sometimes security experts couldn't find all possible paths and surfaces for successful attack when they

really exist. Our method aims to investigating and analyzing end-point product for existing security flaws. It designed to make a new impact estimation approach from issue root cause of issue to the entry points of the target system P . To address the challenges in impact evaluation method uses ELF binary analysis and semantical non-code text vulnerability information retrieving.

The AVIA's work flow is illustrated in Fig 1. The process of computing vulnerability impact on system could be divided in three main stages: (1) feature extraction and

data processing, and (2) risk assessment and impact evaluation.

A. Feature extracting and data processing

To address the problem of automatic assessment of vulnerability impact on target system, data collecting and processing procedure could be divided in two parts: processing data on target system and features which could be extracted from vulnerabilities.

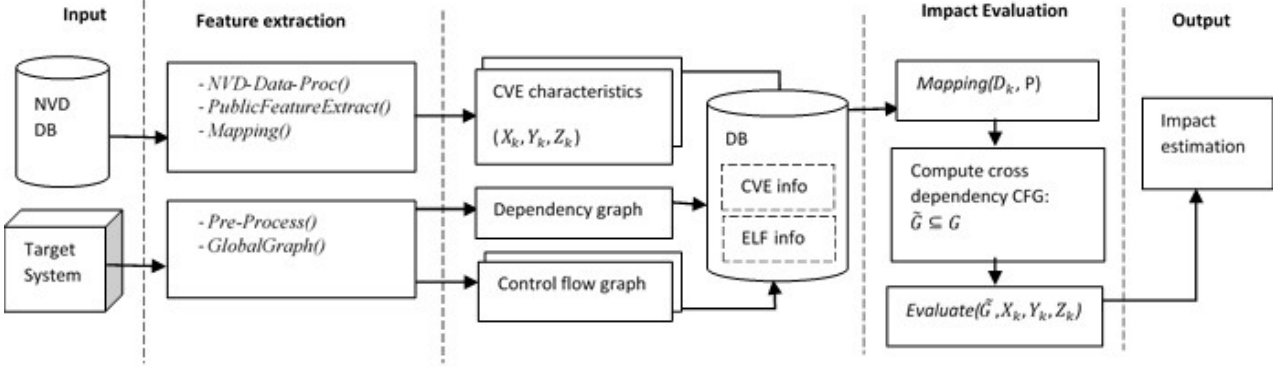


Fig.1. Brief AVIA design

Data processing on target system. As shown in Fig 1, at first we retrieve and analyze all ELF and its environment from file system. In our research we defined P as target product, which could be represented as $P = (C, I, G)$, where C it is a set of binary ELF files. This set contains set of executable ELF C_{exe} and shared object ELF C_{shared} . Each ELF c_i from this set could be proprietary (c_i^{prop}) or open source (c_i^{open}). Using other type of ELF than executable or shared object is out of scope in current paper. C we describe as:

$$C = C_{exe} \cup C_{shared} = \{c_1, c_2, \dots, c_n\}.$$

I it is intermediate representation set of each ELF which could be obtained during $Pre-Process(C)$. This algorithm returns the representation each binary c_i as a set of functions F , dependencies $Deps$ and cross-references graph CFG :

$$I = Pre-Process(C) = \bigcup_{i=1}^n (F, Deps, CFG).$$

During $Pre-Process$ stage we decompile ELF. The popular IDA Pro [24] is a recursive descent disassemble that we couldn't use in terms of our goal, because of lack of scalability on overall computer system and licensing restrictions. Objdump is also does not fit our goal because it requires superfluous amount of work. Capstone [22] disassembly framework was used to decompile and build cross references between functions calls. Each function f_j from set F we represent like $f_j = (Name, Code_{asm}, Type)$

where $Name$ is demangled function name, $Code_{asm}$ disassembled listing of code and $Type$ could has values: entry point, exported or internal. To retrieve dependency list $Deps$ of each executable we dissect it and examine dynamic section with tag DT_NEEDED [23], which holds string table offset to the name of needed library. Note, that symbolic links to ELF also must be included for processing in system according to dynamic linker [26] work. For this task ldd utility is great, but we couldn't use it for security reasons (some versions of ldd may directly execute program to retrieve information). In CFG each node represents a function f_j and direct calls represents edges.

Since ELF binary uses external shared objects which could be vulnerable and have CVEs, there is a need to extract and build cross dependencies graph (G). We compute graph dependencies between executable binaries and shared object files. The analysis process is to retrieve all cross dependencies between ELF on target system:

$$G = GlobalGraph(I) = \bigcup_{i=1}^n (Deps_i, CFG_i).$$

Features extracting from CVE information. Let dangerous of CVE could be expressed as $D_k(CVE) = (X_k, Y_k, Z_k)$.

Where X is a static vulnerability characteristics that are independent of the product that contains CVSS score S , vulnerability type T_{vuln} , attack type T_{at} , short text description $Descr$. So, $X_k = (S, T_{vuln}, T_{at}, Descr)$. All this features could be extracted during $NVD-Data-Proc(CVE)$ parsing.

We define Y_k as vulnerabilities features that could be obtained from public sources and news:

$$Y_k(CVE) = PublicFeatureExtract(CVE) \\ = (Rfs, Expl, Patch, R, Src, V, Rel_{(t)})$$

When CVE become public a lot of open references and news (Rfs) appears over the internet. Analyzing these resources we could obtain existence exploit $Expl$, patch $Patch$, root cause R , source code or git repository Src . CVE provides a references to vulnerable versions and affected products V . One of the most important criterion is the relevance of found data $Rel_{(t)}$. It decreases over time.

But very often this information is presented not an appropriate form. Furthermore, sometimes CVE description contains possible input format (ex. "via a small bit-depth value in an IHDR") which could be future used in automatic proof-of-concept (PoC) generation. Huge amount of CVE information comes in natural language. We used it to extract semantic information: affected product name, vulnerable functions, attack vector and malformed input format (ex. png header, tcp packet etc.). To achieve this aim, several approaches are used. At first, direct extraction of data such as regular expression based string match applied. If it doesn't work well, we propose to retrieve target information using NLP techniques including Part-of-Speech (PoS) Tagging, Phrase Parsing and Syntactic Parsing [18] from CVE descriptions and provided references.

We defined one more type of the vulnerability characteristics taking into account the target product P :

$$Z_k = Mapping(X_k, Y_k, P) = (L_{(c)} Al, At),$$

where $L_{(c)}$ is the list of affected binaries, Al applicability and availability of current vulnerability on target product, At is degree of attainability to this vulnerability from the entry point. Both Al and At are calculated taking into account product profile $Pr(p)$. Product profile describes main software characteristics such as product architecture and infrastructure, network details, user data, critical assets, data repositories etc.

B. Vulnerability impact evaluation process

When all required information was automatically collected, we can make impact analysis of weakness. We define CVE vulnerability impact as $W_i = F_i(D_i)$. We describe common steps to obtain W_i . First, is to check whether vulnerable component and affected binary list $L_{(c)}$ presents on system. If $L_{(c)}$ empty, then $Z_i = 0 \Rightarrow F(D_i) = 0$. In other case we try to compute subgraph $\tilde{G} \subseteq G$. When calculating it, the key role is played by the presence of root cause R (whether G has vulnerable function f). Sometimes several product versions or configurations could not contain vulnerable

function. In this case the process of computing cross dependencies graph is following:

1. Get CFG to the vulnerable function f (if it is exist) in binary c_i .
2. If affected binary file is shared object – get subgraph \tilde{G}' end nodes of which are $\forall f_i(Type) = exported$. Start processing from step 1 with all ELF's which depend on current affected binary c_i . Consider exported functions from current c_i in depended ELF as vulnerable.
3. If affected binary c_i is executable – get subgraph \tilde{G}' according to the binary entry point.
4. Merge \tilde{G}' with result graph \tilde{G} .

Steps 1 – 4 repeat recursively while end-point binaries with entry points are found or until dependency nodes ended. As a result, a graph \tilde{G} of paths from vulnerable function in vulnerable component c_i (ex. shared object) to entry point of the dependent program c_i (end-point binary) is found.

Next step is subgraph \tilde{G} evaluation. At this stage of study we determined the following parameters of graph \tilde{G} assessment:

- number of affected binaries (N_b);
- number of nodes (functions F) (N_f);
- edges number (N_e);
- evaluation of availability and attainability (N_a) using BFS and DFS graph algorithms;
- end-point binary evaluation according to product profile $Pr(p)$.

TABLE I. CVEs WITH COMPUTED IMPACT ON THE TEST TARGET SYSTEM

CVE id	Root cause	\tilde{G} size	CVSS score	Rel(t)	Impact Estimation
CVE-2017-9050	xmlDictAddString	0	5.0	0.9	0
CVE-2016-4448	xmlRelaxNGGetErrorString	130	10	0.8	104
CVE-2015-8710	htmlParseComment	88	7.5	0.7	46.2
CVE-2017-9049	xmlDictComputeFastKey	0	5.0	0.9	0
CVE-2017-9048	xmlSprintfElementContent	1364	7.5	0.9	920

During risk assessment and impact evaluation we consider next units: exploit availability $Expl$, type of

affected binary Src (open source or proprietary), release notes and patch existence $Patch$. Also we take into account version matching V and time coefficient $Rel_{(t)}$. Some parameters from static vulnerability characteristics X_i , such as CVSS score S , attack type T_a and vulnerability type T_v , also included during final impact computation.

Finally, we estimate weakness W of product P by applying additive and multiplicative convolution operations on certain types of measures. Type of convolution could be different and refined to the present time by experimental results.

IV. EXPERIMENTS AND RESULTS

Using AVIA in practice. Our technique turns out to be effective. To verify efficiency we run our implementation proof-of-concept system over 50 CVEs on libxml2 component. 15% of CVEs were not applicable to system because their description doesn't have any mentions about functions names of vulnerable function or they were not present in built shared object.

Over 20% of CVEs have vulnerable functions in the defined shared object, but these functional unused on the target system. 55% we marked as high impact on the system because of cross binary control flow which was found and we get full control flow from entry point of executable binary to the vulnerable function. The remaining 10% of CVEs were marked with medium impact on system, because cross binary control flow was found between shared objects without executable binaries. Table 1 shows brief overview of our results. We take one of the most popular open source library libxml.so and estimate CVEs vulnerability impact on our target test system. At the moment for impact assessment we also take into consideration expert's review of the investigated set of security issues. All verification problems will be considered in a further investigation.

Limitations of AVIA. The present design and implementation of AVIA has several limitations. In our research we consider to work only with executables that are dynamically linked, not stripped and contains information necessary for the dynamic linker. Work with obfuscated, stripped binaries is beyond the scope of the present paper and left as problem for future research.

This method not applicable yet for Linux kernel and kernel objects assessment. Current implementation only deals with arm-based device architectures. Nevertheless porting to arm is not difficult. Part with semantic CVE information retrieving is not fully implemented yet.

V. CONCLUSIONS AND FUTURE WORK

In this paper, for the first time, a new model and method for computing the vulnerability impact and risk assessment on an end-point computer system was designed with taking into account the properties of the vulnerability, the news context, and product features. Our study could be combined

with fuzzing techniques which aims to generate proof-of-concept exploits and help to save a lot of time avoiding redundant runs. Future work will include solutions of the above-mentioned problems and improving method of assessing the vulnerability exposure. The next steps in research is to supplement deep learning for semantic information retrieving, resolve indirect calls in binary analysis, provide experiments using widely range of CVEs.

REFERENCES

- [1] *Dirty Cow (CVE-2016-5195)*, <https://dirtycow.ninja/>
- [2] *Meltdown and Spectre*, <https://meltdownattack.com/>
- [3] *NIST, NVD Statistics Results*, https://nvd.nist.gov/vuln/search/statistics?adv_search=false&form_type=basic&results_type=statistics&search_type=all
- [4] P. John, *Cyber Security Trends: Aiming Ahead of the Target to Increase Security in 2017*, Tech. Rep., SANS Institute, 2017
- [5] *Common Vulnerability Scoring System*, <https://www.first.org/cvss/>
- [6] *OWASP Risk Ranking Methodology*, https://www.owasp.org/index.php/OWASP_Risk_Rating_Methodology
- [7] *OWASP Vulnerability Scanning Tools*, https://www.owasp.org/index.php/Category:Vulnerability_Scanning_Tools
- [8] D. Kennedy, J. O'Gorman, D. Kearns, and M. Aharoni, "Metasploit, The Penetration Tester's Guide", No Strach Press, San Francisco, 2011
- [9] A. Singh, "Metasploit Penetration Testing Cookbook", Packt Publishing, Birmingham, UK, 2012
- [10] *Lynis, Open Source Auditing*, <https://cisofy.com/lynis/>
- [11] *Cyber Grand Challenge*, <https://www.darpa.mil/program/cyber-grand-challenge>
- [12] S. K. Cha, T. Avreginos, A. Rebert, and D. Brumley. *Unleashing Mayhem on Binary Code*. In *Proceedings of the IEEE Symposium on Security and Privacy*, pages 380-394, 2012
- [13] T. Avreginos, S. K. Cha, B. L. Tze Hao, and D. Brumley. *AEG – Automatic Exploit Generation*. In *Proceedings of the 18th Annual Network and Distributed System Security Symposium (NDSS)*, 2011
- [14] L. Sanguino, R. Uetz, "Software Vulnerability Analysis Using CPE and CVE", arXiv preprint arXiv:1705.05347
- [15] *CVE MITRE*, <https://cve.mitre.org/>
- [16] S. Neuhaus, T. Zimmermann, "Security Trend Analysis with CVE Topic Models" Department of Computer Science, University of Calgary, Tech. Rep. 2010-970-19, 2010
- [17] S. Zhang, D. Caragea, and X. Ou, "An Empirical Study on Using the National Vulnerability Database to Predict Software Vulnerabilities", in *The 2011 International Conference on Security and Management (SAM'11)*, 2011
- [18] You, Wei, et al. "SemFuzz: Semantics-based Automatic Generation of Proof-of-Concept Exploits." *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*. ACM, 2017.
- [19] Li, Zhen, et al. "VulDeePecker: A Deep Learning-Based System for Vulnerability Detection." *arXiv preprint arXiv:1801.01681* (2018).
- [20] Viega, John, et al. "ITS4: A static vulnerability scanner for C and C++ code." *Computer Security Applications, 2000. ACSAC'00. 16th Annual Conference*. IEEE, 2000.
- [21] Shoshitaishvili, Yan, et al. "Sok:(state of) the art of war: Offensive techniques in binary analysis." *Security and Privacy (SP), 2016 IEEE Symposium on*. IEEE, 2016.
- [22] Capstone, , <https://www.capstone-engine.org/>
- [23] Youngdale, Eric. "Kernel korner: The ELF object file format by dissection." *Linux Journal* 1995.13es (1995): 15.
- [24] *IDA-Pro*, <https://www.hex-rays.com/index.shtml>
- [25] *Security Development Lifecycle*, <https://www.microsoft.com/en-us/sdl>
- [26] *LD.SO*, <http://man7.org/linux/man-pages/man8/ld.so.8.html>

A Fuzzy Model of Television Rating Control with Trend Rules Tuning based on Monitoring Results

Olexiy Azarov
Computer Facilities Department
Vinnytsia National Technical
University
Vinnytsia, Ukraine
azarov2@vntu.edu.ua

Leonid Krupelnitsky
Computer Facilities Department
Vinnytsia National Technical
University
Vinnytsia, Ukraine
krupost@gmail.com

Hanna Rakytyanska
Soft Ware Design Department
Vinnytsia National Technical
University
Vinnytsia, Ukraine
h_rakit@ukr.net

Abstract—The problem of constructing the recommendation rules for the television rating control is considered. A hybrid approach combining the benefits of semantic training and fuzzy relational equations in simplification of the process of expert recommendation systems construction is proposed. The problem of retaining the television rating can be attributed to the problems of fuzzy resources control. The trends of demand-supply relationships are described by the primary fuzzy knowledge bases. Rules refinement by solving the primary system of fuzzy relational equations allows avoiding labor-intensive procedures for the generation and selection of expert rules.

Keywords—TV channel rating, expert recommendation systems, fuzzy resources control, fuzzy classification knowledge bases, solving fuzzy relational equations

I. INTRODUCTION

The top priority task of the personnel of TV companies is to control the ratio of programs of different genres when forming the broadcast grid in order to increase and maintain the rating of the channel [1]. The rating of a television channel is determined by specialized sociological services and directly affects the cost of advertising time. Content management is modeled by integrating trials, expert recommendations and users' preferences [2]. The recommendation accuracy strongly depends on the mechanisms of the supply and demand regulation [3, 4]. Parametric statistical models are widely used to evaluate the viewers' demand and the popularity of TV programs [5–7]. These models are adjusted to fit the distribution of user ratings in video on demand services dealing specifically with TV content [5]. To enhance the recommendation accuracy, the statistical models aim to learn audience preferences that follow from the rich user content generated in the social networks [6, 7]. The timing and item recommendations are generated via clustering the common interests of a group of people [8, 9]. Finally, the cognitive models describe the behavior of viewers when choosing a TV channel [10, 11]. Such models predict program commitments based on viewer-program emotional relationships reflecting satisfaction and perception toward alternative programs.

The problem of retaining the television rating can be attributed to the problems of fuzzy resources control [1, 12]. In such models, the “demand-supply” relationships are described by fuzzy IF-THEN rules. Experienced managers

make effective administrative decisions based on a comparison of the viewers' demand for the programs of different genres with the rating of the programs offered at the given time [2]. Dependent upon this, a control action is formed, which consists of increasing or decreasing the rating of the programs in the channel broadcast grid. In works [12, 13], it is suggested to build the fuzzy resources control model on the grounds of the general method of nonlinear dependencies identification by means of fuzzy knowledge bases. The method [12, 13] implies the stage of tuning the fuzzy control model using “demand-supply” training data. The tuning stage consists of finding such fuzzy rules weights and such membership functions forms, which provide maximum proximity of the results of fuzzy logic inference to the correct managerial decisions.

The construction of expert recommendation systems is associated with computational costs. Constantly changing preferences of different categories of viewers require the selection and adjustment of the appropriate set of expert rules. At the same time, experts establish the trends of demand-supply relationships, which are subject to further refinement. Such trend dependencies are described by primary fuzzy knowledge bases. The solution to the problem of expert rules refinement may be the use of fuzzy relational equations [14–17], the solutions of which represent the linguistic modification of the primary terms. The obtained solutions can be considered as composite fuzzy rules that connect significance measures of the primary fuzzy terms [18, 19]. The number of rules in the class is determined by the number of solutions, and the form of the membership functions of the composite terms in the rule is determined by significance measures of the primary terms. Refinement of the rule set by solving the primary system of fuzzy relational equations allows avoiding labor-intensive procedures for the generation and selection of expert rules. Therefore, it is important to develop a hybrid approach combining the benefits of semantic training [12, 13] and fuzzy relational equations [14–17] in simplification of the process of expert recommendation systems construction. Following the approach proposed in [14–17], the genetic algorithm is used for tuning the primary fuzzy model and solving the primary system of fuzzy relational equations.

II. STRUCTURE OF THE TV RATING CONTROL MODEL

The fuzzy model of resources management was constructed using the example of the Ukrainian television

channel, presenting programs of such basic genres: political programs and news releases ($s=1$); TV serials ($s=2$); entertaining and sports programs ($s=3$) [20].

For the TV rating control problem, the monitoring and forecast window is a week. The analysis of the TV channel rating is carried out according to the results of monitoring of the TV programs ratings obtained for the previous week. The TV program is compiled for the forthcoming week. The structure of the TV rating control model corresponds to the following hierarchical tree of logic inference [12, 13]:

$$y_s(t, p) = f_s(x_s(t, p), z_s(t, p-1)), s = \overline{1, K}, \quad (1)$$

$$u(t, p) = f_0(z_1(y_1(t, p)), \dots, z_K(y_K(t, p))), \quad (2)$$

where $x_s(t, p)$ is the viewers' demand for the programs of the genre s at the time moment t of the p -th week; $z_s(t, p-1)$ is the rating of the offered programs of the genre s at the time moment t of the $(p-1)$ -th week; $y_s(t, p)$ is a control action for the time moment t of the p -th week, consisting in increasing–decreasing the rating of the offered programs of the genre s ; K is the number of genres of the proposed TV programs; $u(t, p)$ is the rating of the TV channel at the time moment t of the p -th week.

It is supposed that the control action is determined as the difference between the rating values before and after control, i.e. $y_s(t, p) = z_s(t, p) - z_s(t, p-1)$.

The proportion of TV viewers who watch TV at the time moment (t, p) determines the rating of the TV channel. In accordance with expert estimates, the range of the parameters $x_s(t, p)$, $z_s(t, p)$ and $u(t, p)$ is $[0, 20]$ %. The parameter $y_s(t)$ range is $[-20, 20]$ %.

We shall describe the trend dependencies with the help of the primary fuzzy terms: increased (decreased) (I, D) or stable (St) for $x_s(t, p)$, $z_s(t, p)$ and $u(t, p)$; increase (decrease) (I, D) or stay inactive (N) for $y_s(t, p)$. For the composite terms construction, we shall use the linguistic modifiers: sharply (sh), moderately (m), weakly (w). These modifiers describe the semantic intensity of the primary terms D and I [17].

Functional dependencies (1) and (2) are defined by the primary fuzzy relations and rules presented in Table I and Table II, respectively.

TABLE I. PRIMARY FUZZY RULES FOR CONTROL ACTIONS IN EACH GENRE CATEGORY

Primary rule	IF		THEN
	$x_s(t, p)$	$z_s(t, p-1)$	$y_s(t, p)$
H_1	D	I	D
H_2	D	St	
H_3	St	I	
H_4	D	D	N
H_5	St	St	
H_6	I	I	
H_7	St	D	I
H_8	I	St	
H_9	I	D	

It is necessary to transfer the primary fuzzy relations and rules into the composite ones for the modified decision classes of the variables $y_s(t, p)$ and $u(t, p)$.

TABLE II. PRIMARY FUZZY RELATIONS FOR TV RATING CLASSES

IF	THEN $u(t)$			
	D	St	I	
$z_s(t, p)$	D	$m-sh$	w	-
	St	w	$m-sh$	$w-m$
	I	-	$w-m$	$m-sh$

III. THE PROBLEM OF TUNING THE FUZZY CONTROL MODEL

Correlations (1), (2) define the primary fuzzy control model in the form:

$$\mu_y^s = f_s(\mu_x^s, \mu_z^s, \mathbf{W}_s), \quad (3)$$

$$\mu_u = f_0(\mu_z^1, \dots, \mu_z^K, \mathbf{R}), \quad (4)$$

where $\mu_x^s = (\mu_x^{s,D}, \mu_x^{s,St}, \mu_x^{s,I})$, $\mu_z^s = (\mu_z^{s,D}, \mu_z^{s,St}, \mu_z^{s,I})$, $\mu_y^s = (\mu_y^{s,D}, \mu_y^{s,N}, \mu_y^{s,I})$ are the vectors of significance measures of the primary fuzzy terms of the variables x_s , z_s and y_s in correlation (1); $\mu_z^s = (\mu_z^{s,D}, \mu_z^{s,St}, \mu_z^{s,I})$ and $\mu_u = (\mu^D, \mu^{St}, \mu^I)$ are the vectors of significance measures of the primary fuzzy terms of the variables z_s and u in correlation (2); $\mathbf{W}_s = (w_1^s, \dots, w_6^s)$ is the primary rules vector in correlation (1); $\mathbf{R} = (r_{11}^s, \dots, r_{13}^s, \dots, r_{31}^s, \dots, r_{33}^s)$ is the primary fuzzy relational matrix for genre preferences in correlation (2).

We use a bell-shaped membership function model of variable v to arbitrary term T in the form [12, 13]:

$$\mu^T(v) = 1 / (1 + ((v - \beta) / \sigma)^2),$$

where β is a coordinate of function maximum, $\mu^T(\beta) = 1$; σ is a parameter of concentration.

In this case, correlations (3), (4) take the form [14–17]:

$$\mu_y^s(y_s, \mathbf{B}_y^s, \Omega_y^s) = f_s(x_s, z_s, \mathbf{W}_s, \mathbf{B}_x^s, \Omega_x^s, \mathbf{B}_z^s, \Omega_z^s), \quad (5)$$

$$\mu_u(u, \mathbf{B}_u, \Omega_u) = f_0(\mu_z^1, \dots, \mu_z^K, \mathbf{R}), \quad (6)$$

where $\mathbf{B}_y^s = (\beta_y^{s,D}, \beta_y^{s,N}, \beta_y^{s,I})$, $\Omega_y^s = (\sigma_y^{s,D}, \sigma_y^{s,N}, \sigma_y^{s,I})$, $\mathbf{B}_x^s = (\beta_x^{s,D}, \beta_x^{s,St}, \beta_x^{s,I})$, $\Omega_x^s = (\sigma_x^{s,D}, \sigma_x^{s,St}, \sigma_x^{s,I})$, $\mathbf{B}_z^s = (\beta_z^{s,D}, \beta_z^{s,St}, \beta_z^{s,I})$, $\Omega_z^s = (\sigma_z^{s,D}, \sigma_z^{s,St}, \sigma_z^{s,I})$, $\mathbf{B}_u = (\beta_u^D, \beta_u^{St}, \beta_u^I)$, $\Omega_u = (\sigma_u^D, \sigma_u^{St}, \sigma_u^I)$ are the vectors of membership functions parameters of the primary fuzzy terms of the variables x_s , z_s , y_s and u .

It is assumed that some training data sample in the form of P pairs of experimental data can be obtained on the ground of successful managerial decisions $\langle \hat{x}_s^l, \hat{z}_s^l, \hat{y}_s^l \rangle$,

\hat{u}^l), $l=1, P$, where \hat{x}_s^l and \hat{z}_s^l are the control system state parameters in the experiment number l ; \hat{y}_s^l and \hat{u}^l are the control action and TV rating in the experiment number l . The essence of the fuzzy model (5), (6) tuning is as follows. It is necessary to find the relation matrix \mathbf{R} , the rules weights vector \mathbf{W} and the vectors of the membership functions parameters $\mathbf{B}_x^s, \mathbf{\Omega}_x^s, \mathbf{B}_z^s, \mathbf{\Omega}_z^s, \mathbf{B}_y^s, \mathbf{\Omega}_y^s, \mathbf{B}_u, \mathbf{\Omega}_u$, which provide the minimum distance between theoretical and experimental data:

$$\sum_{l=1}^P [f_s(\hat{x}_s^l, \hat{z}_s^l, \mathbf{W}_s, \mathbf{B}_x^s, \mathbf{\Omega}_x^s, \mathbf{B}_z^s, \mathbf{\Omega}_z^s) - \hat{y}_s^l(\hat{y}_s^l, \mathbf{B}_y^s, \mathbf{\Omega}_y^s)]^2 = \min_{\mathbf{W}, \mathbf{B}, \mathbf{\Omega}} \quad (7)$$

$$\sum_{l=1}^P [f_0(\hat{\mu}_Z^1, \dots, \hat{\mu}_Z^K, \mathbf{R}) - \hat{\mu}_u(\hat{u}^l, \mathbf{B}_u, \mathbf{\Omega}_u)]^2 = \min_{\mathbf{R}} \quad (8)$$

We shall denote: $\{D_1, \dots, D_M\}$ and $\{d_1, \dots, d_m\}$ are the modified decision classes of the variables $u(t, p)$ and $y_s(t, p)$, respectively. Given the primary fuzzy model and qualitative output values, the problem of tuning the composite fuzzy model is formulated as follows [14–17]. For each output class $D_J, J=1, M$, the fuzzy causes vector $\mu_{Z,J}^s$ should be found which provides the least distance between observed and model fuzzy effects vectors in correlation (4), and for each output class $d_j, j=1, m$, the fuzzy causes vectors $\mu_{x,j}^s, \mu_{z,j}^s$ should be found which provide the least distance between observed and model fuzzy effects vectors in correlation (3):

$$[f_0(\mu_{Z,J}^1, \dots, \mu_{Z,J}^K, \mathbf{R}) - \mu_u(D_J)]^2 = \min_{\mu_{Z,J}^1, \dots, \mu_{Z,J}^K} \quad (9)$$

$$[f_s(\mu_{x,j}^s, \mu_{z,j}^s, \mathbf{W}_s) - \mu_y^s(d_j^s)]^2 = \min_{\mu_{x,j}^s, \mu_{z,j}^s} \quad (10)$$

The genetic algorithm is used for solving the optimization problems (7)–(10) of tuning the primary fuzzy model and rule-based solutions of primary fuzzy relational equations.

IV. SOLVING FUZZY RELATIONAL EQUATIONS

Let us consider the construction of composite rules for the rating $u(t, p)$ and control action $y_s(t, p), s=1$. The primary system of fuzzy relational equations after tuning has the form:

$$\begin{aligned} \mu^{E_1} &= (\mu^{C_{11}} \wedge 0.90) \vee (\mu^{C_{12}} \wedge 0.30) \vee (\mu^{C_{21}} \wedge 0.70) \vee \\ &\vee (\mu^{C_{22}} \wedge 0.28) \vee (\mu^{C_{31}} \wedge 0.67) \vee (\mu^{C_{32}} \wedge 0.26), \\ \mu^{E_2} &= (\mu^{C_{11}} \wedge 0.30) \vee (\mu^{C_{12}} \wedge 0.93) \vee (\mu^{C_{13}} \wedge 0.46) \vee \\ &\vee (\mu^{C_{21}} \wedge 0.25) \vee (\mu^{C_{22}} \wedge 0.86) \vee (\mu^{C_{23}} \wedge 0.45) \vee \\ &\vee (\mu^{C_{31}} \wedge 0.29) \vee (\mu^{C_{32}} \wedge 0.62) \vee (\mu^{C_{33}} \wedge 0.45), \end{aligned}$$

$$\begin{aligned} \mu^{E_3} &= (\mu^{C_{12}} \wedge 0.50) \vee (\mu^{C_{13}} \wedge 0.96) \vee (\mu^{C_{22}} \wedge 0.50) \vee \\ &\vee (\mu^{C_{23}} \wedge 0.78) \vee (\mu^{C_{32}} \wedge 0.48) \vee (\mu^{C_{33}} \wedge 0.81), \quad (11) \end{aligned}$$

$$\begin{aligned} \mu^{e_1} &= (\mu^{H_1} \wedge 0.89) \vee (\mu^{H_2} \wedge 0.75) \vee (\mu^{H_3} \wedge 0.82) \\ \mu^{e_2} &= (\mu^{H_2} \wedge 0.49) \vee (\mu^{H_3} \wedge 0.45) \vee (\mu^{H_4} \wedge 0.79) \vee \\ &\vee (\mu^{H_5} \wedge 0.90) \vee (\mu^{H_6} \wedge 0.86) \vee (\mu^{H_7} \wedge 0.51) \vee (\mu^{H_8} \wedge 0.54) \end{aligned}$$

$$\mu^{e_3} = (\mu^{H_7} \wedge 0.78) \vee (\mu^{H_8} \wedge 0.80) \vee (\mu^{H_9} \wedge 0.93), \quad (12)$$

$$\begin{aligned} \mu^{H_1} &= \mu^{c_{11}} \wedge \mu^{c_{23}}, \\ \mu^{H_2} &= \mu^{c_{11}} \wedge \mu^{c_{22}}, \\ \mu^{H_3} &= \mu^{c_{12}} \wedge \mu^{c_{23}}, \\ \mu^{H_4} &= \mu^{c_{11}} \wedge \mu^{c_{21}}, \\ \mu^{H_5} &= \mu^{c_{12}} \wedge \mu^{c_{22}}, \\ \mu^{H_6} &= \mu^{c_{13}} \wedge \mu^{c_{23}}, \\ \mu^{H_7} &= \mu^{c_{12}} \wedge \mu^{c_{21}}, \\ \mu^{H_8} &= \mu^{c_{13}} \wedge \mu^{c_{22}}, \\ \mu^{H_9} &= \mu^{c_{13}} \wedge \mu^{c_{21}}, \quad (13) \end{aligned}$$

where variables $u(t, p)$ and $z_s(t, p)$ were described by fuzzy terms $E_1 \div E_3$ and $C_{s1} \div C_{s3} (D, St, I)$; variables $y_s(t, p), x_s(t, p)$ and $z_s(t, p-1)$ were described by fuzzy terms $e_1 \div e_3 (D, N, I), c_{11} \div c_{13}$ and $c_{21} \div c_{23} (D, St, I)$.

The composite rules were built for the classes $D_1 \div D_7 (shD, mD, wD, St, wI, mI, shI)$ and $d_1 \div d_7 (shD, mD, wD, N, wI, mI, shI)$. Fuzzy effects vectors were defined with the help of the membership functions of the fuzzy terms $E_1 \div E_3, D_1 \div D_7$ in Fig. 1,a and $e_1 \div e_3, d_1 \div d_7$ in Fig. 1,b:

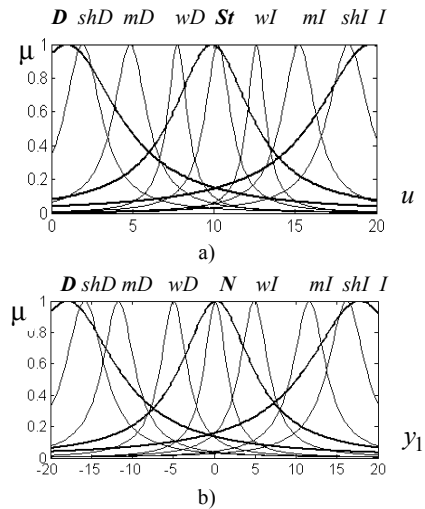


Fig. 1. Membership functions of the fuzzy terms for the variables $u(t, p)$ (a) and $y_1(t, p)$ (b)

$$\begin{aligned} \mu^E(D_1) &= (0.96, 0.22, 0.10); \quad \mu^e(d_1) = (0.95, 0.20, 0.09); \\ \mu^E(D_2) &= (0.63, 0.40, 0.11); \quad \mu^e(d_2) = (0.68, 0.29, 0.10); \end{aligned}$$

$$\begin{aligned} \mu^E(D_3) &= (0.30, 0.75, 0.14); \mu^e(d_3) = (0.33, 0.65, 0.16); \\ \mu^E(D_4) &= (0.20, 0.99, 0.21); \mu^e(d_4) = (0.18, 0.99, 0.22); \\ \mu^E(D_5) &= (0.12, 0.62, 0.30); \mu^e(d_5) = (0.14, 0.71, 0.37); \\ \mu^E(D_6) &= (0.11, 0.37, 0.58); \mu^e(d_6) = (0.10, 0.31, 0.73); \\ \mu^E(D_7) &= (0.10, 0.21, 0.94); \mu^e(d_7) = (0.08, 0.20, 0.98). \end{aligned}$$

The solution set of the fuzzy relational equations (11) is presented in Table 3. The linguistic interpretation of the obtained solutions in the form of the “single input – single output” fuzzy rules is presented in Table 4 [18].

TABLE III. SOLUTIONS OF FUZZY RELATIONAL EQUATIONS FOR TV RATING CLASSES

IF	THEN $u(t)$						
	<i>shD</i>	<i>mD</i>	<i>wD</i>	<i>St</i>	<i>wI</i>	<i>mI</i>	<i>shI</i>
$\mu^{C_{11}}$	0.90–1	0.63	0.30	0.21	0.12	0.11	0.10
$\mu^{C_{12}}$	0.10	0.11	0.30–0.75	0.93–1	0.30–0.62	0.11	0.10
$\mu^{C_{13}}$	0.10	0.11	0.14	0.21	0.50	0.58	0.94
$\mu^{C_{21}}$	0.25–1	0.25–0.63	0.30	0.21	0.12	0.11	0.10
$\mu^{C_{22}}$	0.10	0.11	0.30–0.75	0.30–1	0.30–0.62	0.11	0.10
$\mu^{C_{23}}$	0.10	0.11	0.14	0.21	0.50	0.37–0.58	0.45–1
$\mu^{C_{31}}$	0.29–1	0.29–0.63	0.30	0.21	0.12	0.11	0.10
$\mu^{C_{32}}$	0.10	0.11	0.30–1	0.30–1	0.62–1	0.11	0.10
$\mu^{C_{33}}$	0.10	0.11	0.14	0.21	0.50	0.37–0.58	0.45–1

TABLE IV. COMPOSITE FUZZY RELATIONS FOR TV RATING CLASSES

IF		THEN $u(t)$						
		<i>shD</i>	<i>mD</i>	<i>wD</i>	<i>St</i>	<i>wI</i>	<i>mI</i>	<i>shI</i>
$z_1(t, p)$	<i>shD</i>	0.90	0	0	0	0	0	0
	<i>mD</i>	0	0.63	0.30	0	0	0	0
	<i>wD</i>	0	0	0.75	0.21	0	0	0
	<i>St</i>	0	0	0	0.93	0	0	0
	<i>wI</i>	0	0	0	0.21	0.62	0	0
	<i>mI</i>	0	0	0	0	0.50	0.58	0
	<i>shI</i>	0	0	0	0	0	0	0.94
$z_2(t, p)$	<i>shD</i>	1	0	0	0	0	0	0
	<i>mD</i>	0.70	0.63	0.30	0.30	0	0	0
	<i>wD</i>	0.25	0.25	0.75	0.50	0	0	0
	<i>St</i>	0	0	0	1	0	0	0
	<i>wI</i>	0	0	0	0.50	0.62	0.37	0
	<i>mI</i>	0	0	0	0.30	0.50	0.58	0.45
	<i>shI</i>	0	0	0	0	0	0	0.78
$z_3(t, p)$	<i>shD</i>	1	0	0	0	0	0	0
	<i>mD</i>	0.67	0.63	0.30	0.30	0	0	0
	<i>wD</i>	0.29	0.29	0.62	0.62	0.30	0	0
	<i>St</i>	0	0	1	1	1	0	0
	<i>wI</i>	0	0	0.30	0.62	0.62	0.37	0
	<i>mI</i>	0	0	0	0.30	0.50	0.58	0.45
	<i>shI</i>	0	0	0	0	0	0	0.81

The solution set of the fuzzy relational equations (12), (13) is presented in Table 5. The set of interval rules and linguistic interpretation of the obtained solutions in the form

of the “multiple inputs – single output” fuzzy rules is presented in Table 6 [19].

TABLE V. SOLUTIONS OF FUZZY RELATIONAL EQUATIONS FOR CONTROL ACTIONS

IF						THEN
$x_1(t, p)$			$z_1(t, p-1)$			$y_1(t, p)$
$\mu^{c_{11}}$	$\mu^{c_{12}}$	$\mu^{c_{13}}$	$\mu^{c_{21}}$	$\mu^{c_{22}}$	$\mu^{c_{23}}$	
0.89–1	0–0.20	0–0.09	0.09	0–0.20	0.89–1	d_1
0.68–1	0.10	0–0.10	0.49	0–0.49	0.68–1	d_2
0.68–1	0.49	0–0.10	0.10	0.68–1	0–0.49	
0–0.49	0.68–1	0.49	0–0.10	0.10	0.68–1	
0.65–1	0.16	0–0.16	0.65–1	0.33	0–0.33	d_3
0.33	0.65–1	0–0.16	0.16	0.65–1	0–0.33	
0–0.33	0.33	0.65–1	0–0.16	0.16	0.65–1	
0.90–1	0.22	0–0.22	0.90–1	0.18	0–0.18	d_4
0–0.18	0.90–1	0–0.22	0–0.22	0.90–1	0–0.18	
0–0.18	0.18	0.90–1	0–0.22	0.22	0.90–1	
0.71–1	0.37	0–0.37	0.71–1	0.14	0–0.14	d_5
0–0.14	0.71–1	0.37	0–0.37	0.71–1	0.14	
0–0.14	0.14	0.71–1	0–0.37	0.37	0.71–1	
0.10	0.73–1	0–0.54	0.73–1	0.54	0–0.10	d_6
0–0.10	0.54	0.73–1	0–0.54	0.73–1	0.10	
0–0.10	0.10	0.73–1	0.73–1	0–0.54	0.54	
0–0.08	0–0.20	0.93–1	0.93–1	0–0.20	0.08	d_7

The lower and upper bounds of the interval rules were obtained with the help of the membership functions of the primary fuzzy terms. The membership functions of the primary and composite fuzzy terms for input variables are shown in Fig. 2.

The composite fuzzy rules for relations $y_2(t, p)$ and $y_3(t, p)$ were tuned in a similar way.

TABLE VI. COMPOSITE FUZZY RULES FOR CONTROL ACTIONS

Primary rule	IF		THEN
	$x_1(t, p)$	$z_1(t, p-1)$	$y_1(t, p)$
H_1	0–3.34, <i>shD</i>	16.35–20, <i>shI</i>	d_1
H_1	5.10, <i>mD</i>	14.60–20, <i>mI–shI</i>	d_2
H_2	0–5.10, <i>mD–shD</i>	14.60, <i>mI</i>	
H_3	0–5.10, <i>mD–shD</i>	7.35–12.27, <i>wD–wI</i>	
H_4	8.00–12.57, <i>wD–wI</i>	14.60–20, <i>mI–shI</i>	
H_4	0–5.40, <i>mD–shD</i>	5.42, <i>mD</i>	d_3
H_5	7.86–12.43, <i>wD–wI</i>	12.43, <i>wI</i>	
H_5	7.86, <i>wD</i>	7.18–12.43, <i>wD–wI</i>	
H_6	14.35, <i>mI</i>	14.30–20, <i>mI–shI</i>	
H_4	0–3.27, <i>shD</i>	0–3.31, <i>shD</i>	d_4
H_5	9.20–11.43, <i>St</i>	8.63–11.00, <i>St</i>	
H_6	16.45–20, <i>shI</i>	16.40–20, <i>shI</i>	
H_4	4.90, <i>mD</i>	0–4.96, <i>mD–shD</i>	d_5
H_5	8.16–12.45, <i>wD–wI</i>	7.50, <i>wD</i>	
H_5	12.45, <i>wI</i>	7.50–12.10, <i>wD–wI</i>	
H_6	14.87–20, <i>mI–shI</i>	14.79, <i>mI</i>	
H_7	8.26–12.32, <i>wD–wI</i>	0–4.76, <i>mD–shD</i>	d_6
H_8	15.00–20, <i>mI–shI</i>	7.61–12.00, <i>wD–wI</i>	
H_9	15.00, <i>mI</i>	0–4.76, <i>mD–shD</i>	
H_9	15.00–20, <i>mI–shI</i>	4.76, <i>mD</i>	
H_9	16.75–20, <i>shI</i>	0–3.05, <i>shD</i>	d_7

V. EXAMPLE OF THE TV CHANNEL RATING CONTROL

The values ⟨viewers’ demand $\hat{x}_s(t, p)$, supply before and after control $\hat{z}_s(t, p-1)$, $\hat{z}_s(t, p)$, control action $\hat{y}_s(t, p) = \hat{z}_s(t, p) - \hat{z}_s(t, p-1)$, rating $\hat{u}(t, p)$ ⟩, corresponding to the experienced manager actions were taken as the training data sample. In this case, the TV rating was maintained at a

consistently high level, and the unmet viewers' demand was reduced to a minimum.

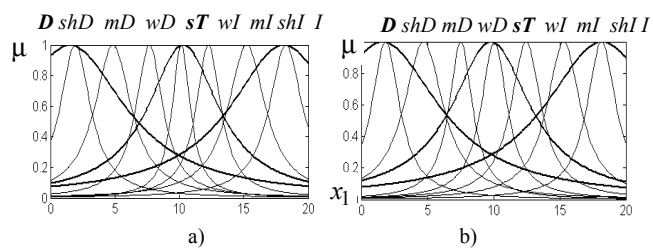


Fig. 2. Membership functions of the fuzzy terms for the input variables $x_s(t, p)$ (a) and $z_s(t, p)$ (b)

The unmet demand after control for each genre category can be defined as $\delta_s(t, p) = z_s(t, p) - x_s(t, p)$. The experimental value $\hat{x}_s(t, p)$ for each genre category is defined as the ratio of viewers who watch such programs broadcast by all TV channels at the time moment (t, p) . Evaluation of the rating of TV programs in the channel broadcasting network is carried out with the help of the authors' monitoring system [21]. The experimental values $\hat{z}_s(t, p-1)$, $\hat{z}_s(t, p)$ and $\hat{u}(t, p)$ are determined on the basis of the weekly top 20 rating [20]. The sample includes data from 2015 to 2017. The training sample fragment is presented in Fig. 3 in the form of the dynamics of the demand and supply change for each genre during the day. Management is carried out at the level of each air-hour, i.e. $t \in [0...24]$. Comparison of the model and experimental control, as well as the unmet demand after control is presented in Fig. 4.

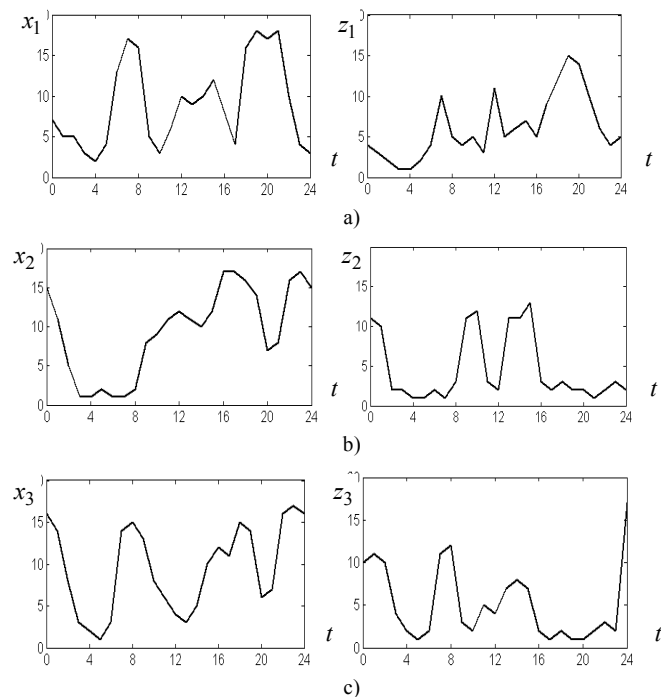


Fig. 3. The training sample fragment: the dynamics of the demand and supply change for genres $s=1$ (a); $s=2$ (b); $s=3$ (c)

For the political genre, the offered programs have balanced the viewers' demand (Fig. 4, a). The demand for

the genre of television serials from 16 to 17 hours has been satisfied with the program of the sports and entertainment genre (Fig. 4, b). Some popular serials have been offered instead of the entertainment programs from 18 to 19 and from 22 to 23 hours (Fig. 4, c).

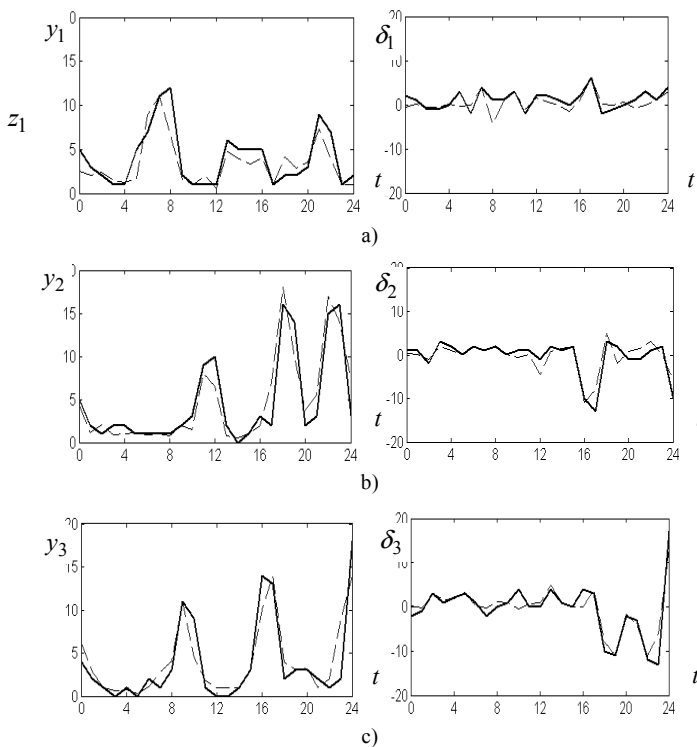


Fig. 4. The model (—) and experimental (---) control action and the unmet demand after control for genres $s=1$ (a); $s=2$ (b); $s=3$ (c)

Comparison of the model and experimental rating during the day is shown in Fig. 5, a. The dynamics of the average weekly rating change at the level of each air-hour for $p \in [1...14]$ weeks is shown in Fig 5, b. When compiling the average weekly rating, the time range is $t \in [8...23]$ hours, since the programs in the range $t \in [0...7]$ do not fall into the weekly top 20 rating.

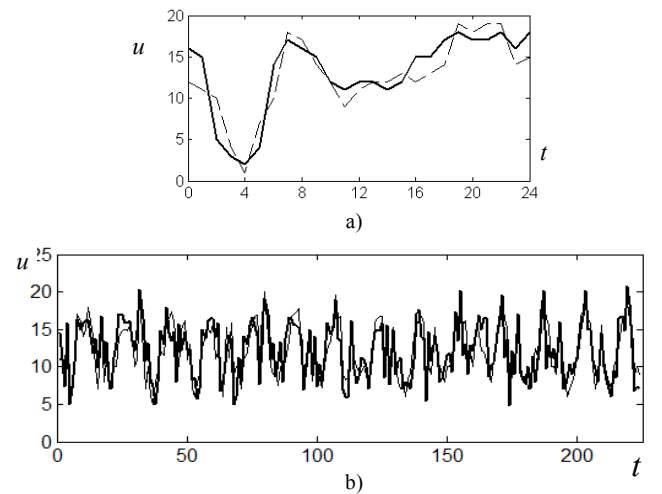


Fig. 5. The dynamics of the daily (a) and average weekly (b) TV rating change for the model (—) and experimental (---) control action

VI. CONCLUSIONS

The proposed approach can find application in the automated recommendation systems in the TV domain. Further research is required to develop models, which are tuned to predict the TV programs popularity among different demographic groups. The policy of family channels with diversified content categories implies the balanced ratings for all social and age groups. In this case, the timing and item recommendations should be augmented by genre preference relationships based on viewers' demographics. Besides that, supplementary factors influencing the demand and supply values (collaborative filtering, items similarity, purchase prices of new programs, advertising revenue) can be taken into account with the help of the primary relations and rules.

ACKNOWLEDGMENTS

The paper was prepared within the 58-D-369 "Technologies of the construction of intelligent analog-digital systems for monitoring and analysis of multimedia information" project.

REFERENCES

- [1] S. Žilic Fišer, *Successful Television Management: the Hybrid Approach*. Peter Lang GmbH: Int. Verlag der Wissenschaften, 2015.
- [2] D. Schuurman, L. D. Marez, P. Veevaete, and T. Evens, "Content and context for mobile television: integrating trial, expert and user findings," *Telematics and Informatics*, vol. 26 (3), pp. 293–305, August 2009.
- [3] D. Vêras, T. Prota, A. Bispo, R. Prudêncio, and C. Ferraz, "A literature review of recommender systems in the television domain," *Expert Systems with Applications*, vol. 42 (22), pp. 9046–9076, December 2015.
- [4] J. Oh, S. Kim, J. Kim, and H. Yu, "When to recommend: a new issue on TV show recommendation," *Information Sciences*, vol. 280, pp. 261–274, October 2014.
- [5] F. Fraile and J. C. Guerri, "Simple models of the content duration and the popularity of television content," *Journal of Network and Computer Applications*, vol. 40, pp. 12–20, April 2014.
- [6] Y. Zhang, W. Chen, and Z. Yin, "Collaborative filtering with social regularization for TV program recommendation," *Knowledge-Based Systems*, vol. 54, pp. 310–317, December 2013.
- [7] Y. Xu and J. Yin, "Collaborative recommendation with user generated content," *Eng. Appl. Artif. Intell.*, vol. 45, pp. 281–294, October 2015.
- [8] Z. Wang and L. He, "User identification for enhancing IP-TV recommendation," *Knowledge-Based Systems*, vol. 98, pp. 68–75, April 2016.
- [9] L. Boratto, S. Carta, and G. Fenu, "Investigating the role of the rating prediction task in granularity-based group recommender systems and big data scenarios," *Information Sciences*, vol. 378, pp. 424–443, February 2017.
- [10] E. Panova, A. Raikov, and O. Smirnova, "Cognitive television viewer rating," *Procedia Comput. Sci.*, vol. 66, pp. 328–335, 2015.
- [11] J.-S. Lin, Y. Sung, and K.-J. Chen, "Social television: examining the antecedents and consequences of connected TV viewing," *Computers in Human Behaviour*, vol. 58, pp. 171–178, May 2016.
- [12] A. Rotshtein and H. Rakytyanska, *Fuzzy Evidence in Identification, Forecasting and Diagnosis. Studies in Fuzziness and Soft Computing*, vol. 275, Heidelberg: Springer, 2012.
- [13] A. Rotshtein and A. Rakytyanskaya, "Inventory control as an identification problem based on fuzzy logic," *Cybernetics and Systems Analysis*, vol. 42 (3), pp. 411–419, May 2006.
- [14] A. Rotshtein and H. Rakytyanska, "Expert rules refinement by solving fuzzy relational equations," In *Proc. of the VIth IEEE Conference on Human System Interaction*. 6-8 June, 2013, Sopot, Poland, pp. 257–264, 2013.
- [15] A. Rotshtein and H. Rakytyanska, "Optimal design of rule-based systems by solving fuzzy relational equations," In: S. Hippe Z., L. Kulikowski J., Mroczek T., Wtorek J. (eds) *Issues and Challenges in Artificial Intelligence. Studies in Computational Intelligence*, vol. 559, pp. 167–178, Springer, 2014.
- [16] H. Rakytyanska, "Fuzzy classification knowledge base construction based on trend rules and inverse inference," *Eastern-European Journal of Enterprise Technologies*, vol. 1(3), pp. 25–32, 2015.
- [17] H. Rakytyanska, "Classification rule hierarchical tuning with linguistic modification based on solving fuzzy relational equations," *Eastern-European Journal of Enterprise Technologies*, vol. 1(4), pp. 50–58, 2018.
- [18] A. Rotshtein and H. Rakytyanska, "Adaptive diagnostic system based on fuzzy relations," *Cybernetics and Systems Analysis*, vol. 45(4), pp. 623–637, July 2009.
- [19] A. Rotshtein and H. Rakytyanska, "Fuzzy logic and the least squares method in diagnosis problem solving," In: Sarma R. (Ed.) *Genetic diagnoses*. New York: Nova Science Publishers, pp. 53–97, 2011.
- [20] <http://inter.ua/ru/about/rating>.
- [21] O. D. Azarov, L. V. Krupelnitsky, V. Y. Steiskal, and O. A. Bilokon', "Specialized and Measuring Equipment of Own Design and Production for TV and Radio Broadcasting," *Catalog of the Scientific and Technical Center «Analog-Digital Systems»*, Vinnitsya: VNTU, 2015. <http://ot.vntu.edu.ua/katalog>.

Mathematical Modeling of Two-Dimensional Deformation-Relaxation Processes in Environments with Fractal Structure

Yaroslav Sokolovskyy
Department of Information Technologies
Ukrainian National Forestry University, UNFU
Lviv, Ukraine
sokolovskyyar@yahoo.com

Mokrytska Olha
Department of Information Technologies
Ukrainian National Forestry University, UNFU
Lviv, Ukraine
mokrytska@nltu.edu.ua

Maryana Levkovych
Department of Information Technologies
Ukrainian National Forestry University, UNFU
Lviv, Ukraine
maryana.levkovych@gmail.com

Vitalij Atamanyuk
Department of Electromechanics and Electronics
Hetman Petro Sahaidachnyi National Army Academy, NAA
Lviv, Ukraine
atamanvitalv@gmail.com

Abstract—In the work, the general mathematical model of two-dimensional viscoelastic deformation using the fractional integro-differential apparatus is constructed. The relations in the differential and integral forms are given to present two-dimensional Kelvin's and Voigt's rheological models. The algorithm of a numerical method for solving the problem, based on the use of finite-difference schemes, was developed. The analytical expressions to describe deformations of one-dimensional fractal models are given, and on the basis of which the identification of fractional-differential parameters is carried out. The influence of fractal parameters on the dynamics of deformation and stress variation for different rheological models is investigated.

Keywords—two-dimensional mathematical model; derivative of fractional order; deformation-relaxation processes; numerical method; statistical criterion.

I. INTRODUCTION

Experimental and numerical studies have shown that many information, biological, physical, technological processes have a complex fractal structure. Mathematical models of processes of visco-elastic deformation and heat-mass transfer in the environments with a fractal structure, which are characterized by memory effects, spatial non-locality and self-organization, are based on the use of the mathematical apparatus of fractional integro-differential operators. Differential equations of fractional order describe the evolution of physical systems with residual memory which occupy an intermediate position between Markov's systems and systems that are characterized by total memory. In particular, the fractional index indicates the share of system states that are stored throughout the process of its operation. In the works [1], [2], the apparatus of fractional integro-differential operators is used to construct adequate mathematical models of heat-mass transfer processes and viscoelastic deformation in the environments with fractal structure. Also, an important factor is that the fractal parameters allow for a more complete description of different processes, as opposed to integer ones. The fractal analysis is used for analyses, decision-making and prediction of systems with complex and multiscale descriptions [3, 9].

One example of such systems is the human brain, which can be idealized as a complex dynamic system consisting of many interacting agents.

II. PROBLEM FORMULATION

The mathematical rheological model of two-dimensional viscoelastic deformation using the derivatives of fractional order is described by means of equilibrium equations with a fractional order γ ($0 < \gamma \leq 1$) by spatial coordinates x_1 and x_2 :

$$C_{11} \left(\bar{R}_{11} \frac{\partial^\gamma \varepsilon_{11}}{\partial x_1^\gamma} - \tilde{R}_{11} \right) + C_{12} \left(\bar{R}_{12} \frac{\partial^\gamma \varepsilon_{22}}{\partial x_1^\gamma} - \tilde{R}_{12} \right) + 2C_{33} \left(\bar{R}_{33}^2 \frac{\partial^\gamma \varepsilon_{12}}{\partial x_2^\gamma} - \tilde{R}_{33}^2 \right) = 0, \quad (1)$$

$$C_{21} \left(\bar{R}_{21} \frac{\partial^\gamma \varepsilon_{11}}{\partial x_2^\gamma} - \tilde{R}_{21} \right) + C_{22} \left(\bar{R}_{22} \frac{\partial^\gamma \varepsilon_{22}}{\partial x_2^\gamma} - \tilde{R}_{22} \right) + 2C_{33} \left(\bar{R}_{33}^1 \frac{\partial^\gamma \varepsilon_{12}}{\partial x_1^\gamma} - \tilde{R}_{33}^1 \right) = 0. \quad (2)$$

where \bar{R}_{ij} , \tilde{R}_{ij} are corresponding values of integrals:

$$\int_0^t R_{ij}(t-z, T, U) dz = \bar{R}_{ij},$$

$$\int_0^t R_{ij}(t-z, T, U) \frac{\partial^\gamma \varepsilon_{T1, T2}}{\partial x_k^\gamma} dz = \tilde{R}_{ij}, \quad (k=1, 2)$$

$$\int_0^t R_{ij}(t-z, T, U) \frac{\partial^\gamma \varepsilon_{T3}}{\partial x_2^\gamma} dz = \tilde{R}_{33}^2, \quad \int_0^t R_{ij}(t-z, T, U) \frac{\partial^\gamma \varepsilon_{T3}}{\partial x_1^\gamma} dz = \tilde{R}_{33}^1,$$

R_{ij} are the relaxation kernels of fractional-differential models which depend on time t , temperature T and moisture U ; $\varepsilon^T = (\varepsilon_{11}, \varepsilon_{22}, \varepsilon_{12})$ is vector of deformations, components of which are dependent on time t and spatial variables x_1 and x_2 , $(t, x_1, x_2) \in D, D = [0, \tilde{T}] \times [0, l_1] \times [0, l_2]$, $\varepsilon_T = (\varepsilon_{T1}, \varepsilon_{T2}, \varepsilon_{T3})^T$ is the vector of deformations, components of which are dependent on changes in temperature ΔT and moisture content ΔU :

$$\begin{aligned} \varepsilon_{T1} &= \alpha_{11}\Delta T + \beta_{11}\Delta U, & \varepsilon_{T2} &= \alpha_{22}\Delta T + \beta_{22}\Delta U, \\ \varepsilon_{T3} &= 0, \end{aligned}$$

$\alpha_{11}, \alpha_{22}, \beta_{11}, \beta_{22}$ are coefficients of thermal expansion and moisture-dependent shrinkage;

C_{ij} are components of the elastic tensor of an orthotropic body

$$\begin{aligned} C_{11} &= \frac{E_{11}}{(1-\nu_1\nu_2)}, & C_{12} &= \frac{\nu_2 E_{11}}{(1-\nu_1\nu_2)}, \\ C_{21} &= \frac{\nu_1 E_{22}}{(1-\nu_1\nu_2)}, & C_{22} &= \frac{E_{22}}{(1-\nu_1\nu_2)}, & 2C_{33} &= \mu, \end{aligned}$$

μ is the shear modulus in the plane $\left(\mu = \frac{E_{12}}{2(1-\nu_{12})}\right)$,

E_{11}, E_{22}, E_{12} are Young's moduli, ν_1, ν_2 are Poisson's ratios.

Set the following boundary conditions:

$$\varepsilon_{ij}|_{x_j=0} = 0, \quad \varepsilon_{ij}|_{x_j=l_j} = 0, \quad (3)$$

and initial conditions accordingly:

$$\varepsilon_{ij}|_{t=0} = 0. \quad (4)$$

The relationship between stress components $\sigma^T = (\sigma_{11}, \sigma_{22}, \sigma_{12})$ and deformations $\varepsilon^T = (\varepsilon_{11}, \varepsilon_{22}, \varepsilon_{12})$ for two-dimensional fractional-differential rheological models, respectively, can be written as follows:

Voigt's model

$$\sigma_{11} = \frac{E_{11}}{(1-\nu_1\nu_2)} D_t^\alpha (\varepsilon_{11} - \varepsilon_{T1}) + \frac{\nu_2 E_{11}}{(1-\nu_1\nu_2)} D_t^\alpha (\varepsilon_{22} - \varepsilon_{T2}) + 2E\tau^\beta (D_t^\beta (\varepsilon_{11} - \varepsilon_{T1}) + D_t^\beta (\varepsilon_{22} - \varepsilon_{T2})), \quad (5)$$

$$\sigma_{22} = \frac{\nu_1 E_2}{(1-\nu_1\nu_2)} D_t^\alpha (\varepsilon_{11} - \varepsilon_{T1}) + \frac{E_{22}}{(1-\nu_1\nu_2)} D_t^\alpha (\varepsilon_{22} - \varepsilon_{T2}) + 2E\tau^\beta (D_t^\beta (\varepsilon_{11} - \varepsilon_{T1}) + D_t^\beta (\varepsilon_{22} - \varepsilon_{T2})), \quad (6)$$

$$\sigma_{12} = \mu D_t^\alpha (\varepsilon_{12} - \varepsilon_{T3}) + E\tau^\beta D_t^\beta (\varepsilon_{12} - \varepsilon_{T3}), \quad (7)$$

E is the modulus of elasticity of the elastic element of Voigt's body, $-0 \leq \alpha < \beta \leq 1$;

Kelvin' model

$$\sigma_{11} + \frac{E_1 \tau^\alpha}{E} D_t^\alpha \sigma_{11} = C_{11}(\varepsilon_{11} - \varepsilon_{T1}) + C_{12}(\varepsilon_{22} - \varepsilon_{T2}) + \frac{2E_1 E_2 \tau^\beta}{E} (D_t^\beta (\varepsilon_{11} - \varepsilon_{T1}) + D_t^\beta (\varepsilon_{22} - \varepsilon_{T2})), \quad (8)$$

$$\sigma_{22} + \frac{E_1 \tau^\alpha}{E} D_t^\alpha \sigma_{22} = C_{21}(\varepsilon_{11} - \varepsilon_{T1}) + C_{22}(\varepsilon_{22} - \varepsilon_{T2}) + \frac{2E_1 E_2 \tau^\beta}{E} (D_t^\beta (\varepsilon_{11} - \varepsilon_{T1}) + D_t^\beta (\varepsilon_{22} - \varepsilon_{T2})), \quad (9)$$

$$\begin{aligned} \sigma_{12} + \frac{E_1 \tau^\alpha}{E} D_t^\alpha \sigma_{12} &= 2C_{33} (\varepsilon_{12} - \varepsilon_{T3}) + \\ &+ \frac{E_1 E_2 \tau^\beta}{E} D_t^\beta (\varepsilon_{12} - \varepsilon_{T3}), \end{aligned} \quad (10)$$

In the case, when in the relations (5) - (7) we assume that $\alpha = 0, \beta = 1$, we obtain a classical two-dimensional Voigt's model in the case of orthotropy. The relations (8) - (10) will describe the classical Kelvin's model at fractal values $\alpha = 1, \beta = 1$.

For the integral representation of the relations (5) - (10), we take into account the properties of fractional derivatives [4], the definition of the fractional derivative ϑ , ($0 \leq \vartheta < 1$):

$$D_t^\vartheta f(t) = \frac{1}{\Gamma(1-\vartheta)} D_t \int_0^t (t-\xi)^{-\vartheta} f(\xi) d\xi, \quad (11)$$

as well as the method of the Laplace transformation [5].

Thus, the relations describing the relationship between stress and deformation (5) - (13) can be rewritten after the corresponding transformations in the integral form::

for Voigt's model

$$\begin{aligned} \sigma_{11} &= \frac{E_{11}}{(1-\nu_1\nu_2)\Gamma(1-\alpha)} (D_t \int_0^t (t-\xi)^{-\alpha} [(\varepsilon_{11}(\xi) - \varepsilon_{T1}(\xi)) + \\ &+ \nu_2 (\varepsilon_{22}(\xi) - \varepsilon_{T2}(\xi))] d\xi) + \\ &+ \frac{2E\tau^\beta}{\Gamma(1-\beta)} (D_t \int_0^t (t-\xi)^{-\beta} [(\varepsilon_{11}(\xi) - \varepsilon_{T1}(\xi)) + \\ &+ (\varepsilon_{22}(\xi) - \varepsilon_{T2}(\xi))] d\xi), \end{aligned} \quad (12)$$

$$\begin{aligned} \sigma_{22} &= \frac{E_{22}}{(1-\nu_1\nu_2)\Gamma(1-\alpha)} (D_t \int_0^t (t-\xi)^{-\alpha} [\nu_1 (\varepsilon_{11}(\xi) - \varepsilon_{T1}(\xi)) + \\ &+ (\varepsilon_{22}(\xi) - \varepsilon_{T2}(\xi))] d\xi) + \\ &+ \frac{2E\tau^\beta}{\Gamma(1-\beta)} (D_t \int_0^t (t-\xi)^{-\beta} [(\varepsilon_{11}(\xi) - \varepsilon_{T1}(\xi)) + \\ &+ (\varepsilon_{22}(\xi) - \varepsilon_{T2}(\xi))] d\xi), \end{aligned} \quad (13)$$

$$\sigma_{33} = \frac{\mu}{\Gamma(1-\alpha)} \left(D_t \int_0^t (t-\xi)^{-\alpha} (\varepsilon_{12}(\xi) - \varepsilon_{T3}(\xi)) d\xi \right) + \frac{E\tau^\beta}{\Gamma(1-\beta)} \left(D_t \int_0^t (t-\xi)^{-\beta} (\varepsilon_{12}(\xi) - \varepsilon_{T3}(\xi)) d\xi \right), \quad (14)$$

for Kelvin's model

$$\sigma_{11} = C_1 G(t) + A \int_0^t G(t-\xi) [C_{11} (\varepsilon_{11}(\xi) - \varepsilon_{T1}(\xi)) + \frac{2E_1 E_2 \tau^\beta}{(E_1 + E_2)} D_t^\beta (\varepsilon_{11}(\xi) - \varepsilon_{T1}(\xi))] d\xi + \quad (15)$$

$$+ A \int_0^t G(t-\xi) [C_{12} (\varepsilon_{22}(\xi) - \varepsilon_{T2}(\xi)) + \frac{2E_1 E_2 \tau^\beta}{(E_1 + E_2)} D_t^\beta (\varepsilon_{22}(\xi) - \varepsilon_{T2}(\xi))] d\xi,$$

$$\sigma_{22} = C_2 G(t) + A \int_0^t G(t-\xi) [C_{21} (\varepsilon_{11}(\xi) - \varepsilon_{T1}(\xi)) + \frac{2E_1 E_2 \tau^\beta}{(E_1 + E_2)} D_t^\beta (\varepsilon_{11}(\xi) - \varepsilon_{T1}(\xi))] d\xi + \quad (16)$$

$$+ A \int_0^t G(t-\xi) [C_{22} (\varepsilon_{22}(\xi) - \varepsilon_{T2}(\xi)) + \frac{2E_1 E_2 \tau^\beta}{(E_1 + E_2)} D_t^\beta (\varepsilon_{22}(\xi) - \varepsilon_{T2}(\xi))] d\xi,$$

$$\sigma_{12} = C_3 G(t) + A \int_0^t G(t-\xi) [2C_{33} (\varepsilon_{12}(\xi) - \varepsilon_{T3}(\xi)) + \frac{E_1 E_2 \tau^\beta}{(E_1 + E_2)} D_t^\beta (\varepsilon_{12}(\xi) - \varepsilon_{T3}(\xi))] d\xi, \quad (17)$$

where $A = \frac{(E_1 + E_2)}{E_1 \tau^\alpha}$, $G(t) = t^{\alpha-1} E_{\alpha, \alpha}(-At^\alpha)$,

$$C_1 = C_2 = \sigma^{(\alpha-1)}(0+) - 2E_2 \tau^{\beta-\alpha} (\varepsilon_{11}^{(\beta-1)}(0+) - \varepsilon_{T1}^{(\beta-1)}(0+)) + (\varepsilon_{22}^{(\beta-1)}(0+) - \varepsilon_{T2}^{(\beta-1)}(0+)),$$

$$C_3 = \sigma^{(\alpha-1)}(0+) - E_2 \tau^{\beta-\alpha} (\varepsilon_{12}^{(\beta-1)}(0+) - \varepsilon_{T3}^{(\beta-1)}(0+)),$$

$$f^{(\alpha-1)}(0+) = \lim_{x \rightarrow 0} \frac{1}{\Gamma(1-\alpha)} \int_0^x \frac{f(\xi)}{(x-\xi)^\alpha} d\xi.$$

III. NUMERICAL METHOD AND ALGORITHM OF MATHEMATICAL MODEL SOLUTION

To implement a numerical method, we introduce spatiotemporal mesh into the domain D:

$$\begin{aligned} \varpi_{\Delta t, h_1, h_2} &= \{(t^k, x_{1(n)}, x_{2(m)}) : x_{1(n)} = (n-1)h_1, \\ x_{2(m)} &= (m-1)h_2, t^k = k\Delta t, n = 1, \dots, N; h_1 = \frac{l_1}{N-1}; \\ m &= 1, \dots, M; h_2 = \frac{l_2}{M-1}; k = 0, 1, \dots, K; \Delta \tau = \frac{\tilde{T}}{K}\}. \end{aligned} \quad (18)$$

Given the Riemann-Liouville formula [6]:

$$\frac{\partial^\gamma f(x)}{\partial x^\gamma} \Big|_{x^n} = \frac{1}{\Gamma(1-\gamma)} \left(\frac{f(x^n)}{(x^{n+1} - x^n)^\gamma} + \int_{x^n}^{x^{n+1}} \frac{f'(\xi)}{(x^{n+1} - \xi)^\gamma} d\xi \right), \quad (19)$$

the difference approximation of the fractional derivative γ ($0 < \gamma \leq 1$) by coordinates x_1, x_2 can be written as follows [2]:

$$\frac{\partial^\gamma u}{\partial x_i^\gamma} \Big|_{x_{i(n)}} \approx \frac{u_{n+1} - \mathcal{U}_n}{\Gamma(2-\gamma)h_i^\gamma}, \quad (20)$$

where $h_i = x_{i(n+1)} - x_{i(n)}$, ($i=1,2$), $\Gamma(\cdot)$ is gamma function.

Taking into account the formula (20), the finite-difference approximation of the system of differential equations (1) - (2) will take the following form:

$$\begin{aligned} &\frac{C_{11}}{\Gamma(2-\gamma)h_1^\gamma} \bar{R}_{11} (\varepsilon_{11(n+1,m)}^k - \mathcal{U}_{11(n,m)}^k) - C_{11} \tilde{R}_{11} + \\ &+ \frac{C_{12}}{\Gamma(2-\gamma)h_1^\gamma} \bar{R}_{12} (\varepsilon_{22(n+1,m)}^k - \mathcal{U}_{22(n,m)}^k) - C_{12} \tilde{R}_{12} + \\ &+ \frac{2C_{33}}{\Gamma(2-\gamma)h_2^\gamma} \bar{R}_{33}^2 (\varepsilon_{12(n,m+1)}^k - \mathcal{U}_{12(n,m)}^k) - 2C_{33} \tilde{R}_{33}^2 = 0, \end{aligned} \quad (21)$$

$$\begin{aligned} &\frac{C_{21}}{\Gamma(2-\gamma)h_2^\gamma} \bar{R}_{21} (\varepsilon_{11(n,m+1)}^k - \mathcal{U}_{11(n,m)}^k) - C_{21} \tilde{R}_{21} + \\ &+ \frac{C_{22}}{\Gamma(2-\gamma)h_2^\gamma} \bar{R}_{22} (\varepsilon_{22(n,m+1)}^k - \mathcal{U}_{22(n,m)}^k) - C_{22} \tilde{R}_{22} + \\ &+ \frac{2C_{33}}{\Gamma(2-\gamma)h_1^\gamma} \bar{R}_{33}^1 (\varepsilon_{12(n+1,m)}^k - \mathcal{U}_{12(n,m)}^k) - 2C_{33} \tilde{R}_{33}^1 = 0. \end{aligned} \quad (22)$$

The boundary (3) and the initial (4) conditions in finite-difference form will be written:

$$\varepsilon_{11,22,12}^k \Big|_{(1,m)} = 0, \quad \varepsilon_{11,22,12}^k \Big|_{(N,m)} = 0, \quad (23)$$

$$\varepsilon_{11,22,12}^k \Big|_{(n,1)} = 0, \quad \varepsilon_{11,22,12}^k \Big|_{(n,M)} = 0, \quad (24)$$

$$\varepsilon_{11,22,12}^0 \Big|_{(n,m)} = 0. \quad (25)$$

In each spatio-temporal interval of the difference equations (21) and (22), it is necessary to find the vector of deformations $\varepsilon^T = (\varepsilon_{11}, \varepsilon_{22}, \varepsilon_{12})$. Taking into account the previous studies [2], we believe that the vector of deformations $\varepsilon_T = (\varepsilon_{T1}, \varepsilon_{T2}, \varepsilon_{T3})^T$, the components of which are due to temperature changes ΔT and moisture content ΔU ... are known at each time step

The numerical solution algorithm can be written as follows:

1. With the initial conditions (25) for n, m , where $(1 \leq n \leq N, 1 \leq m \leq M)$ we will find: $\varepsilon_{11n,m}^0, \varepsilon_{22n,m}^0, \varepsilon_{12n,m}^0$

2. For any time interval k , where $(0 < k \leq K)$, the values of the deformation components will be found as follows:

2.1. From the boundary conditions (23) - (24) we find $\varepsilon_{11(1,m)}^k, \varepsilon_{22(1,m)}^k, \varepsilon_{12(1,m)}^k, \varepsilon_{11(N,m)}^k, \varepsilon_{22(N,m)}^k, \varepsilon_{12(N,m)}^k$ for m , where $(1 \leq m \leq M)$;

$\varepsilon_{11(n,1)}^k, \varepsilon_{22(n,1)}^k, \varepsilon_{12(n,1)}^k, \varepsilon_{11(n,M)}^k, \varepsilon_{22(n,M)}^k, \varepsilon_{12(n,M)}^k$ for n , where $(1 \leq n \leq N)$.

2.2. Assuming that $n=1$, and $m \in (1, M)$, from relation (21) we obtain:

$$-C_{21}\varepsilon_{T1} + C_{21}\tilde{R}_{21} - C_{22}\varepsilon_{T2} + C_{22}\tilde{R}_{22} - 2C_{33}\varepsilon_{T3} + \frac{2C_{33}}{\Gamma(2-\gamma)h_1^\gamma} \left(1 - \bar{R}_{33}^1\right) \varepsilon_{12(2,m)}^k + 2C_{33}\tilde{R}_{33}^1 = 0. \quad (26)$$

As a result of elementary mathematical transformations we find $\varepsilon_{12(2,m)}^k$.

2.3. Similar to item 2.2., assuming that $n \in [2, N)$, and $m \in (1, M)$, we find $\varepsilon_{12(n,m)}^k$ from the relation:

$$\varepsilon_{12(n,m)}^k = \frac{\Gamma(2-\gamma)h_1^\gamma}{2C_{33}(1-\bar{R}_{33}^1)} (C_{21}\varepsilon_{T1} - C_{21}\tilde{R}_{21} + C_{22}\varepsilon_{T2} - C_{22}\tilde{R}_{22} + 2C_{33}\varepsilon_{T3} - 2C_{33}\tilde{R}_{33}^1). \quad (27)$$

2.4. Taking into account items 2.2. and 2.3., the values of the deformation component ε_{12} are found throughout the spatial-temporal interval. In future, taking into account the previous results, the deformation components ε_{11} and ε_{22} are searched for as follows. Assuming that in the relation (21) $n=1$, $m=2$ we obtain:

$$\frac{C_{11}(1-\bar{R}_{11})}{\Gamma(2-\gamma)h_1^\gamma} \varepsilon_{11(2,2)}^k - C_{11}\varepsilon_{T1} + C_{11}\tilde{R}_{11} + \frac{C_{12}(1-\bar{R}_{12})}{\Gamma(2-\gamma)h_1^\gamma} \varepsilon_{22(2,2)}^k - C_{12}\varepsilon_{T2} + C_{12}\tilde{R}_{12} - 2C_{33}\varepsilon_{T3} + 2C_{33}\tilde{R}_{33}^2 = 0. \quad (28)$$

In relation (22), we assume that $n=2$, $m=1$, from which expression (22) will take the form:

$$\frac{C_{21}(1-\bar{R}_{21})}{\Gamma(2-\gamma)h_2^\gamma} \varepsilon_{11(2,2)}^k - C_{21}\varepsilon_{T1} + C_{21}\tilde{R}_{21} + \frac{C_{22}(1-\bar{R}_{22})}{\Gamma(2-\gamma)h_2^\gamma} \varepsilon_{22(2,2)}^k - C_{22}\varepsilon_{T2} + C_{22}\tilde{R}_{22} - 2C_{33}\varepsilon_{T3} + 2C_{33}\tilde{R}_{33}^1 = 0. \quad (29)$$

As a result, we obtained two equations for two unknowns - $\varepsilon_{11(2,2)}^k, \varepsilon_{22(2,2)}^k$, whose results are found simply by means of elementary mathematical transformations.

2.5. Taking into account the preceding items and sub-items, let us assume that in the expression (21) $n=2$, $m=2$ and we obtain

$$\begin{aligned} & \frac{C_{11}}{\Gamma(2-\gamma)h_1^\gamma} (1-\bar{R}_{11}) (\varepsilon_{11(3,2)}^k - \gamma \varepsilon_{11(2,2)}^k) - C_{11}\varepsilon_{T1} + C_{11}\tilde{R}_{11} + \\ & + \frac{C_{12}}{\Gamma(2-\gamma)h_1^\gamma} (1-\bar{R}_{12}) (\varepsilon_{22(3,2)}^k - \gamma \varepsilon_{22(2,2)}^k) - C_{12}\varepsilon_{T2} + C_{12}\tilde{R}_{12} + \\ & + \frac{2C_{33}}{\Gamma(2-\gamma)h_2^\gamma} (1-\bar{R}_{33}^2) (\varepsilon_{12(2,3)}^k - \gamma \varepsilon_{12(2,2)}^k) - 2C_{33}\varepsilon_{T3} + 2C_{33}\tilde{R}_{33}^2 = 0. \end{aligned} \quad (30)$$

In relation (22) for $n=3$, $m=1$, we obtain:

$$\begin{aligned} & \frac{C_{21}}{\Gamma(2-\gamma)h_2^\gamma} (1-\bar{R}_{21}) (\varepsilon_{11(3,2)}^k - \gamma \varepsilon_{11(3,1)}^k) - C_{21}\varepsilon_{T1} + C_{21}\tilde{R}_{21} + \\ & + \frac{C_{22}}{\Gamma(2-\gamma)h_2^\gamma} (1-\bar{R}_{22}) (\varepsilon_{22(3,2)}^k - \gamma \varepsilon_{22(3,1)}^k) - C_{22}\varepsilon_{T2} + C_{22}\tilde{R}_{22} + \\ & + \frac{2C_{33}}{\Gamma(2-\gamma)h_1^\gamma} \left[(1-\bar{R}_{33}^1) (\varepsilon_{12(4,1)}^k - \gamma \varepsilon_{12(3,1)}^k) - (\varepsilon_{T3(4,1)}^k - \gamma \varepsilon_{T3(3,1)}^k) \right] + 2C_{33}\tilde{R}_{33}^1 = 0. \end{aligned} \quad (31)$$

As a result, we obtained two equations for two unknowns - $\varepsilon_{11(3,2)}^k, \varepsilon_{22(3,2)}^k$

2.6. Similar to sub-items 2.5 and 2.4, assuming that in expression (21) $n=3, \dots, N-2$, and $m=2$, while in relation (22) $n=4, \dots, N-1$, $m=1$, we obtain:

$$\begin{aligned} & \frac{C_{11}(1-\bar{R}_{11})}{\Gamma(2-\gamma)h_1^\gamma} (\varepsilon_{11(m+1,2)}^k - \gamma \varepsilon_{11(n,2)}^k) - C_{11}\varepsilon_{T1} + C_{11}\tilde{R}_{11} + \\ & + \frac{C_{12}(1-\bar{R}_{12})}{\Gamma(2-\gamma)h_1^\gamma} (\varepsilon_{22(m+1,2)}^k - \gamma \varepsilon_{22(n,2)}^k) - C_{12}\varepsilon_{T2} + C_{12}\tilde{R}_{12} + \\ & + \frac{2C_{33}}{\Gamma(2-\gamma)h_2^\gamma} (1-\bar{R}_{33}^2) (\varepsilon_{12(n,3)}^k - \gamma \varepsilon_{12(n,2)}^k) - 2C_{33}\varepsilon_{T3} + 2C_{33}\tilde{R}_{33}^2 = 0. \end{aligned} \quad (32)$$

$$\begin{aligned} & \frac{C_{21}(1-\bar{R}_{21})}{\Gamma(2-\gamma)h_2^\gamma} (\varepsilon_{11(n,2)}^k - \gamma \varepsilon_{11(n,1)}^k) - C_{21}\varepsilon_{T1} + C_{21}\tilde{R}_{21} + \\ & + \frac{C_{22}(1-\bar{R}_{22})}{\Gamma(2-\gamma)h_2^\gamma} (\varepsilon_{22(n,2)}^k - \gamma \varepsilon_{22(n,1)}^k) - C_{22}\varepsilon_{T2} + C_{22}\tilde{R}_{22} + \\ & + \frac{2C_{33}}{\Gamma(2-\gamma)h_1^\gamma} (1-\bar{R}_{33}^1) (\varepsilon_{12(m+1,1)}^k - \gamma \varepsilon_{12(n,1)}^k) - 2C_{33}\varepsilon_{T3} + 2C_{33}\tilde{R}_{33}^1 = 0. \end{aligned} \quad (33)$$

Taking into account the above written, we obtain the following solutions - $\varepsilon_{11(n,2)}^k, \varepsilon_{22(n,2)}^k$ for any n , where $n \in (1, N)$.

2.7. Arguing as in items 2.4. - 2.6., we will find all the values $\varepsilon_{11(n,m)}^k, \varepsilon_{22(n,m)}^k$ for m , where $m \in (3, M-1)$. To do this, m in relation (21) is such that $m \in (3, M-1)$, and in relation (22) m is such that $m \in (2, M-2)$.

Thus, the constructed algorithm for solving the numerical method is based on difference schemes and reduces the problem of finding the deformation components

$\varepsilon^T = (\varepsilon_{11}, \varepsilon_{22}, \varepsilon_{12})$ on the spatiotemporal mesh to elementary systems of two equations with two unknowns.

IV. ALGORITHM FOR IDENTIFYING FRACTIONAL DIFFERENTIAL PARAMETERS AND ANALYSIS OF NUMERICAL MODELING RESULTS

For one-dimensional fractal models of Voigt and Kelvin, the explicit form of the expression describing the deformation can be represented as follows:

$$\varepsilon_F(t) = \frac{\sigma_0}{E\tau^\beta \Gamma(\beta)} [2t^\beta - (t-t_1)^\beta h(t-t_1)] - \frac{\sigma_0}{E\tau^{2\beta-\alpha} (2\beta-\alpha)\Gamma(2\beta-\alpha)} [2t^{2\beta-\alpha} - (t-t_1)^{2\beta-\alpha} h(t-t_1)] \quad (34)$$

$$\varepsilon_K(t) = \frac{(E_1 + E_2)\sigma_0}{E_1 E_2 \tau^\beta \Gamma(\beta)} [2t^\beta - (t-t_1)^\beta h(t-t_1)] - \frac{(E_1 + E_2)\sigma_0}{E_1 E_2 \tau^{2\beta} 2\beta \Gamma(2\beta)} [2t^{2\beta} - (t-t_1)^{2\beta} h(t-t_1)] + \frac{\sigma_0}{E_2 \tau^{\beta-\alpha} \Gamma(1-\alpha)\Gamma(\beta)} [t^{\beta-\alpha} - (t-t_1)^{\beta-\alpha} h(t-t_1)] - \frac{\sigma_0}{E_2 \tau^{2\beta-\alpha} \Gamma(1-\alpha)\Gamma(2\beta)} [t^{2\beta-\alpha} - (t-t_1)^{2\beta-\alpha} h(t-t_1)] \quad (35)$$

where σ_0 is the value of the stress at the initial time, $h(t)$ is Heaviside's function ($h(t)=0$ at $t < 0$, $h(t)=1$ at $t \geq 0$), t_1 is unloading time.

Having a certain set of experimental and input data [7], the identification of fractional differential parameters of relations (34), (35) was carried out using the method of minimum squares:

$$\sum_{i=1}^n (\varepsilon_i - \varepsilon_{F,K}(t_i, \alpha, \beta))^2 \Rightarrow \min. \quad (36)$$

Clarification of the identification parameters is carried out using the method of coordinate descent [8].

A quantitative estimate of the difference in the results obtained for fractal creep equations (34), (35) is obtained by using a statistical criterion based on the correlation coefficient $\rho_{\varepsilon\bar{\varepsilon}}$:

$$\rho_{\varepsilon\bar{\varepsilon}} = \frac{\sum_{i=0}^n (\varepsilon_i - \varepsilon_{av})(\bar{\varepsilon}_i - \bar{\varepsilon}_{av})}{\sqrt{\sum_{i=0}^n (\varepsilon_i - \varepsilon_{av})^2 \sum_{i=0}^n (\bar{\varepsilon}_i - \bar{\varepsilon}_{av})^2}}, \quad (37)$$

where n is the number of points for which the comparison of variables is carried out, $\varepsilon_i, \bar{\varepsilon}_i$ are the deformation values for each model in accordance with the determined expressions and experimental data $\bar{\varepsilon}_{av} = \frac{1}{n+1} \sum_{i=0}^n \bar{\varepsilon}_i$.

The statistical criterion, accordingly, will take the form:

$$\Delta = \frac{|\rho_{\varepsilon\bar{\varepsilon}}|}{\sqrt{1 - \rho_{\varepsilon\bar{\varepsilon}}^2}} \sqrt{n-2}, \quad (38)$$

Let us present the results of identification of the fractional-differential Voigt's and Kelvin's models (Fig. 1,2) for moisture content $W = 15\%$ at temperature $T = 23^\circ C$, with modulus of elasticity according to wood samples $E = 16600 \text{ MPa}$ and $E = 13800 \text{ MPa}$. The values of the identified fractal parameters and the statistical criterion for Voigt's model are $\alpha = 0.22, \beta = 0.2394, \Delta = 2.2351$, and for Kelvin's model - $\alpha = 0.905, \beta = 0.91, \Delta = 7.38$

A numerical experiment to determine the dependence of stress σ and deformation ε upon time t for Voigt's, Maxwell's and Kelvin's models will be given for birch wood species, whose modulus of elasticity, E , is ($E = 10200 \text{ MPa}$). Fig. 3 shows the stress curves of the fractional-differential Voigt's model. The parameters α and β changed as follows: the fractional parameter α will be fixed closer to 1 (let α take on a value of 0.9), with the fractal parameter β being variable ($\beta = 0.3; 0.5; 0.7$), the fractional parameter β will be fixed closer to 0 (let β take on a value of 0.1), then α will change ($\alpha = 0.4; 0.6; 0.8$). For fractal Kelvin's and Maxwell's models, the parameters α, β (Fig.4) take on the following values: $\alpha = 0.4$ and $\beta = 0.5; 0.7; 0.9$.

The following conclusions can be drawn from the graphic dependencies: stress curves and deformations increase with increasing time; Kelvin's model takes the largest value of deformation in fractal parameters; among the deformation curves, the deformation takes the largest value with the fractal parameter $\beta = 0.5$, and for fractional differential Maxwell's model - with $\beta = 0.9$.

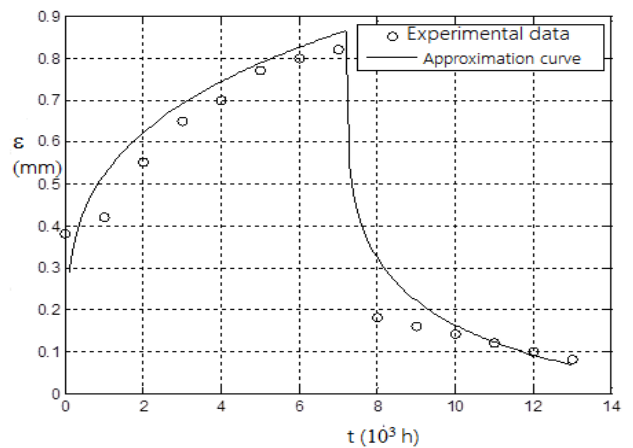


Fig. 1. Identification of the fractal parameters of Voigt's model

V. CONCLUSIONS

Considered are two-dimensional mathematical models of deformation and relaxation processes in fractal environment using the apparatus of fractional integro-differentiation. An algorithm for numerical solution to a two-dimensional mathematical model of viscoelastic deformation is obtained and the influence of fractional-differential parameters on deformation and stress for rheological fractal models is investigated. The algorithmic aspects and identification results for mathematical models with fractional-differential parameters are presented.

REFERENCES

- [1] J. Sokolovskyy, V. Shymanskyi, M. Levkovich, and V. Yarkun, "Mathematical and Software providing of research of deformation and relaxation processes in environments with fractal structure," XII international scientific and technical conference "Computer science and informational technologies" CSIT 2017, Lviv, Ukraine, pp. 24-2705-08 September, 2017.
- [2] Ya. I. Sokolovskyy, and M. V. Levkovich, "Numerical method for the study of nonisothermic moisture transfer in the environments with fractal structure," Bulletin of the National University "Lviv Polytechnic" Computer Science and Information Technologies, no. 843, pp. 288-296, 2015. (in Ukrainian)
- [3] L. Livi, A. Sadeghian, and A. Di Ieva, Fractal Geometry Meets Computational Intelligence: Future Perspectives. In: Di Ieva A. (eds) The Fractal Geometry of the Brain. Springer Series in Computational Neuroscience. Springer, New York, NY, 2016
- [4] V. V. Vasilyev, and L. A. "Simak Fractional calculus and approximation methods in the modeling of dynamic systems," Scientific publication Kiev, National Academy of Sciences of Ukraine, p.256, 2008
- [5] V. Uchajkin, Method of fractional derivatives. Ulyanovsk: Publishing house «Artishok», 2008.
- [6] I. Podlubny, Fractional Differential Equations, vol. 198. Mathematics in Science and Engineering, Academic Press, San Diego, Calif, USA, 1999.
- [7] Liu Tong, "Creep of wood under a large span of loads in constant and varying environments," Pt.1: Experimental observations and analysis, Holz als Roh- und Werkstoff, vol. 51, pp. 400-405, 1993.
- [8] E. Ogorodnikov, V. Radchenko, and L. Ugarova, "Mathematical modeling of hereditary deformational elastic body on the basis of structural models and of vehicle of fractional integro-differentiation Riman-Liuvil", Vest. Sam. Gos. Techn. Un-ty. Series. Phys.-math. sciences, vol. 20, no. 1, pp. 167-194, 2016.
- [9] O. Riznik, I. Yurchak, E. Vdovenko, and A. Korchagina, "Model of stegosystem images on basis of pseudonoise codes", VIth International Conference on Perspective Technologies and Methods in MEMS Design, Lviv, pp. 51-52, 2010.

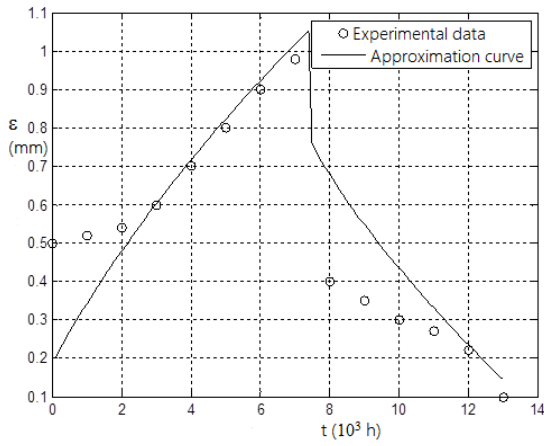


Fig. 2. Identification of the fractal parameters of Kelvin's model

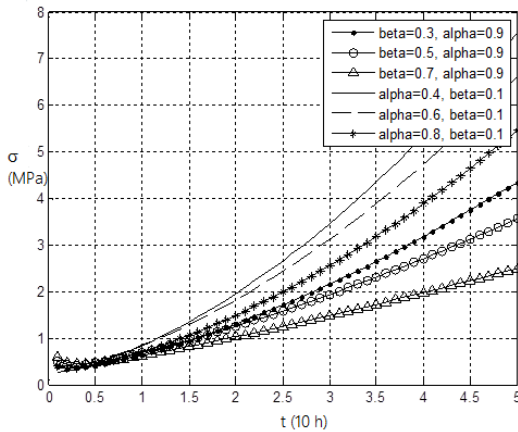


Fig. 3. Dependence of Voigt's model on fractional-differential parameters

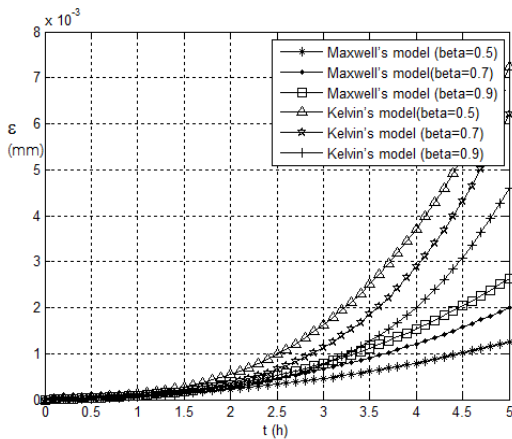


Fig. 4. Dependence of Maxwell's and Kelvin's models on fractional-differential parameters

Energy Efficient Clustering Protocol for Heterogeneous Wireless Sensor Network: A Hybrid Approach using GA and K-means

Shashi Bhushan
School of Computer and Information
Science, IGNOU
New Delhi, India
s.bhushan2k5@gmail.com

Raju Pal
Dept. of CSE
Jaypee Institute of Information
Technology, Noida, U.P., India
raju3131.pal@gmail.com

Svetlana G Antoshchuk
Institute of Computer System
ONPU
Odessa, Ukraine
asgonpu@gmail.com

Abstract—A hybrid approach combining genetic algorithm(GA) and K-means algorithm, called KGA is proposed in this paper for design of clustering protocol with energy efficiency for non-homogeneous wireless sensor network. The problem of optimal clustering can be considered as a problem for searching for an optimal number of clusters in a big search space such that WSN metrics are optimized. In the proposed protocol, distance between clusters, distance within clusters and a number of cluster heads are employed to search for optimal number of clusters and cluster heads. Maximization of energy saving and lifetime of a network are the two important metrics. The KGA is designed with a hybrid approach to population initialization scheme and objective function. The superiority of the protocol over other heuristic and meta-heuristic techniques is extensively demonstrated on several parameters: energy efficiency, network life time and throughput

Keywords— *Optimal clustering, WSN, Genetic Algorithm, K-means.*

I. INTRODUCTION

A WSN is a large network of sensor nodes characterized by sensing capability, limited wireless range for communication, limited processing power, limited storage and battery capacities. Sensor nodes are deployed in a large Sensor nodes sense physical properties of environment such as humidity, temperature, sound, etc., collect the sense data and forward it to the base station for further processing. They have potential application in large domains: environment monitoring, disaster warning system, health system and military application [1]. This has been enabled by the availability of cheaper and smarter sensors. One of the major challenge in WSN, however, is how to maximize the energy saving and lifetime of the network.

Clustering is one of the most efficient technique for maximizing energy saving and increase the lifetime of the network. K-means is one of the most popular data clustering techniques which partitions the data into K-clusters such that there is more similarity within clusters but less similarity between clusters [2]. Within each cluster there is a node called a cluster head which aggregates data received from sensor nodes within a cluster and forwards it to the base station. Clustering has been the most popular approaches for scalability of the network as well as designing energy efficient routing protocol for WSN. Clustering and routing with energy efficiency are the two widely used problems in WSN which belong to optimization class of problems. A

number of heuristic and meta-heuristic techniques have been proposed for the above problems with the objectives of extending the lifetime of a network and maximize energy saving. LEACH [3] and SEP [4] are the two main prominent heuristic techniques for clustering routing in WSN. All sensor nodes in LEACH are homogeneous whereas in SEP, some nodes maintain heterogeneity in terms of energy. LEACH utilizes randomized rotation of CHs for proper load distribution of energy loads among all the nodes in a cluster. However it depends upon the probability model. In SEP, CH selection probabilities of nodes are weighted by initial energy of each node compared to the others. In [5] a meta-heuristic technique using GA [6] is used to generate energy efficient hierarchical clusters. Its fitness function include several distance parameters that is, distances between sensor nodes and the base station, sensor nodes and a cluster head and cluster heads and a base station as well as standard deviation in cluster distance, transfer energy and number of transmissions. In [7], three cluster quality parameters have been added to derive fitness function in order to measure quality of energy efficient clustering which includes: intra-cluster distance, inter-cluster distance and a number of CHs. Higher CHs result in more energy consumption compared to few CHs. In [8], GA based protocol is proposed to extend the lifetime of a network and improve the stability period by finding optimal number of CHs and their locations based on minimizing the energy consumption of all the sensor nodes. PSO [10] is applied for both clustering and routing algorithms in [9]. For finding out a routing path, there is a tradeoff between transmission distance and a number of hop counts for selection of a routing path. For the clustering algorithms, all the CHs which are in use heavily as a next hop relay node in data forwarding are assigned a few nodes to balance the energy consumption. In [11] GA is used for both clustering and routing algorithms. Residual energy of the gateways and distance from sensor nodes to their CHs are used for clustering whereas for selecting the routing path, the residual energy of the gateways along with a tradeoff between transmission distance and a number of hops are used. Unlike the above protocols where population initialization is random, the proposed protocol follows the deterministic approach to population initialization. The population is seeded with K-means algorithm to have good quality genes in the beginning. GA is further used to form clustering.

The main contribution of the papers is the following:

- Population initialization and clustering through a mix of deterministic and random approaches.
- Performance comparison with high performance heuristic and meta-heuristic techniques

The rest of the paper is organized as follows: Section 2 describes GA, section 3 formulates a problem, section 4 gives a detailed description of the proposed algorithm, KGA, and Section 5 presents analysis of the result. Section 6 concludes the paper with future work.

II. PRELIMINARIES

A. An Overview of Genetic Algorithm (GA)

GA is an optimization technique based on the principles of evolution and natural genetics and randomized search. It performs search operation in very complex, large and multimodal landscape in order to provide a near optimal solution to a problem. Initially population representing individual solutions is generated randomly in a search space. Each solution is a chromosome. A fitness function is used to measure quality of solutions in the population. Through use of biologically inspired operations like crossover and mutation on the selected population, a new generation of a good solution is selected. This process (of selection, crossover and mutation) is continued for fixed number of iterations or till a termination condition is fulfilled. The following algorithm explains steps in writing GA.

Algorithm 1: Genetic Algorithm
 $t = 0$;first iteration
initialize $P(t)$;initial population
evaluate $P(t)$;evaluate population using fitness function
while (termination_ condition not fulfilled) do
 $t = t+1$
choose $P(t)$ from $P(t-1)$; based on the evaluation result
modify $P(t)$;using cross-over and mutation operators
evaluate $P(t)$

End while

B. Energy Model

The following equations are used to compute the transmission and reception energy consumption as proposed in [7]

$$E_{Trans}(k, dist) = E_{elec} \cdot k^{\epsilon_{fs}} \cdot k \cdot d^2, dist \leq d_0 \quad (1a)$$

$$E_{Trans}(k, dist) = E_{elec} \cdot k + \epsilon_{mp} \cdot k \cdot d^4, dist > d_0 \quad (1b)$$

$$E_{Rcv}(k, dist) = E_{elec} \cdot k \quad (2)$$

where $E_{Trans}(l, d)$ is k bit transmission energy consumption at a distance of $dist$, $E_{Rcv}(k, dist)$ is receiving energy consumption for k bits data. Several factors such as encoding and modulation techniques and filtering influence E_{elec} which is the consumed energy by circuits while it is transmitting and receiving data. ϵ_{fs} and ϵ_{mp} refer to energy consumption required by the amplifier for the free space and multipath fading respectively which depend upon (i) distance between a transmitter and a receiver and (ii) an acceptable bit error rate.

III. KGA

This section covers the details of proposed method namely KGA from the initials of K-means and GA. It is a new hybrid evolutionary algorithm combining GA and K-means to solve clustering problem of WSN. Most of the routing protocols based on GA choose random population but these may not be efficient for sparsely deployed sensor networks because it may lead to the selection of CHs with low density. This may affect the final clustering result. A solution to this limitation is selection of good quality CHs in the population. A combination of meta-heuristic approach for supporting high exploration and deterministic approach for selection of good quality CHs can provide better clustering solution for HWSNs. In the proposed method, the population of GA is seeded with K-means in order to get good clustering result.

A. Problem Statement

The problem being studied in this paper is how to achieve optimal number of clusters in a large heterogeneous sensor network in order to maximize its lifetime by minimizing the energy consumption. Optimal clustering is an NP hard problem [12]. Various meta-heuristic protocols are available in literature to solve such kind of problems. Genetic algorithm is the most popular meta-heuristic method. In this paper a novel clustering method based on GA is proposed which considers the clustering parameters: intra-cluster distance, inter-cluster distance and number of clusters.

The main objective of clustering WSNs is to minimize the overall energy consumption in the network. Energy consumption of a node is directly proportional to the communication distance between the nodes. Long distance transmission between sensor nodes and a base station always consumes more energy. Energy consumption required in transmitting data by a sensor node S_i to the corresponding CH _{i} is represented as:

$$E_{Trans}(s_i, CH_i) \propto d^\lambda. \quad (3)$$

Where d is the distance between sender and receiver and λ is a path loss component, whose value lies between 2 and 4, i.e. $2 \leq \lambda \leq 4$. If d is minimized, the total intra-cluster distance is also minimized. Then, E_{Trans} is also minimized, resulting in increasing the network lifetime.

Efficient clustering plays a key role to minimize the communication distance between the nodes. To improve the quality of clustering a novel fitness function is proposed which is described in section III-C. Moreover, to overcome the problem of long distance communication between CH to sink, multihop routing is also considered. Multihop communication is formulated as follows:

- WSN = $\{C_1, C_2 \dots C_N\}$; set of clusters
- $C_i = \{S_1, S_2 \dots S_j\}$; set of sensor nodes. j is $|C_i|$
- $dist(C_i, C_j)$; distance between two CHs
- $next-hop(C_i) = \{C_j | C_j \in dist(C_j, BS) < dist(C_i, BS)\}$

B. Chromosomes Structure and Population Initialization

The performance of GA is based on the efficient population initialization. In KGA, K-means is used to seed the population so that better CHs can be considered even in the case of sparsely deployed sensor network. The cluster-heads returns by K-means are represented as 1s in a chromosome and the rest of the sensor nodes are represented by 0s denoting normal sensor nodes. Dead nodes are represented as -1s. The detailed algorithm is given in Algorithm 2.

Algorithm 2 Population Initialization
For each chromosome i in Population P
For each node j in a chromosome i
If node(j).energy >0
 $k = \text{round}((\text{Alive nodes} * 0.2));$
Else if $k > 0$
 $\text{ids} = \text{apply K-means with sensor nodes and } k$
 $\text{chromosome}(i,j) = 1$ if $i \in \text{ids}$
 $\text{chromosome}(i,j) = 0$ if otherwise
End If
End for
End for

B. Fitness Function

In KGA, the fitness function is defined based on intra and inter cluster distances and the number of alive nodes in the network. The objective is to minimize it. The function is defined in Eq. (7).

$$\text{Minimize } f = \alpha(D_{\text{intra}}/D_{\text{inter}}) + \beta * |WSN|, \quad (4)$$

where, α and β are weighted coefficients such that $\alpha + \beta = 1$. D_{intra} and D_{inter} are the intra-cluster and inter-cluster distances given in Eq. (5) and (6) respectively and $|WSN|$ gives the number of cluster heads in the network given in Eq. (4).

$$D_{\text{intra}} = \sum_{i=1}^{CHs} \sum_{\forall S_j \in C_i} d(S_j, CH_i), \text{ where } j = 1, \dots, n. \quad (5)$$

$$D_{\text{inter}} = \sum_{\forall C_i, C_j, C_i \neq C_j} d(CH_i, CH_j). \quad (6)$$

C. Clustering Algorithm

GA based clustering algorithm in KGA has two main phases: (i) Setup phase which is a clustering formation phase and (ii) Steady state phase in which intra-cluster and inter-cluster communication takes place through multi-hop communication. In set up phase cluster head selection is guided by GA. The steady state is of our protocol is similar to LEACH. The details steps are defined in Algorithm 3.

IV. RESULT ANALYSIS

In this section the performance of the proposed protocol is analysed against other existing routing protocols as SEP, IHCR and ERP in terms of network lifetime, residual energy and throughput with 10% and 20% advanced nodes. The simulation is performed in MATLAB 2016a. The discussion on each considered performance metrics is given in following subsections.

Algorithm 3 KGA

Setup Phase

1. Initialize Parameters of WSN fields
2. Initialize Heterogeneous WSN and Energy Model

3. Creation of Random Sensor Network
4. Initialize Evolutionary Algorithm Parameters
5. Seed the population of GA with K-means
6. Perform GA based clustering
7. Select the best CHs based on fitness value
8. CHs broadcast a message to all the remaining sensor nodes
9. The nodes select their CHs based on distance
10. Formation of CHs is complete in the first round

Steady State Phase

Sensor nodes start sensing the environment and transmitting data to their CHs as per their TDMA schedule. After receiving data, CHs aggregate and send data to the BS in multi-hop manner.

A. Network Lifetime

A Network lifetime can be shown by capturing the number of alive nodes at each round till every node in the network dies. Figure 1 and Figure 2 depict the comparative network life of each considered algorithm over 10% and 20% of node heterogeneity respectively. In each of the scenario KGA outperforms the other considered protocols.

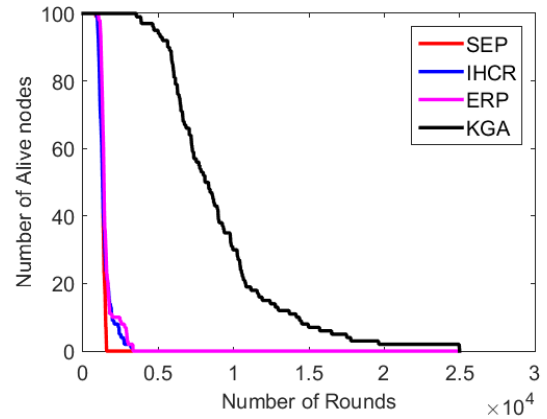


Fig. 1. Network Lifetime (10% advanced node)

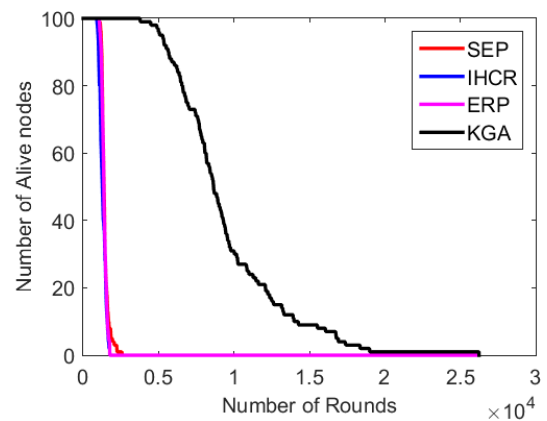


Fig. 2. Network Lifetime (20% advanced node)

To clearly depict the effectiveness of proposed algorithm, quantitative analysis has been performed and presented in Table 1 and 2 for 10% and 20% of advanced nodes (AV) in the network respectively. It can be observed from the table 1 that in KGA 10 nodes died in 5623 rounds while in other protocols, nodes died very early as 1153, 1047, and 1205

round for SEP, HCR, and ERP respectively. Similar in Table 2, nodes stay longer in KGA as compared to other protocols. Further, Table 3 and Table 4 have been presented to visualize the depletion of normal nodes(NRs) in comparison with advanced nodes(AVs) in the network. First 6000 round have been considered for comparative analysis and network statistics have been captured in different intervals of rounds like 10%, 20% and so on. Table 3 represents the statistics for the network with 10% advanced nodes. From the table, it is visible that almost all nodes died in HCR and ERP till 3000 rounds while in KGA no advanced node has died till 6000 round. Similarly, Table 4 represents the statistics for the network with 20% advanced nodes. From the table, it can be seen that in KGA, no advanced nodes have died in 6000 rounds.

TABLE I. NETWORK LIFETIME WITH 10% ADVANCED NODES

M=0.1				
% nodes died	SEP	HCR	ERP	KGA
10	1153	1047	1205	5623
20	1204	1099	1256	6130
30	1232	1159	1295	6602
40	1278	1255	1364	7263
50	1300	1303	1390	8097
60	1328	1372	1432	8972
70	1370	1459	1535	9943
80	1445	1940	1683	10788
90	1494	1956	2445	14170
100	1563	3220	3317	25002

TABLE II. NETWORK LIFETIME WITH 20% AVS

%dead nodes	SEP	HCR	ERP	KGA
10	1185	1050	1190	5648
20	1219	1146	1258	6540
30	1250	1208	1312	7605
40	1284	1276	1364	8071
50	1323	1353	1408	8648
60	1368	1430	1480	9288
70	1424	1569	1572	10034
80	1529	1928	1887	12085
90	1791	2529	2747	13971
100	2236	3536	3673	262389

TABLE III. ROUND HISTORY OF DEAD NODES WITH 10% AV AND NR NODES FOR A TOTAL OF 6000 ROUNDS

%rounds	HCR		ERP		KGA	
	AV	NR	AV	NR	AV	NR
10	0	0	0	0	0	0
20	0	35	0	9	0	0
30	0	86	0	89	0	0
40	4	90	0	90	0	0
50	8	90	7	90	0	0
60					0	1
70					0	3
80					0	5
90					0	8
100					0	16

TABLE IV. NUMBER OF DEAD NODES WITH 20% AV AND NR NODES FOR A TOTAL OF 6000 ROUNDS

%rounds	HCR		ERP		KGA	
	AV	NR	AV	NR	AV	NR
10	0	0	0	0	0	0
20	0	20	0	4	0	0

%rounds	HCR		ERP		KGA	
	AV	NR	AV	NR	AV	NR
30	0	80	0	78	0	0
40	2	85	0	80	0	0
50	6	85	5	80	0	0
60					0	0
70					0	1
80					0	2
90					0	8
100					0	13

B. Residual Energy

Figures 3 and 4 shows the comparison of KGA with the other protocols in terms of RE versus number of rounds with 10% and 20% advanced nodes respectively. There is a less steepness of the curve due to fairness in the energy load distribution and gradual dissipation of energy in the proposed protocol. The result is further validated through Tables 5-6.

C. Throughput

Figures 5 and 6 demonstrate the number of data packets sent to the base station by CH nodes per round with 10% and 20% node heterogeneity respectively. From these figures it can be observed that in KGA cluster heads send more packets to the base station compared to the other protocols.

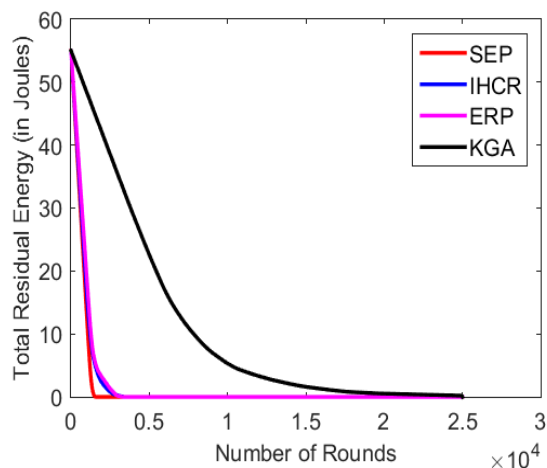


Fig. 3. Residual energy with 10% advanced nodes

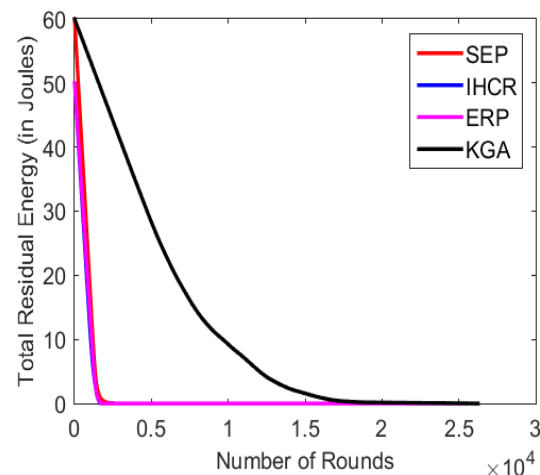


Fig. 4. Residual energy with 20% advanced nodes

V. CONCLUSION

Clustering and routing are two important aspects in WSN. GA has been extensively used to solve problems related to these aspects. The novel idea presented in the proposed work is related to hybridization of population initialization. In this paper a hybrid approach combining GA and K-means is proposed to do clustering of WSN. Initial population of GA is seeded with K-means to have good quality CHs. The fitness function is defined over the parameters like intra-cluster distance, inter-cluster distance and number of clusters. The experimental results show that the performance of the proposed protocol is better than IHCR, SEP and ERP in terms

of network life time, residual energy with 10% and 20% advanced nodes and throughput. In the future, the proposed work will be tested on multiple WSNs and a large population size.

REFERENCES

- [1] Ian F. Akyildiz, et al., "A survey on sensor networks," IEEE Communications magazine, vol. 40.8, pp. 102-114, 2002.
- [2] J. Han, and M. Kamber, Data Mining: Concepts and Techniques, 2 ed. Morgan Kaufman Publishers, 2006
- [3] W. B. Heinzelman, A. P. Chandrakasan, H. Balakrishnan, "An application-specific protocol architecture for wireless microsensor networks," IEEE Transactions on wireless communications, vol. 1(4) pp. 660-670, Oct. 2002
- [4] Georgios Smaragdakis, Ibrahim Matta, and Azer Bestavros, "SEP: A stable election protocol for clustered heterogeneous wireless sensor networks," Boston University Computer Science Department, May 31 2004.
- [5] A. W. Matin, and S. Hussain, "Intelligent hierarchical cluster-based routing," in: Proceedings of the international workshop on mobility and scalability in wireless sensor networks (MSWSN) in IEEE international conference on Distributed Computing in Sensor Networks (DCOSS), pp. 165-172, 2006.
- [6] Zbigniew Michalewicz, Genetic algorithms + data structures = evolution programs. Springer, 2009
- [7] Bara'a A. Attea, and Enan A. Khalil. "A new evolutionary based routing protocol for clustered heterogeneous wireless sensor networks," Applied Soft Computing vol. 12.7, pp. 1950-1957, 2012.
- [8] Mohammed Abo-Zahhad, et al. "A new energy-efficient adaptive clustering protocol based on genetic algorithm for improving the lifetime and the stable period of wireless sensor networks," International Journal of Energy, Information and Communications, vol. 5.3 pp. 47-72, 2014.
- [9] Pratyay Kula, and Prasanta K. Jana, "Energy efficient clustering and routing algorithms for wireless sensor networks: Particle swarm optimization approach," Engineering Applications of Artificial Intelligence, vol. 33 pp. 127-140, 2014.
- [10] James Kennedy, "Particle swarm optimization." Encyclopedia of machine learning. Springer US, 2011. 760-766
- [11] Suneet K. Gupta, and Prasanta K. Jana. "Energy efficient clustering and routing algorithms for wireless sensor networks: GA based approach," Wireless Personal Communications, vol. 83.3 pp. 2403-2423, 2015.
- [12] Stefano Basagni, et al., "A generalized clustering algorithm for peer-to-peer networks," in Workshop on Algorithmic Aspects of Communication. 1997.

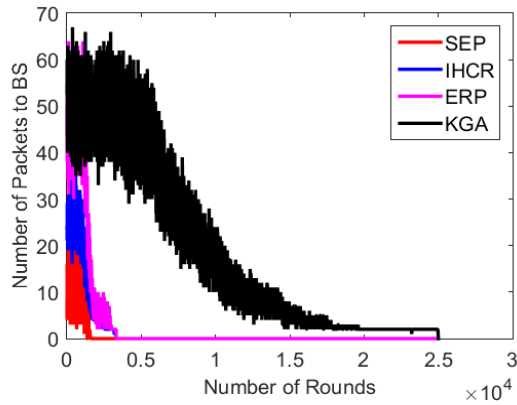


Fig. 5. Throughput (10% advanced nodes)

TABLE V. RESIDUAL ENERGY WITH 10% ADVANCED NODES

%rounds	SEP	HCR	ERP	KGA
10	27.66	26.89782	28.28127	51.05516
20	4.33	5.493893	6.849335	47.11319
30	—	2.014	3.06148	43.15388
40	—	0.088	0.986013	39.19862
50	—	0.002	0.229928	35.25073
60	—	—	—	31.31175
70	—	—	—	27.49672
80	—	—	—	23.78303
90	—	—	—	20.23968
100	—	—	—	16.91759

TABLE VI. RESIDUAL ENERGY WITH 20% ADVANCED NODES

%rounds	SEP	HCR	ERP	KGA
10	29.72	31.86016	33.49057	56.16828
20	5.11	10.43817	12.21157	52.34371
30	—	3.06148	3.82058	48.51257
40	—	0.986013	1.654297	44.68288
50	—	0.229928	0.158129	40.85083
60	—	—	—	37.00909
70	—	—	—	33.22163
80	—	—	—	29.49268
90	—	—	—	25.95671
100	—	—	—	22.78291

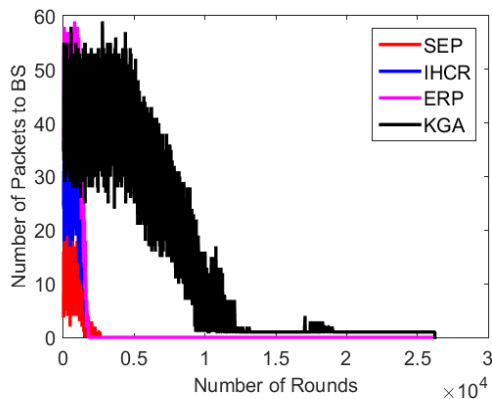


Fig. 6. Throughput (20% advanced nodes)

Hybridization of the SGTM Neural-like Structure through Inputs Polynomial Extension

Pavlo Vitynskyi

*Department of Publishing Information Technologies,
Lviv Polytechnic National University,
Lviv, Ukraine,
pavlo.vitynsky@gmail.com*

Ivan Izonin

*Department of Publishing Information Technologies,
Lviv Polytechnic National University,
Lviv, Ukraine,
ivanizonin@gmail.com*

Roman Tkachenko

*Department of Publishing Information Technologies,
Lviv Polytechnic National University,
Lviv, Ukraine,
roman.tkachenko@gmail.com*

Hakan Kutucu

*Department of Computer Engineering,
Karabuk University,
Karabuk, Turkey,
hakankutucu@karabuk.edu.tr*

Abstract—In this paper, a new approach for increasing the approximation accuracy with the use of computational intelligence tools is described. It is based on the compatible use of the neural-like structure of the Successive Geometric Transformations Model and the inputs polynomial extension. To implement such an extension, second degree Wiener polynomial is used. This combination improves the method accuracy for solving various tasks, such as classification and regression, including short-term and long-term prediction, dynamic pricing, as well as image recognition and image scaling, e-commerce. Due to the use of SGTM neural-like structure, the high speed of the system is maintained in both training and using modes. The simulation of the described approach is carried out on real data, the time results of the neural-like structure work and the accuracy results (MAPE, RMSE, R) are given. A comparison of the operation of the method with existing ones, such as Support vector regression, Classic linear SGTM neural-like structure, Linear regression (using Stochastic Gradient Descent), Random Forest, Multilayer Perceptron, AdaBoost are made. The advantages of the developed approach, in particular with regard to the highest accuracy among existing ones were experimentally established.

Keywords—approximation, Wiener polynomial, neural-like structures, Successive Geometric Transformation Model, input's extension

I. INTRODUCTION

To solve a number of data processing tasks, such as time series predicting [1, 2]; computer network traffic modeling [3]; development of the prevention subsystems for the smart home system [4]; development of information technology for environmental monitoring [5]; medical diagnostics [6]; image processing [7, 8, 9]; dynamic pricing; there is a critical need for an accurate and fast solution of the approximation task.

One of the promising classes of methods for solving this problem are methods based on machine learning. An effective solution to the problem of data approximation, in particular on the basis of the use of ANN, is based on their ability to approximate nonlinear functions. ANN as a universal approximator is capable of reproducing dependencies of any complexity [2].

Existing methods based on machine learning provide rather good results [10]. However, the use of this tools for solving the problem is characterized by the fact that its effectiveness largely depends on the computational capabilities of a particular type of ANN, which to some extent follows from their architecture [11]. In addition, the classic neural network tools, with iterative learning algorithms, is quite slow for solving such tasks. This imposes a number of limitations, mostly on time for ANN class that can be applied while working online. Another disadvantage of a number of methods is the dimension of a training set that will provide sufficient approximation quality. In some cases, it is not enough.

All this leads to the improvement of existing and the development of new methods for increasing the accuracy of approximation in the tasks of information processing, in particular, using efficient algorithms of machine learning.

In this work, we propose a new hybrid neural-like structure of the Successive Geometric Transformations Model (SGTM NLS) for solving this task. The feature of the proposed approach is the use of Wiener polynomial for the functional extension of its inputs. This ensures high accuracy of approximation with little time spent on training.

II. LINEAR NEURAL-LIKE STRUCTURE OF SUCCESSIVE GEOMETRIC TRANSFORMATIONS MODEL

The basis for modeling using the SGTM NLS is the basic principle of representing hypersurfaces of response in orthogonal coordinate systems (both straight-line and curvilinear) that coincide with the major dimensions of hypersurfaces.

The peculiarity of the linear-type SGTM neural-like structure is that the hypersurfaces of the response are hyperplanes [12]. In this case, the additional dimension of the model is completely determined by the noise components and round-off errors.

The training and using procedures of this computational intelligence tool are the same types [12]. Detailed training can be seen in [13]. The result of the application of the linear SGTM is that the basic measurements of the hyperplanes coincide with the results obtained using known PCA methods. However, a number of advantages characterize the

method based on SGTM, in particular, it is fast due to non-iterative, without errors accumulation and noticeable dimensionality limitations. It is no need to perform iterative adaptation or to solve the systems of normal equations [14].

In usage mode of the trained model, there is a fundamental possibility of analyzing the coordinates (components of the model) for the predictability and extraction of noise components. The topology of the SGTM NLS is shown on Fig. 1.

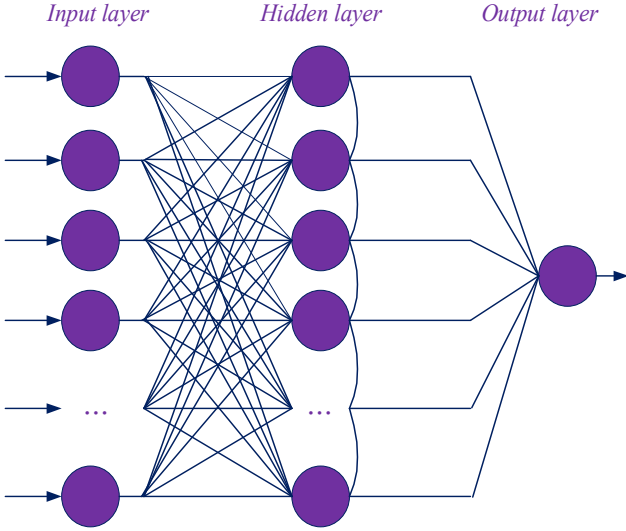


Fig. 1. The topology of the SGTM linear neural-like structure

The peculiarity of the SGTM neural-like structure is that their both variants (as software [12] as hardware [14]) effective implementation it is possible in particularly using parallel and distributed computing [15, 16].

III. WIENER POLYNOMIAL

One of the most effective methods for solving approximation tasks is based on the application of the Wiener polynomial [17]. According to the Weierstrass first theorem, this polynomial provides a simulation of continuous dependencies with arbitrarily high accuracy. Wiener's polynomial can be written as follows:

$$\begin{aligned}
 Y(x_1, \dots, x_n) = & a_i + \sum_{i=1}^n a_i x_i + \sum_{i=1}^n \sum_{j=1}^n a_{i,j} x_i x_j + \\
 & + \sum_{i=1}^n \sum_{j=1}^n \sum_{l=1}^n a_{i,j,l} x_i x_j x_l + \dots \\
 & \dots + \sum_{i=1}^n \sum_{j=1}^n \sum_{l=1}^n \dots \sum_{z=k-1}^n a_{i,j,l,\dots,z} x_i x_j x_l \dots x_z
 \end{aligned} \quad (1)$$

where n is number of variables and k is the degree of the Wiener polynomial.

A rather complicated problem, in this case, is the determination of the polynomial's coefficients using the various variants of the linear regression methods. It is because of the high computational complexity and the fact that this task belongs to the class of almost degenerate tasks.

The difference of this work is that we determine the polynomial's coefficients not using the least squares method,

but using the hybrid variant of the SGTM NLS in supervised mode.

IV. HYBRID NEURAL-LIKE STRUCTURE OF SUCCESSIVE GEOMETRIC TRANSFORMATION MODEL

The aim of the SGTM NLS hybridization is to provide accelerated formation of the inputs weights of neural-like structure, which will be the Wiener polynomial's coefficients and their application for solving various tasks.

The structure of the hybrid version of SGTM NLS contains two blocks (Fig. 2). The first one (Generator of the Wiener polynomial inputs) is intended to form the Wiener polynomial's members. This occurs according to (1) for a given the number of input variables (n) and the polynomial's degree (k).

Wiener polynomial's members $x_i, x_i x_j, x_i x_j x_l, \dots, \dots, x_i x_j x_l \dots x_z$ are formed sequentially for $i = \overline{1, n}, j = \overline{i, n}, l = \overline{j, n}, \dots, z = \overline{k-1, n}$ and $k = 1, 2, 3, \dots$

The next block of hybrid structure is a linear version of SGTM (Fig. 1). At its inputs, we serve the members of the Wiener polynomial, formed of the primary inputs (the values of the independent variables of a specific task) and the given polynomial degree.

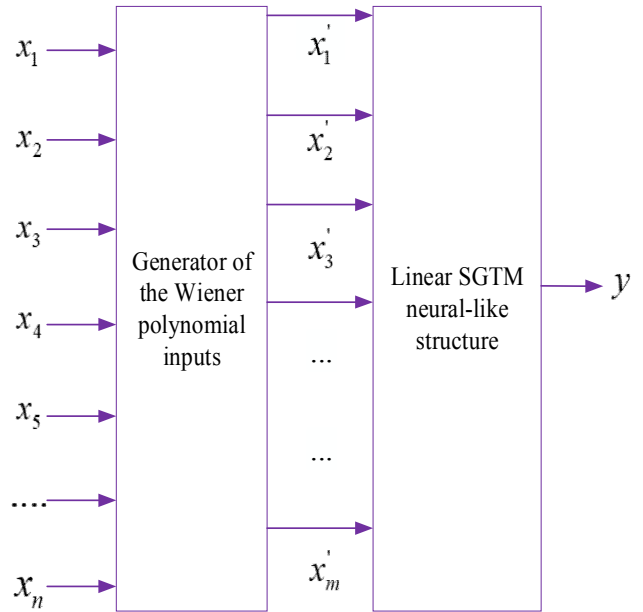


Fig. 2. Scheme of the hybrid SGTM neural-like structure

Such inputs polynomial extension of the linear version of SGTM NLS, allows increasing the accuracy of training and prediction procedures. In addition, this approach provides an opportunity to present a result in a compact form of the Wiener polynomial. This is possible given the repetition of the solution provided by the chosen learning tool. To do this, it is necessary to make a decomposition of the linear SGTM NLS [12] and synthesize the Wiener polynomial coefficients using the algorithm described in [13].

The next step in the procedure is to train SGTM NLS according to a known algorithm [12, 13], in which the neural-like structure implicitly forms the Wiener polynomial

coefficients. They will be used to solve various regression problems.

V. MODELLING

The method's simulation was carried out to solve the regression task. The task was to simulate (predict) the solar radiation of Libya. Data for the task was collected in 25 cities of Libya (Fig. 3) in the period 2010-2015 [18]. The training sample has 1900 vectors, each of which contains the seven inputs and one output attributes (Table I).

TABLE I. TRAINING DATA'S ATTRIBUTES

Variable	Attributes
x_1	Month
x_2	Elevation
x_3	Mean Temperature
x_4	Relative Humidity
x_5	Mean sunshine duration/h
x_6	Longitude
x_7	Latitude
y	Daily solar radiation

The entire sample was randomly divided into proportions of 80% and 20% [19] for the implementation of training and testing procedures respectively.

Basics parameters of the hybrid SGTM neural-like structure are m inputs, m neurons in the hidden layer, 1 output, where:

$$m = n + \frac{n(n+1)}{2} . \quad (2)$$

The Wiener polynomial degree that used for experiments was $k = 2$.

The quality assessment results of the developed hybrid structure were obtained using such indicators [20, 21]:

- tool absolute percentage error (MAPE) [22];

$$MAPE = \frac{1}{m} \sum_{i=1}^m \left| \frac{y_i - y'_i}{y_i} \right| * 100 . \quad (3)$$

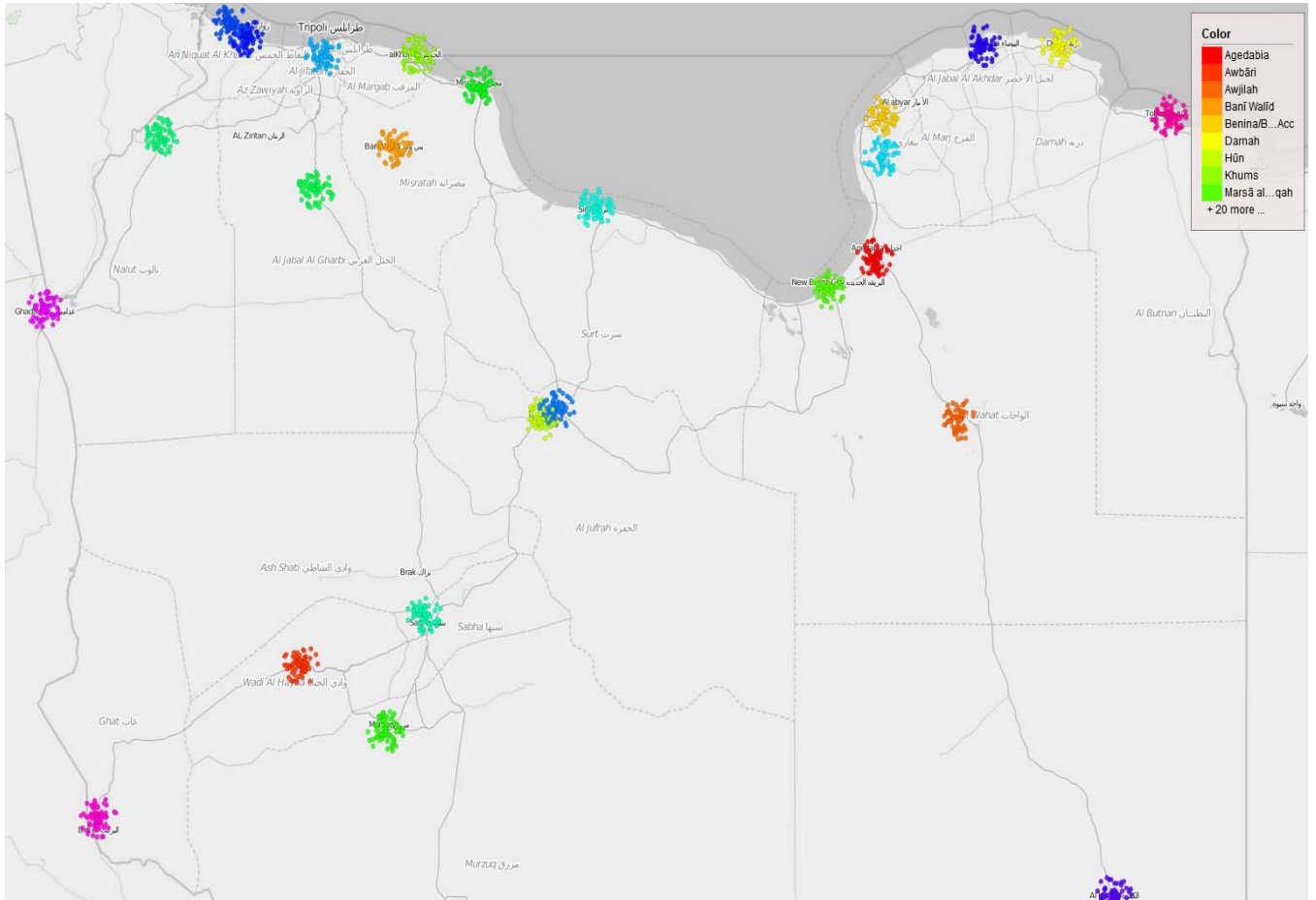


Fig. 3. Map of the Libya's cities.

- root tool squared error (RMSE) [23]:

$$RMSE = \sqrt{\frac{\sum_{i=1}^m (y_i - y'_i)^2}{m}} , \quad (4)$$

where y_i is a real daily solar radiation value and y'_i is the predicted value, $i = 1, m$.

- linear correlation coefficient (R) [24]:

$$R = \frac{m(\sum xy) - (\sum x)(\sum y)}{\sqrt{(m\sum x^2 - (\sum x)^2)(m\sum y^2 - (\sum y)^2)}}. \quad (5)$$

The software solution for implementing the described approach was developed using Python. The NumPy library, which contains linear algebra operations and provides high-performance, was used for work with data arrays.

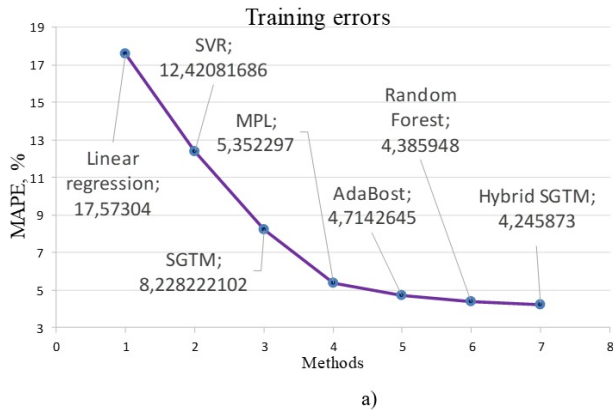
The results of the quality assessment of the described approach based on (3), (4), (5) are given in Table II. Results visualization of the described approach is shown in Fig. 6(g).

VI. COMPARISON OF THE SIMULATED RESULTS

Comparison of the results of the developed approach occurred with known methods. Their parameters that were used for simulation are given in Table III.

TABLE II. MODELLING RESULTS

N	Indicator	Result
1	MAPE, %	4,217
2	RMSE	0,298
3	R	0,98
4	Training time, seconds	0,1349350



Implementations of all compared methods were taken from the scikit-learn machine learning library in Python.

TABLE III. EXISTING METHODS

N	Method	Parameters
1	Support vector regression	kernel='rbf', gamma='auto', coef0=0.0, epsilon=0.001, max_iter=200
2	Classic linear SGTM neural-like structure	n inputs (primary data from Table I), n neurons in the hidden layer, 1 output
3	Linier Regression (with Stochastic Gradient Descent)	loss = 'squared_loss', alpha=0.0001
4	RandomForest	max_depth=5, random_state=0
5	Multilayer Perceptron	hidden_layer_sizes=(100, 40, 20), activation='relu', solver='adam', alpha=0.0001, batch_size='auto', max_iter=200
6	AdaBoost	max_depth=4, n_estimators=300

In Fig. 4, the results of comparison with other methods in training and testing modes are shown. As can be seen from Fig. 4, the best results for both modes are obtained using the developed method (based on MAPE). The same results are confirmed by other indicators.

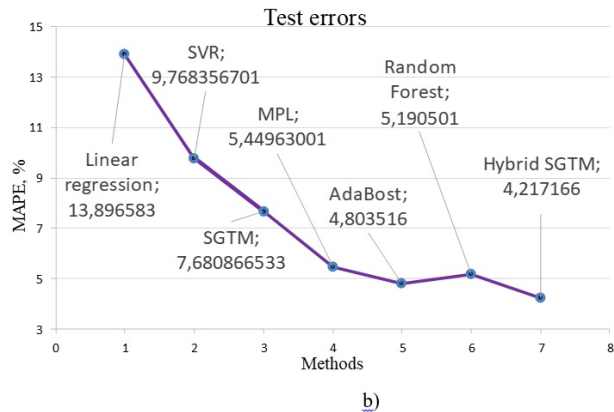


Fig. 4. Accuracy comparison of different methods through MAPE: a) training errors: b) test errors.

The quality evaluation of the result obtained while comparing with other methods was based beside (3) and (4) on the comparison of the training procedure duration [22].

In Fig. 5, the duration of the training procedure of the all methods in seconds are shows. As can be seen from Fig. 5, the smallest training time shown the common SGTM NLS.

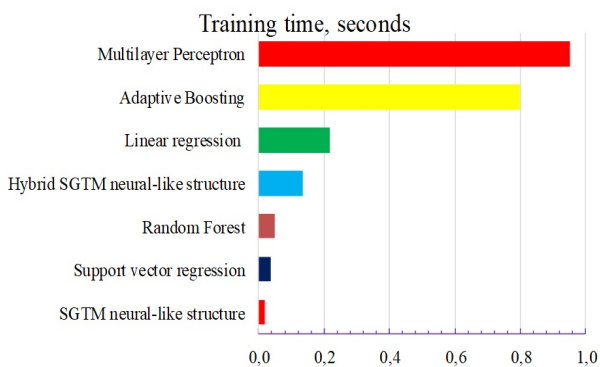


Fig. 5. Time comparison of different methods training procedures.

The proposed hybrid structure for providing the best result according to MAPE is 6 times slower than the classical

SGTM neural-like structure (without polynomial inputs extension), but 7 times faster than the MPL.

The results of various methods in the form of the scatter plots [25] are visualized in Fig. 6. On the x-axis, the daily solar radiation is shown, and the results of the corresponding regression method are presented on the y-axis. In Fig. 6(g), it is also confirmed that the most accurate results of solar energy prediction are obtained by the developed method, based on the Hybrid SGTM NLS.

VII. CONCLUSION

In this paper, a developed approach to increase the approximation accuracy based on the use of computational intelligence tools is described. The Wiener polynomial usage for inputs extension of the neural-like structure greatly increases the accuracy of such tool. The non-iterative training procedure that is provided by the chosen computational intelligence tool maintains a high-performance training process. The compatible usage of these two approaches to solve a number of the task allows us to combine the above-mentioned advantages.

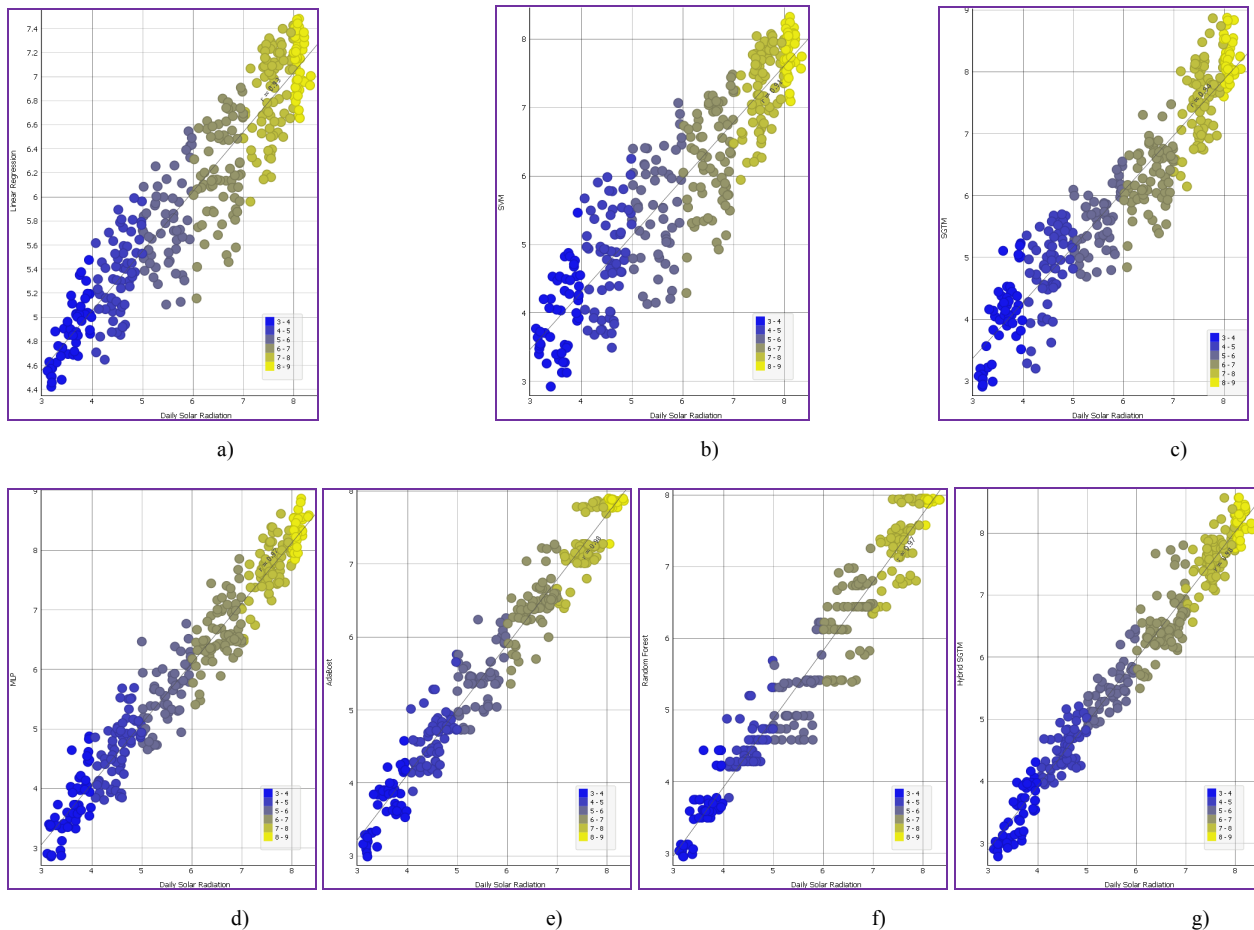


Fig. 6. Visual comparison of the various regression methods: a) Linear regression; b) Support Vector Regression; c) SGTM neural-like structure; d) Multilayer perceptron; e) AdaBoost; f) Random Forest; g) Hybrid SGTM neural-like structure.

The simulation of the proposed approach on real data shows the high accuracy of its work based on various indicators (MAPE, RMSE, R, Training time). Comparison of the results with modern methods (Support vector regression, Classic linear SGTM neural-like structure, Linear regression (with Stochastic Gradient Descent), Random Forest, Multilayer Perceptron, AdaBoost) shows its advantages, including the highest accuracy amongst all. At the same time, the time resources for implementing the training procedure in comparison with the classical methods are rather low.

The approach proposed in this paper can be used to solve a number of tasks that are critical to the time of performance (including training procedures) and performance indicators.

ACKNOWLEDGMENT

The authors' team are grateful to the reviewers for the relevant suggestions and comments that have significantly improved the quality of the article's presentation.

REFERENCES

[1] O. Mulesa, F. Geche, A. Batyuk, and V. Buchok, "Development of Combined Information Technology for Time Series Prediction," *Advances in Intelligent Systems and Computing II*. CSIT 2017. *Advances in Intelligent Systems and Computing*, vol 689, Springer, Cham, pp. 361-373, 2018.

[2] Z. Hu, Y. Bodyanskiy, O. Tyshchenko and O. Boiko, "An Evolving Cascade System Based on a Set of Neo-Fuzzy Nodes," *International*

Journal of Intelligent Systems and Applications (IJISA), vol.8, no.9, pp.1-7, 2016.

[3] I. Dronyuk and O. Fedevych, "Traffic Flows Ateb-Prediction Method with Fluctuation Modeling Using Dirac Functions," *Computer Networks*. CN 2017. *Communications in Computer and Information Science*, vol 718. Springer, Cham, pp. 3-13, 2017. doi.org/10.1007/978-3-319-59767-6_1

[4] V. Teslyuk, V. Beregovskiy, P. Denysyuk, T. Teslyuk and A. Lozynskiy, "Development and Implementation of the Technical Accident Prevention Subsystem for the Smart Home System," *International Journal of Intelligent Systems and Applications(IJISA)*, vol.10, no.1, pp.1-8, 2018. DOI: 10.5815/ijisa.2018.01.01.

[5] N. Shakhovska and O. Shamuratov, "The structure of information systems for environmental monitoring," *XIth Int. Scientific and Technical Conference Computer Sciences and Information Technologies (CSIT)*, Lviv, pp. 102-107, 2016. doi: 10.1109/STC-CSIT.2016.7589880

[6] I. Pliss and I. Perova "Diagnostic Neuro-Fuzzy System and Its Learning in Medical Data Mining Tasks in Conditions of Uncertainty about Numbers of Attributes and Diagnoses," *Automatic Control and Computer Sciences*, vol. 51(6), pp. 391-398, 2017. DOI: 10.3103/S0146411617060062

[7] M. Nazarkevych, R. Oliarnyk, H. Nazarkevych, O. Kramarenko and I. Onyshchenko "The method of encryption based on Ateb-functions," *In Data Stream Mining and Processing (DSMP)*, IEEE First International Conference, pp. 129-133, 2016.

[8] O. Riznik, I. Yurchak, E. Vdovenko and A. Korchagina, "Model of stegosystem images on the basis of pseudonoise codes," *Vlth International Conference on Perspective Technologies and Methods in MEMS Design*, Lviv, pp. 51-52, 2010.

[9] M. Nazarkevych, R. Oliarnyk, O. Troyan and H. Nazarkevych. "Data protection based on encryption using Ateb-functions," *In Scientific*

- and Technical Conference “Computer Sciences and Information Technologies (CSIT), pp. 30-32, 2016.
- [10] J. Wainer, Comparison of 14 different families of classification algorithms on 115 binary datasets. arXiv:1606.00930. June 2016.
- [11] Y. Bodyanskiy, O. Vynokurova, I. Pliss, G. Setlak and P. Mulesa, “Fast learning algorithm for deep evolving GMDH-SVM neural network in data stream mining tasks,” IEEE First International Conference on Data Stream Mining and Processing (DSMP), Lviv, pp. 257- 262. 2016. doi: 10.1109/DSMP.2016.7583555
- [12] R. Tkachenko, , I. Izonin, “Model and Principles for the Implementation of Neural-Like Structures based on Geometric Data Transformations,” Advances in Computer Science for Engineering and Education. ICCSEEA2018. Advances in Intelligent Systems and Computing. Springer, Cham, vol.754, pp.578-587, 2019. https://doi.org/10.1007/978-3-319-91008-6_58
- [13] R. Tkachenko, P. Tkachenko, I. Izonin and Y. Tsymbal, “Learning-Based Image Scaling Using Neural-Like Structure of Geometric Transformation Paradigm”, Advances in Soft Computing and Machine Learning in Image Processing. Studies in Computational Intelligence, Springer, Cham, N. 1, vol. 730, pp. 537-567, 2018. doi.org/10.1007/978-3-319-63754-9_25
- [14] U. Polishchuk, P. Tkachenko, R. Tkachenko and I. Yurchak, "Features of the auto-associative neurolike structures of the geometrical transformation machine," 5th Int. Conf. on Perspective Technologies and Methods in MEMS Design, Zakarpattya, Ukraine, pp. 66-67, 2009.
- [15] I. Tsmots, V. Teslyuk, T. Teslyuk and I. Ihnatyev, “Basic Components of Neuronetworks with Parallel Vertical Group Data Real-Time Processing”, Advances in Intelligent Systems and Computing II. CSIT 2017. Advances in Intelligent Systems and Computing, vol. 689, Springer, Cham, pp. 558 – 576, 2018.
- [16] O. Riznyk, I. Yurchak and O. Povshuk, “Synthesis of optimal recovery systems in distributed computing using ideal ring bundles”, XII Intern. Conf. on Perspective Technologies and Methods in MEMS Design (MEMSTECH), Lviv, pp. 220-222, 2016. doi:10.1109/MEMSTECH.2016.7507545
- [17] R. Tkachenko, Z. Duriagina, I. Lemishka, I. Izonin and A. Trostianchyn “Development of machine learning method of titanium alloys properties identification in additive technologies”, EasternEuropean Journal of Enterprise Technologies, Vol. 3, Iss. 12 (93), 2018, pp. 23-31. DOI: 10.15587/1729-4061.2018.134319
- <http://ijarcet.org/wp-content/uploads/IJARCET-VOL-5-ISSUE-4-919-923.pdf>
- [18] H. Kutucu and A. Almryad “Modeling of solar energy potential in Libya using an artificial neural network model” In: 2016 IEEE First International Conference on Data Stream Mining & Processing (DSMP), Lviv, 356-359, 2016.
- [19] Z. Hu, Y. Bodyanskiy and O. Tyshchenko, “A Deep Cascade Neural Network Based on Extended Neo-Fuzzy Neurons and its Adaptive Learning Algorithm,” IEEE First Ukraine Conference on Electrical and Computer Engineering (UKRCON), Kyiv, Ukraine, pp.801-805, May 29 – June 2, 2017.
- [20] Y. Bodyanskiy, O. Vynokurova, I. Pliss and D. Peleshko, “Hybrid Adaptive Systems of Computational Intelligence and Their On-line Learning for Green IT in Energy Management Tasks,” Green IT Engineering: Concepts, Models, Complex Systems Architectures. Studies in Systems, Decision and Control, vol 74. Springer, Cham, 2017
- [21] Y. Bodyanskiy, G. Setlak, D. Peleshko and O. Vynokurova, “Hybrid generalized additive neuro-fuzzy system and its adaptive learning algorithms,” IEEE 8th International Conference on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications (IDAACS), Warsaw, pp. 328-333, 2015. doi: 10.1109/IDAACS.2015.7340753
- [22] Ye. Bodyanskiy, O. Tyshchenko and D. Kopalani, “Adaptive learning of an evolving cascade neo-fuzzy system in data stream mining tasks,” Evolving Systems, 7(2), pp.107-116, 2016.
- [23] F. Geche, V. Kotsovsky, A. Batyuk, S. Geche and M. Vashkeba, "Synthesis of Time Series Forecasting Scheme Based on Forecasting Models System," 11th International Conference on ICT in Education, Research and Industrial Applications: Integration, Harmonization and Knowledge Transfer Lviv, Ukraine, May 14-16, CEUR Workshop Proceedings, vol. 1356, pp. 121-136, 2015
- [24] I. Perova and Ye. Bodyanskiy, “Fast medical diagnostics using autoassociative neuro-fuzzy memory,” International Journal of Computing, 16 (1), 34-40, 2017.
- [25] G. Setlak, Y. Bodyanskiy, O. Vynokurova and I. Pliss, "Deep evolving GMDH-SVM-neural network and its learning for Data Mining tasks," 2016 Federated Conference on Computer Science and Information Systems (FedCSIS), Gdansk, pp. 141-145, 2016.

Analysis of EEG using Multilayer Neural Network with Multi-Valued Neurons

Igor Aizenberg
Manhattan College
Riverdale
New York, USA
igor.aizenberg@manhattan.edu

Zain Khaliq
Manhattan College
Riverdale
New York, USA
zkhaliq01@manhattan.edu

Abstract—There is a wealth of analysis techniques that can be used in analyzing data of such a nature as EEG (Electroencephalogram), yet there are still many more ways and possibilities of analysis techniques to consider in order to produce a method that far exceeds the capabilities of the prevalent method. Since a multilayer neural network with multi-valued neurons (MLMVN) was successfully used earlier to decode EEG signals in a brain/computer interface (BCI) by analysis of their Fourier transform, it seemed very attractive to use it as a tool for EEG analysis. This work aims to further investigate how a complex-valued machine learning tool can be used to analyze EEG in the frequency domain. Our goal was to check how Fourier transform and complex wavelet transform of EEG can be analyzed using MLMVN in order to diagnose epilepsy, its remission or absence. We worked with a commonly used benchmark data set of epilepsy-related EEGs. The analysis of the transformed data was performed to determine a set of relevant statistical characteristics of DTCWT and Fourier transform components, which were then used as inputs of the MLMVN. The obtained results show a very high efficiency of the proposed approach.

Keywords—Complex-Valued Neural Networks, Multi-Valued Neuron, Multilayer Neural Network with Multi-Valued Neurons, MLMVN, EEG, Fourier transform

I. INTRODUCTION

We would like to use here MLMVN to analyze EEG in the frequency domain. MLMVN is a representative of complex-valued neural networks (CVNN) family. There is plenty of work done that states the use of CVNN, for example, a good observation is given in [1]-[3]. Traditionally CVNNs have been very successful in solving a number of real-world problems. We should mention such applications as detection of landmines [4], prediction of winds and their profiles [5], analysis of bio-medical images [6], prediction of oil production [7], frequency domain analysis of signals in EEG-based BCIs [8].

MLMVN is on the one hand a feedforward neural network, topologically identical to a multilayer perceptron (MLP). But on the other hand, MLMVN, being built from multi-valued neurons (MVNs) has its unique properties and important advantages over MLP. MLMVN was introduced in [9] as a 2-layer network. Then it was further developed [10] where MLMVN with an arbitrary number of hidden layers was introduced. Every particular property of MLMVN and its favorable distinctions over MLP are dictated by the utilization of the multi-valued neuron (MVN) as its essential unit. MVN was initially suggested in [11] as a k -valued threshold element and then re-introduced as a discrete MVN in [12].

MLMVN was effectively utilized in numerous applications. It was applied, for instance, for image deblurring through recognition of point-spread function and its specific parameters [13], long term time series prediction [7], analysis of signals in EEG-based BCIs [8], satellite information reversal for assurance of meteorological information profiles in the environment [15], solving various classification problems [3], [10], [16], and system identification [17]. MLMVN generalization capability in solving problems with discrete output, particularly classification and pattern recognition problems, was improved by a modified learning algorithm with soft margins [16]. To speed up a learning process and maintain simultaneously big learning sets and a high generalization capability, a batch learning algorithm was proposed in [17] and further developed in [18]. This algorithm as it is described in [18] was used in all experiments described in this paper.

EEG is used to collect the data about brain electrical activity. Then analysis of these data can be used to discover a certain dysfunction of some groups of neurons in the brain. Particularly, EEG is used to diagnose epilepsy in its different stages and examine patients with this diagnosis in remission. In the context of computing, computer science and computer engineering, EEG is used in building brain/computer interfaces, which help people with disabilities caused by some brain dysfunctions to perform certain tasks. A seminal work on electrical activity in the brain was published in 1875 by Caton [19]. His ideas were significantly developed 50 years later by Berger [20], [21]. It was succeeded to him to detect electrical activity in the brain using special electrodes placed on the head. Corresponding signals were acquired and recorded using a galvanometer connected to these special electrodes. It was noticed that electrical activity of the brain may change, for example, when eyes are open or closed. With these developments, the presence of EEG signals was scientifically proven. EEG signals are used in diagnostics, controlling of the anesthesia stage during surgical procedures, studies of sleep disorders, sleep psychology, and investigation of migraine. These signals are measured using a BCI. It consists of special electrodes which are used for measuring the electrical activity of the brain from the head surface.

Evaluation of EEGs is a specific job. It can typically be performed only by medical doctors whose area of specialization is EEG analysis. It is important to mention that EEG signals are not stable and they change continuously. They change their phases, frequencies and magnitudes. This makes interpretation of EEGs a challenging task. Medical doctors, depending on how different is their practical

experience, may interpret a same EEG differently. Hence it becomes quite important to use some intelligent computing tool, which should be able to analyze and interpret EEG signals. In [8] Manyakov et. al. suggested analyzing EEG by the analysis of its Fourier transform using MLMVN. This approach was pretty natural, since MLMVN works with complex-valued inputs and can be therefore used to analyze complex-valued information. This analysis was used in [8] to decode EEG signals in a brain/computer interface. In [22], [23] it was suggested to use another complex-valued neural network to analyze the dual-tree complex wavelet transform (DTWCT) [24] of EEG. This method was used to diagnose epilepsy. We would like to use here MLMVN as a sophisticated intelligent EEG signal classifier for epilepsy patients, in order to improve their treatment and distinguish EEGs of patients with epilepsy in remission and of those people who are healthy. This intelligent EEG signal classifier would also help treat those who are not yet fully have been the victims of epilepsy that is patients from a group of risk.

II. DATA GATHERING AND FEATURES EMPLOYED

A. EEG data set

To study EEG signals, we used a commonly employed benchmark dataset [25] containing data for five different classes. Class A consisting of all the EEG's of healthy patients, B, C, and D consists of EEG's of patients with remission, while Class E consist of EEG's of patents that are unhealthy in terms of epilepsy. Hence our dataset consists of five subsets. Each of these subsets contains EEGs of 100 volunteers recorded during a period of 23.6 seconds and sampled in 4097 samples. Subsets A and B contain a single channel EEGs recorded from healthy volunteers. Subset C contains EEGs recorded from hippocampal opposing hemisphere of sick patients before seizures [22]. Subset D consists of EEG recordings obtained from the epileptogenic region in the sick patients before seizures [22]. Subset E consists of EEGs containing seizures recorded from sick volunteers [22]. As it was mentioned above, the original data contain 4097 samples in each EEG. We did not use the last sample, thus we worked with EEG containing 4096 samples.

B. Features used for classification

To analyze a 4096-sampled EEG in the Fourier or DTCWT domain, it is necessary to have a relevant set of features using which it should be possible to perform this analysis. In fact, for example, a Fourier transform of such a 4096-sampled EEG contains 2048 frequencies, but EEGs from all five subsets A-E do not decisively differ from each other in medium and high frequencies. This means that the activity of a brain in those frequencies mostly is not different in different groups of sick and healthy people. However, this activity differs in EEGs of sick and healthy people in low frequency domain. Fig. 1 depicts magnitudes of the first 256 Fourier transform coefficients taken from representatives of each of A-E subsets.

Nevertheless, the analysis even of 256 spectral coefficients using MLMVN or any other machine learning tool should not be efficient. EEGs of different healthy and sick people are different from each other. For example, EEGs of sick people show some abnormal activities in some frequencies, but while shapes of these activities are similar, certain frequencies where these abnormal patterns can be found, are distinct, even if they are close to each other. The same properties are demonstrated by DTWCT transforms of

EEG. This means that it is necessary to extract some specific statistical features from the frequency domain data. We should look for some targeting features, which are similar for representatives from the same class of EEGs, but different for representatives from other classes. It was proposed in [22] where classification of EEGs based on DTCWT was studied to use the following 5 statistical characteristics as features for classification. These characteristics include (see Table I) mean (complex), engineered complex "minimum", engineered complex "maximum", engineered complex "standard deviation", and engineered complex "median". We call the last four complex-valued characteristics "engineered" because they are complex numbers artificially synthesized by finding corresponding real-valued characteristics separately over real and imaginary parts of the corresponding complex numbers followed by creation complex numbers from them by their pairing.

TABLE I. STATISTICAL CHARACTERISTICS USED

Characteristic	Mathematical expression
Minimum (complex)	$\min[\text{Re}(x_n)] + i \min[\text{Im}(X_n)]$
Maximum (complex)	$\max[\text{Re}(x_n)] + i \max[\text{Im}(X_n)]$
Expectation (mean) (complex)	$\frac{1}{N} \sum_{n=1}^N x_n$
Standard deviation (complex)	$\text{Re} \left[\sqrt{\frac{\sum_{n=1}^N (X_n - AM)^2}{N-1}} \right] + i \text{Im} \left[\sqrt{\frac{\sum_{n=1}^N (X_n - AM)^2}{N-1}} \right]$
Median (complex odd)	$\left(\frac{N+1}{2}\right)^{\text{th}}$ Re value + $i \left(\frac{N+1}{2}\right)^{\text{th}}$ Im value
Median (complex even)	$i((N/2) \uparrow \text{th} \text{Imvalue} + (N/2+1) \uparrow \text{th} \text{Imvalue})/2$

However, these characteristics except of mean are in fact artificial. They represent statistical characteristics of real and imaginary parts of complex numbers separately. Being merged into complex numbers they may not represent actual behavior of complex numbers or may represent it incompletely. A major issue here is that when we consider real and imaginary parts of complex numbers as separate abstract real numbers, we may completely lose circular nature of phase and very important information contained in phase. This issue becomes especially sensitive when we need to find characteristics relevant to some complex-valued random process or such a selection of complex numbers as Fourier transform or DTCWT. This problem is comprehensively studied in [26] where there is a very wide observation of first and second order moments suggested by different authors (particularly in [27], [28], [29]) relevant to complex-valued random processes and problems related complex-valued signal and data processing is presented. Let x and y are samples of the random variable. Then their *covariance* as the central moment is defined as follows [26]:

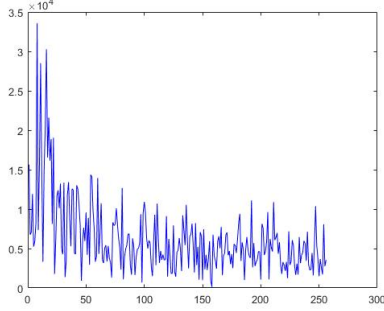
$$\text{cov}(x, y) = E \left[(x - E(x)) \overline{(y - E(y))} \right], \quad (1)$$

where E is an expectation and bar stands for complex conjugation.

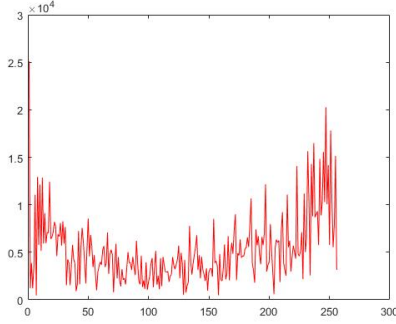
Based on (1) the *pseudo-variance* of x [26] should be defined as follows

$$\tilde{\sigma}_x^2 = \text{cov}(x, \bar{x}) = E \left[(x - E(x))(\bar{x} - E(\bar{y})) \right], \quad (2)$$

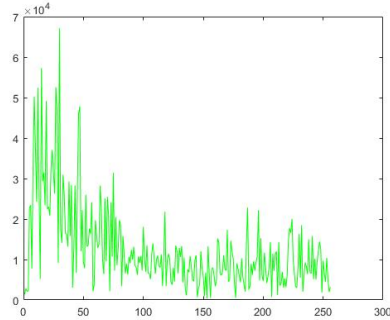
where bar stands for complex conjugation. Evidently $\tilde{\sigma}_x^2 \in \mathbb{C}$ that is it is in fact complex because $\tilde{\sigma}_x^2 = \text{cov}(x, \bar{x}) = \sigma_{x_{\text{re}}}^2 - \sigma_{x_{\text{im}}}^2 + 2i \text{cov}(x_{\text{re}}, x_{\text{im}}) \in \mathbb{C}$ (where i is an imaginary unit) whenever $x \in \mathbb{C}$.



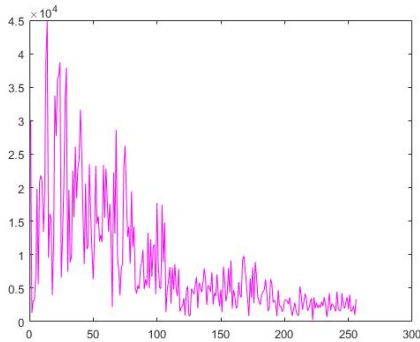
a) Magnitude of the Fourier transform of the representative from subset A (first 256 Fourier coefficients)



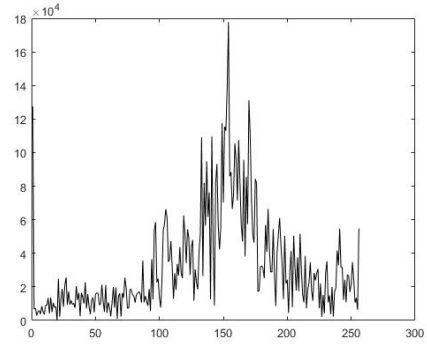
b) Magnitude of the Fourier transform of the representative from subset B (first 256 Fourier coefficients)



c) Magnitude of the Fourier transform of the representative from subset C (first 256 Fourier coefficients)



d) Magnitude of the Fourier transform of the representative from subset D (first 256 Fourier coefficients)



e) Magnitude of the Fourier transform of the representative from subset E (first 256 Fourier coefficients)

Fig. 1. Magnitudes of the Fourier transforms of EEGs

Then it follows from (2) that the *pseudo-variance* of N samples x_1, \dots, x_N of random variable X (the *pseudo mean square deviation*) should be defined as

$$\tilde{\sigma}_X^2 = \left(\sum_{k=1}^N \tilde{\sigma}_{x_k}^2 \right) / N. \quad (3)$$

The *complex correlation* coefficient is used as a measure for the degree of impropriety of x [26]. It is defined [29] as follows

$$\rho = \frac{\text{cov}(x, \bar{x})}{\text{cov}(x, x)}, \quad (4)$$

where bar again stands for complex conjugation. Evidently $\rho \in \mathbb{C}$ because $\text{cov}(x, \bar{x}) \in \mathbb{C}$ whenever $x \in \mathbb{C}$.

Let us now define the sample *complex correlation* coefficient ρ_X for N samples x_1, \dots, x_N of random variable X . Taking into account (4) we obtain the following

$$\rho_X = \frac{\sum_{k=1}^N \text{cov}(x_k, \bar{x}_k)}{\sum_{k=1}^N \text{cov}(x_k, x_k)}. \quad (5)$$

Empirically it looks that the complex pseudo-variance (3) and the complex correlation (5) should better represent closeness and distinctions among samples of complex variables X and Y than synthesized “median”, “min”, “max”, and “standard deviation” from Table I. It is very important that (3) and (5) are complex numbers and they in fact should not significantly differ from each other if X and Y are similar in probabilistic/statistical terms, but they should be quite different from each other when X and Y are probabilistically/statistically different from each other. This should also be true when X and Y represent Fourier or DTCWT spectra of EEGs taken from healthy people, patients with epilepsy and patients with epilepsy in remission. Hence we will use in our experiments the following three features to describe each EEG (its Fourier or DTCWT transform):

- 1) *Complex expectation (mean)* $E(X) = \left(\sum_{k=1}^N x_k \right) / N$
- 2) *Complex pseudo-variance (3)*

3) Complex correlation coefficient (5)

III. EXPERIMENTAL TESTING

Our goal was to perform the same classification experiments, which were performed in [22], but using Fourier transform along with DTCWN for EEG representation, a set of three features, which was just determined instead of those features listed in Table I, and MLMVN as a machine learning tool instead of another complex-valued neural network used in [22]. In all our experiments, we used 10-fold cross-validation (as it was done in [22]).

A. Experiment 1

The goal in this experiment is to classify only healthy people and patients with epilepsy (clusters A and E from the dataset) based on their DTCWT and Fourier transform. Thus, here we have to solve a 2-class classification problem. We used 10-fold cross-validation in this experiment.

B. Experiment 2

This experiment extends Experiment 1 involving the cluster D (sick patients before seizures). We again used both DTWCT and Fourier transform to represent EEGs. Hence in this experiment we are dealing with a 3-class classification

problem with the following three classes: 1) healthy class (all representatives from subset A were included); 2) patients with epilepsy having seizure-free intervals (all representatives from subset D were included); 3) patients with epilepsy having seizure (all representatives from subset E were included). We again used 10-fold cross-validation here.

C. Experiment 3

Experiment 3 was operated by involvement all five subsets from the dataset. In this experiment, we classify EEGs as belonging to the following three classes. The first class was formed from subsets A and B (all healthy people), the second one was formed from subsets C (sick patients before seizures) and D (patients with epilepsy having seizure-free intervals), and the third one was formed from the representatives of class E (patients with seizures). We used 10-fold cross-validation applied to the data extracted from the corresponding EEGs using both DTWCT and Fourier transform.

IV. RESULTS

Our experimental results are excellent. All of them are summarized below in Table II.

TABLE II. EXPERIMENTAL RESULTS

Experi-ment	Fourier Transform			Dual-tree complex wavelet transform (DTCWT)					
	MLMVN topology	# of learning iterations (mean over 10 experiments)	Classification accuracy (mean over 10 experiments)	Level 1			Level 2		
				MLMVN topology	# of learning iterations (mean over 10 experiments)	Classification accuracy (mean over 10 experiments)	MLMVN topology	# of learning iterations (mean over 10 experiments)	Classification accuracy (mean over 10 experiments)
1) A-E	3-2-1	71	100%	3-1	12	100%	3-3-1	6	100%
2) A-D-E	3-2-3	83	100%	2-3	423	100%	3-2-3	27	100%
3) AB-CD-E	3-4-3	84	100%	4-3	210	100%	3-4-3	26	100%

We used MLMVN-SM-LLS as it is presented in [18] with a very slight modification in the learning rule described in [30], which makes it possible to use arbitrary complex-valued input in MVN. We employed MLMVN with a single hidden layer and the output layer. A network topology is represented everywhere as n - N - M where n is the number of network inputs, N is the number of hidden neurons and M is the number of output neurons.

In Experiments 2 and 3 where we worked on solving 3-class classification problems, we used 3 output neurons performing binary classification. The output of the network was determined in such a case using the winner takes it all technique suggested with regard to MVN in [31] and employed for MLMVN-SM-LLS in [18]. The global threshold of 0.78 radian was used for soft margins in MLMVN. After publication of this paper, software and data used here will be available online¹. Hence, the use of the features, which we suggested in this paper, and MLMVN as a machine learning tool improves the results presented in [22]. We got a stable 100% classification rate in all our experiments for both DTWCT and Fourier transform using

only three features. Very small networks were enough to use to get the 100% classification accuracy. In fact, the largest network, which we employed in Experiment 3, contains only 7 neurons (4 hidden neurons and 3 output neurons). The learning process converges very quickly. Actually, this is basically a real time immediate convergence. It is important that while the authors of [22] were skeptical on the use of Fourier transform as a source of features for EEG analysis, we have shown here that it also can be used (while the learning process for DTWCT requires less iterations for its convergence).

V. DISCUSSION AND CONCLUSIONS

We suggested to use complex expectation (mean), complex pseudo-variance and complex correlation coefficient as three features to classify EEGs in the frequency domain. It was justified why these features characterize complex-valued data better than those applied separately to real and imaginary parts. We have also shown that along with DTWCT, which was earlier proven as a very good space for EEG representation, Fourier transform can also be used for the same purposes. It was also shown that MLMVN can successfully be used as an intelligent EEG classifier. The proposed approach should be further developed and tested using other appropriate EEG data. It should be attractive to use it not only for epilepsy diagnostics

¹ Software and data are available here
<https://www.freewebs.com/igora/Downloads.htm>

and epilepsy-related EEG analysis, but for classification of EEGs related to other health issues. This work also confirms high importance of the frequency domain representation of data, which are related to brain activity (since biological neurons exchange information with each other in terms of frequencies of spikes generated). It also shows how important it is to use a proper tool (a complex-valued neural network that is MLMVN in our particular case) for frequency domain data analysis.

REFERENCES

- [1] A. Hirose, *Complex-Valued Neural Networks*, 2nd Edn.. Springer, Berlin, Heidelberg, 2012.
- [2] D. Mandic and V. Su Lee Goh, *Complex Valued Nonlinear Adaptive Filters: Noncircularity, Widely Linear and Neural Models*. John Wiley & Sons, 2009.
- [3] I. Aizenberg, I., *Complex-Valued Neural Networks with Multi-Valued Neurons*. Berlin: Springer-Verlag Publishers, 2011.
- [4] Y.Nakano, and A.Hirose, "Improvement of Plastic Landmine Visualization Performance by use of ring-CSOM and Frequency-Domain Local Correlation," *IEICE Transactions on Electronics*, vol. E92-C, iss. 1, pp. 102-108, Jan. 2009.
- [5] S. L. Goh, M. Chen, D. H. Popovic, K. Aihara, D. Obradovic and D. P. Mandic, "Complex Valued Forecasting of Wind Profile," *Renewable Energy*, vol. 31, pp. 1733-1750, Sep. 2006.
- [6] A. Handayani, A.B.Suksmono, T.L.R.Mengko, and A.Hirose, "Blood Vessel Segmentation in Complex-Valued Magnetic Resonance Images with Snake Active Contour Model," *International Journal of E-Health and Medical Communications*, vol. 1, iss. 1, pp. 41-52, Jan. 2010.
- [7] I. Aizenberg, L. Sheremetov, L. Villa-Vargas, and J. Martinez-Muñoz, "Multilayer Neural Network with Multi-Valued Neurons in Time Series Forecasting of Oil Production," *Neurcomputing*, vol. 175, part B, pp. 980-989, Jan. 2016.
- [8] N.V.Manyakov, I. Aizenberg, N. Chumerin, and M. Van Hulle, "Phase-Coded Brain-Computer Interface Based on MLMVN", book chapter in *Complex-Valued Neural Networks: Advances and Applications* (A. Hirose – Ed.), Wiley, 2012, pp. 185-208.
- [9] I. Aizenberg, C. Moraga, and D. Paliy, "A Feedforward Neural Network based on Multi-Valued Neurons", In *Computational Intelligence, Theory and Applications*. *Advances in Soft Computing*, XIV, (B. Reusch - Ed.), Springer, Berlin, Heidelberg, New York, pp. 599-612, 2005.
- [10] I. Aizenberg, and C. Moraga, "Multilayer Feedforward Neural Network based on Multi-Valued Neurons (MLMVN) and a backpropagation learning algorithm," *Soft Computing*, 11, iss. 2, pp. 169-183, Jan. 2007.
- [11] N. N. Aizenberg, Yu. L. Ivaskiv, D. A. Pospelov, and G.F. Hudiakov, "Multivalued Threshold Functions. Synthesis of Multivalued Threshold Elements," *Cybernetics and Systems Analysis*, vol. 9, no. 1, pp. 61-77, January 1973.
- [12] N.N. Aizenberg and I.N. Aizenberg, "CNN Based on Multi-Valued Neuron as a Model of Associative Memory for Gray-Scale Images," *Proceedings of the Second IEEE International Workshop on Cellular Neural Networks and their Applications*, Munich, pp.36-41, October 14-16, 1992.
- [13] I. Aizenberg, D. Paliy, J. Zurada, and J. Astola, "Blur Identification by Multilayer Neural Network based on Multi-Valued Neurons", *IEEE Transactions on Neural Networks*, vol. 19, no. 5, pp. 883-898, May 2008.
- [14] O. Fink, E. Zio, and U. Weidmann, "Predicting Component Reliability and Level of Degradation with Complex-Valued Neural Networks," *Reliability Engineering & System Safety*, vol. 121, pp. 198–206, 2014.
- [15] I. Aizenberg, A. Luchetta and S. Manetti, "A modified learning algorithm for the multilayer neural network with multi-valued neurons based on the complex QR decomposition," *Soft Computing*, vol. 16, iss. 4, pp. 563-575, Apr. 2012.
- [16] I. Aizenberg, "MLMVN with Soft Margins Learning," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 25, no. 9, pp. 1632-1644, September 2014.
- [17] I. Aizenberg, A. Luchetta, S. Manetti., and C. Piccirilli., "System Identification using FRA and a modified MLMVN with Arbitrary Complex-Valued Inputs," *Proceedings of the 2016 IEEE International Joint Conference on Neural Networks (IJCNN 2016)*, Vancouver, pp. 4404-4411, July, 2016.
- [18] E. Aizenberg and I. Aizenberg, "Batch LLS-based Learning Algorithm for MLMVN with Soft Margins," *IEEE Symposium Series of Computational Intelligence (SSCI-2014)*, pp. 48-55, December, 2014.
- [19] R. Caton, "The electric currents of brain", *British Medical Journal*, vol. 2, pp. 278, 1875.
- [20] H. Berger, "Über das Elektroenkephalogramm des Menschen", *Arch Psychiatr Nervenkr*, vol. 87, pp. 527-570, 1929.
- [21] E. D. Adrian, and B. H. C. Matthews, "The Berger rhythm: potential changes from the occipital lobes in man," *Brain: A Journal of Neurology*, vol. 57, pp.355-385, 1934.
- [22] M. Peker, B. Sen, and D. Delen, "A Novel Method for Automated Diagnosis of Epilepsy Using Complex-Valued Classifiers," *IEEE Journal of Biomedical and Health Informatics* vol. 20, no. 1, pp. 108–118, January 2016.
- [23] B. Sen, M. Peker, F.V. Celebi and A. Cavusoglu, "A comparative study on classification of sleep stage based on EEG signals using feature selection and classification algorithms," *Journal of Medical Systems*, vol. 38, no. 18, 2014. doi: 10.1007/s10916-014-0018-0,
- [24] I. W.Selesnick, R. G.Baraniuk, and N. G.Kingsbury, "The Dual-Tree Complex Wavelet Transform", *IEEE Signal Processing Magazine*, Vol. 22, No 6, pp. 123–151, December 2005.
- [25] R. G. Andrzejak, G. Widman, K. Lehnertz, C. Rieke, P. David and C. E. Elger, "The epileptic process as nonlinear deterministic dynamics in a stochastic environment: an evaluation on mesial temporal lobe epilepsy," *Epilepsy Research*, vol. 44, pp. 129-140, 2001.
- [26] S. Trampitsch, "Complex-Valued Data Estimation", Master Thesis, Alpen-Adria-Universität Klagenfurt, Fakultät für Technische Wissenschaften, Austria, 2013, available online at http://www.uni-klu.ac.at/tewi/downloads/masterthesis_TrampitschStefan_Final_Versi_on.pdf
- [27] T. Adili, P. Schreier, and L.L. Scharf. "Complex-valued signal processing, The proper way to deal with impropriety," *IEEE Transactions on Signal Processing*, vol. 59, no. 11, pp. 5101-5125, November 2011.
- [28] S. M. Kay. *Statistical Signal Processing: Estimation Theory*, volume 1. Prentice Hall PTR, 2010.
- [29] P. J. Schreier and L. L. Scharf. "Second-order analysis of improper complex random vectors and processes," *IEEE Transactions on Signal Processing*, vol.51, no. 3, pp.s 714-725, March 2003.
- [30] I. Aizenberg, "Hebbian and Error-Correction Learning for Complex-Valued Neurons," *Soft Computing*, vol. 17, no. 2, pp. 265-273, Feb. 2013.
- [31] I. Aizenberg, N. Aizenberg, C. Butakov, and E. Farberov, "Image Recognition on the Neural Network based on Multi-Valued Neurons," *Proceedings of the 15th International Conference on Pattern Recognition*, Barcelona, Spain, IEEE Computer Society Press, vol. 2. pp. 993-996, September 3-8, 2000.

Business-Oriented Feature Selection for Hybrid Classification Model of Credit Scoring

Galyna Chornous
Department of Economic Cybernetics
Taras Shevchenko National University of Kyiv
Kyiv, Ukraine
chornous@univ.kiev.ua

Ihor Nikolskyi
Department of Economic Cybernetics
Taras Shevchenko National University of Kyiv
Kyiv, Ukraine
ihor.nikolskyi@gmail.com

Abstract — Application of predictive models on the basis of data mining confirmed its expediency in solving many economic problems. One of the crucial issues is the assessment of the borrower's creditworthiness on the basis of credit scoring models. This paper proposed an ensemble-based technique combining selected base classification models with business-specific feature selection add-on to increase the classification accuracy of real-life case of credit scoring. As the model limitations have been used easy-understandable algorithms on open-source software (R programming). The statistical results proved that hybrid approach for user-defined variables can be more than useful for ensemble binary classification model. It is shown that a great improvement can be reached by applying hybrid approach to feature selection process on additional variables (more descriptive ones that were built on initial features) for this real-life case with limited computational resources.

Keywords — hybrid, ensemble, feature selection, binary classification, stacking, major voting, credit scoring, R programming

I. INTRODUCTION

Application of predictive models on the basis of data mining (DM) confirmed its expediency in solving many economic problems: forecasting changes in stock indexes, price and production management, analysis of insurance risks, diagnostics of bankruptcy, credit card fraud, etc.

One of the crucial issues is the assessment of the borrower's creditworthiness on the basis of credit scoring models. The approach advantage is to avoid subjectivity in making credit decisions, to justify relevant decisions based on the knowledge extracted through the use of intelligent analysis of the accumulated data sets of borrowers.

A lot of statistical and machine learning techniques such as Logistic Regression, Support Vector Machines, Bayesian Networks, Neural Networks and Decision Tree algorithms are common tools for measuring creditworthiness [1; 2; 3].

The latest researches demonstrate that the implementation of separate stand-alone models for solving complex tasks does not always lead to success. Currently, the search for "strong" individual model is no longer relevant for most of researchers: the point of interest is about large ensembles of "weak" methods and algorithms [4; 5; 6].

The integration of various methods gives opportunity to deal with the limitations of each individual method, and in turn provides new opportunities for supporting the decision-making process within a single architecture and leads to the creation of hybrid intelligence systems (HIS). The

methodological support of HIS is based on a combination of different methods of AI, operations research, decision making theory and system analysis, the formation of ensemble models and the use of hybrid algorithms [7].

Forecasting the borrower's creditworthiness is one of the typical tasks of DM, it involves a binary classification of borrowers. To support the solution of the classification problem, the researchers offer a modern, complex implementation of methods and models such as advanced machine learning tools, such as ensembles of classifiers and hybrid classifiers, provide higher accuracy and performance. However, not all companies have proper human and computational resource for developing and implementation of state-of-the-art solutions. It limits the power of data science, i.e. a live issue is to find a way to improve performance of free common open-source solutions.

That's why the idea of this work is to find a way to dramatically improve model performance with limited computational resources, which is typical for Ukrainian business. The idea is to build a hybrid feature selection add-on for a typical ensemble model for binary classification.

II. RELATED WORKS

The assessment of the borrower's creditworthiness is one of the crucial issues for the financial institutions. The credit classification scoring is a great technique for understanding the risk of individual borrowers, gauging overall risk exposure and building data-driven, risk-adjusted strategies for existing customers. Therefore, great attention of researchers is paid to this issue in recent years.

We have studied a lot of publications regarding performance evaluation of DM algorithms on different tools, that were presented during the last 3 years. Some of them are described below.

L. Zernova [8] studied the relationship between the concepts of individual borrower risk, creditworthiness, the techniques of estimating creditworthiness, refined the concept and factors of creditworthiness, developed a methodology of scoring. N. Siddiqi [9] showed the most recent trends that widen credit scoring functionality and new in-depth analyses. It deals with defining infrastructure for in-house credit scoring, validation, governance, and Big Data. The authors [10] overview ideas of the statistical and operations research methods used in building credit and behavioral scorecards, as well as the advantages and disadvantages of each approach. These studies do not include specific developments in the hybrid approach, but present some examples of them.

H. Chen etc. [11] advanced Bayesian algorithm for credit assessment. The new trial ensembles logistic regression analysis (LRA), cluster and MLP-NN in Bayesian approach as an advanced classifier. S. Dahiya etc. [12] proposed an hybrid modeling technique using seven individual models (the NNs, C5.1, CART Tree, QUEST, CHAID, LR and SVM) to increase the model performance. Feature selection has also been used for selecting important attributes for classification. In this particular case Chi-Square statistic was adopted for choosing the most important ones out of all basic features. A.G. Armaki etc. [13] combined traditional and ensemble methods and come up with a hybrid meta-learner model. The structure of the model is based on the traditional hybrid model of ‘classification + clustering’. They propose several versions of the hybrid model by using various combinations of classification and clustering algorithms. S.H. Van etc. [14] built a creditworthiness classification model based on parallel Gradient Boosted Model, filter and wrapper approaches to estimate the credit score from the input features. Selected scoring variables are combined by feature importance (Gini index) and Information Value. H. Xiao etc. [15] proposed a hybrid classification method based on supervised clustering for credit scoring. Clusters from different classes are then pairwise combined to form a number of training subsets. In each training subset, a specific base classifier is constructed. M. Ala'raj and M.F. Abbod [16] presents a new hybrid ensemble credit scoring model through the combination of two data pre-processing methods based on Gabriel Neighbourhood Graph editing and Multivariate Adaptive Regression Splines in the hybrid modelling phase.

Despite the generous amount of studies and high quality results, the issue of which classifier is the best remains open, since it is very difficult to develop an all-in-one model that adapts to each set of data or set of attributes. Especially in the face of a shortage of human and computational resources for the implementation of the most up-to-date solutions.

III. METHODOLOGY

A. Dataset and tools

The main task of this work is the construction of an algorithm for binary classification in the context of determining the creditworthiness by personal data.

The data was collected by Dream Housing Finance Company in all home loans. They have presence across all urban, semi urban and rural areas. Customer first apply for home loan after that company validates the customer eligibility for loan. The given problem is to identify the customer segments, those are eligible for loan amount so that they can specifically target these customers. The data was collected for Analytics Vidhya loan prediction hackathon [17]. Dataset consists of 614 observations and the structure is presented in the Table I. The data was collected from online registration forms.

To perform all stages of the experiment one of the most popular DM tools was chosen – R programming. R is an efficient tool for evaluating great variety of DM techniques such as preprocessing, classification, clustering etc. It for was used for all stages of the experiment: data retrieving, data cleaning, feature selection, model building and validating the results. The base libraries are caret and caretEnsemble.

TABLE I. DATASET STRUCTURE

Variable	Type	Description
Loan_ID	Factor	Applicant ID
Gender	Factor	Gender
Married	Factor	Marital status
Dependents	Numeric	Number of dependents
Education	Factor	Education status
Self_Employed	Factor	Employment status
ApplicantIncome	Numeric	Income level
CoapplicantIncome	Numeric	Co-applicant income level
LoanAmount	Numeric	Asked loan
Loan_Amount_Term	Numeric	Payout period
Credit_History	Factor	Positive or negative history
Property_Area	Factor	Residence status
Loan_Status	Factor	Approved / disapproved

B. Concepts used

The concepts used in the experiment are feature selection, classification and ensemble modeling (ensemble stacking).

1) Feature selection

Feature selection or selection of information attributes is the process of selecting the most significant features for their further use in machine learning and statistics. Selection of attributes helps to solve the following tasks:

- Avoid overtraining the model by removing irrelevant characteristics.
- Avoid the "Curse of Dimension".
- Simplify the model and reduce the time for its training by removing redundant or insignificant features from the input data.

It is important to distinguish the feature selection from the dimensionality reduction. In the second case, the set of originally collected features is transformed into another set, which is more suitable for machine learning.

In this experiment Information Gain, Chi-Squared and Mean Decrease Gini methods.

- *Information Gain.* The entropy characterizes the impurity of an arbitrary collection of samples. Information Gain is the expected reduction in entropy caused by partitioning the samples according to a given feature which is the way of measuring association between inputs and outputs.
- *Chi-Squared test.* Pearson’s chi-squared test is a common statistical tool used for categorical data to understand the probability that any observed difference between the sets arose by chance and to test the independence of two events. it is used to find whether the occurrence of a certain feature and the occurrence of a specific class are independent. Thus, after estimating the following quantity for each feature they are ranked by the score. Higher scores states that the null hypothesis (H0) of autonomy is not confirmed which means that the occurrence of the term and class are dependent.
- *Mean Decrease Gini coefficient.* Mean decrease in the Gini impurity criterion is used as a way to estimate variable importance. The concept implies that every time a feature is used to split a node, the Gini coefficient for the child nodes are calculated to be

compared with the original node. The greater decreases, the stronger relationship with the output [18].

2) Feature engineering

The concept of feature engineering is to construct the process of using given knowledge of the data to select or build new features which are more suitable for machine learning algorithms. Feature engineering is a necessary stage of machine learning. It requires both analytical creativity and powerful computational resources. However, a good feature selection add-on can be automated through a combination of basic routine methods.

Feature engineering is more about human side of machine learning, but its usage is essential in applied machine learning. The basic process of feature engineering consists of the following steps:

- Consolidating fundamental ideas about feature set;
- Creating concepts for new features;
- Implementation of the ideas to a given dataset;
- Testing model sensitivity and performance with new features;
- Adjusting new features;
- Looping through this process until the best combination of features is found.

3) Binary Classification

Today binary classification is a widely used method in such spheres as medicine, telecommunication, economics, trading, sociology etc.

Binary classification is a common task in machine learning universe. This is an example of supervised learning technique, a method of machine learning where the classes are known beforehand and is used to classify new observations into predefined categories. Speaking about binary classification, there are only two classes.

The task of binary (binomial) classification is classifying the given elements into two groups (predicting which group each one belongs to) using a classification engine. A set of parameters (numeric, factor etc.) define the context of each observation which is an input for classification rule.

Accuracy metrics in term of binary classification engines calculate the two classes of correct predictions and two classes of errors. Typical metrics are accuracy (ACC), precision, recall, false positive rate, F1-measure. Every metric evaluates a certain aspect of the forecasting model. Accuracy (ACC) estimates the share of correct predictions. Precision estimates the share of true positives predictions among all observations that are predicted as positive. Recall estimates how many true positives were predicted as positive. F-measure is the harmonic mean of precision and recall.

Typical methods for performing binary classification are Decision Trees, Random Forest, Support Vector Machines, Neural Networks, Logistic Regression [5].

4) Stacking ensemble

Simultaneous global optimization of all basic algorithms used to construct the ensembles is a complex multicriteria problem. Such an optimization requires knowledge of the

internal structure of the algorithms, which complicates the use of standard teaching methods. In practice, the implementation of such a strategy is embodied in certain improvements of consistent and parallel approaches.

One of the corresponding implementations is level aggregation or stacking. Unlike bagging and booting, stacking is usually not used to combine models of the same type but applies to models built using various learning algorithms. Staking tries to learn each model using a meta-learning algorithm, which allows to find the best combination of outputs of the basic algorithms [19] (Fig. 1).

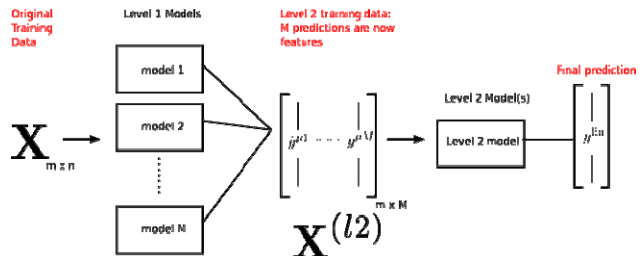


Fig. 1. Basic concept of stacking ensemble

Work with the ensemble includes the choice of basic neural network architecture and the number of networks in the ensemble, training (with the help of, for example, bagging or other algorithm), work with the received model (data processing, copying / serialization of the model, etc.).

The ensembles are widely used in practice, especially in predicting bankruptcy, determining credit scores etc. Examples of the implementation of ensembles are given in the studies.

IV. EXPERIMENTAL RESULTS

A. Proposed experiment

As stated above the key of the experiment is to improve performance of a multicriterial ensembled model for solving the task of binary classification on the case of determining the creditworthiness by personal data. Model concept is presented in Fig. 2.

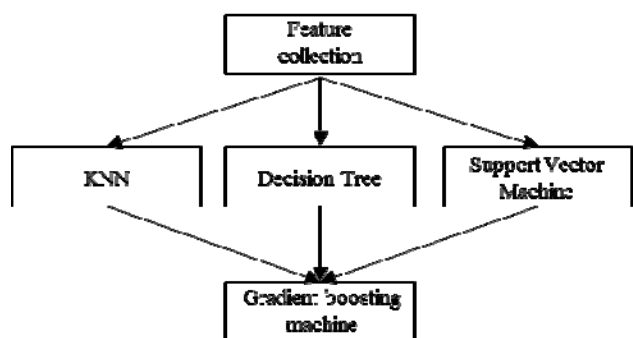


Fig. 2. Proposed model

The different steps that were carried out for conducting the experiment have been shown in this section. The high-level description of framework used to build the hybrid model is given in the Fig. 3.

To make the paper be outstanding form the variety of models for binary classification the following steps have been made:

- Applying machine learning to real-life case with limitations of Ukrainian business reality.
- Including heuristic analysis for data cleaning process to consider the origin of data
- Adding new features, which are more business informative than initial ones.
- Combination of several different feature selection methods, which gives the ability to correctly process both factor and numeric variables.
- Improving performance of common algorithms by combination of technical and analytical approaches.

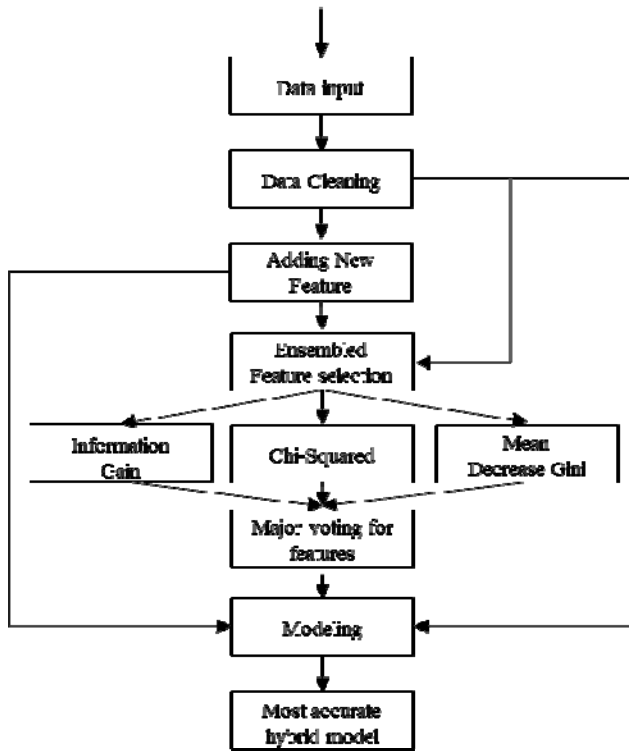


Fig. 3. Proposed architecture for the experiment.

According to the experiment plan the first tasks were about retrieving and cleaning data. As the idea of experiment is also aimed at thorough data cleaning, the various data preprocessing techniques were performed to remove noisy and redundant data from the database. The procedure of cleaning included removing observations with empty features and extreme values, filling missed values with median (for numeric features) and most popular (factor features) values. Exceptional approach was applied to features “Married” and “Self_Employed” (“No” if missed) due the nature of these fields.

Feature selection is also a data preprocessing technique to select the relevant attributes for the experiment. Feature engineering is crucial for model optimization.

The modeling stage implies looping through different combinations of features used for binary classification to find the most accurate approach.

B. Statistics results

This section presents the statistical results of all stages of the experiment. The underlying idea is to adjust business

logic to the process of feature selection by adding new feature interaction rules, which are presented in the Table II.

The next step was to apply hybrid approach for feature selection by major voting of mean decrease Gini, information gain and Chi-squared coefficient methods. The statistical results are presented in the Table III. In case if modified features outperform basic ones, basic variables are excluded from the collection of important features.

TABLE II. FEATURE INTERACTION RULES

New features	Interaction
IncomePerPerson	=ApplicantIncome / (Dependents+1)
MonthlyPayment	=Loan_Amount / Loan_Amount_Term
Friend	=if(CoapplicantIncome>0; True; False)
Family_backup	=if(AND(CoapplicantIncome>0;Married="Yes"); True;False)

The results in the Table IV show that modified features almost all tend to describe the nature of data better and be more informative in terms of data. So, the target collection of important features consists of Credit_History, Friend, Family_backup, MonthlyPayment, IncomePerPerson.

The next table describes comparison of basic methods for classification with three feature collections: basic features, all features and important features.

As we see, the main hypothesis is proven: ensemble model can benefit from hybrid approach for feature selection, especially with analytically defined variables, which results in slightly higher accuracy of a model.

TABLE III. FEATURE INTERACTION RULES

Feature	Importance	Information gain	Chi.Squared coefficient
Credit_History	0,039	0,099	0,457
Friend	0,037	0,099	0,457
Family_backup	0,022	0,062	0,346
ApplicantIncome	0,019	0,006	0,113
MonthlyPayment	0,018	0,005	0,105
LoanAmount	0,018	0,003	0,076
IncomePerPerson	0,015	0,001	0,037
CoapplicantIncome	0,012	0	0,029
Married, Dependents, Gender, Self_Employed, Loan_Amount_Term, Education, Property_Area	< 0,01	0	0

TABLE IV. EVALUATION OF MODEL PERFORMANCE

	Accuracy	F-measure
All initial features	0,692	0,679
All features	0,674	0,671
Ensembled approach		
Important initial features	0,776	0,753
Important modified features	0,841	0,832

V. CONCLUSION

This paper proposes an ensemble-based technique combining selected base classification models with business-

specific feature selection add-on to increase the classification accuracy of real-life case of credit scoring.

The model limitations were to use commonplace easy-understandable algorithms on open-source software (R programming).

The key points of this study are applying machine learning to real-life case with limitations of Ukrainian business reality; including heuristic analysis for data cleaning process to consider the origin of data; adding new features, which are more business informative than initial ones; combination of several different feature selection methods, which gives the ability to correctly process both factor and numeric variables; improving performance of common algorithms by combination of technical and analytical approaches.

The statistical results proved that hybrid approach for user-defined variables can be more than useful for ensemble binary classification model. That means that a great improvement can be reached by applying hybrid approach to feature selection process on additional variables (more descriptive ones that were built on initial features) for this real-life case with limited computational resources.

REFERENCES

- [1] A.Q. Kadhim, G.A. El-Refae, and S.F. El-Itter, "Neural Networks in Bank Insolvency Prediction," *International Journal of Computer Science and Network Security*, vol 10, no 5, pp. 240–245, 2010.
- [2] T. Pavlenko, and O. Chernyak, "Credit risk modeling using bayesian networks," *International Journal of Intelligent Systems*, vol. 25, issue 4, pp.326–344, 2010.
- [3] G.O. Chornous, *Proactive Management of Socio-Economic Systems Based on Intellectual Data Analysis: Methodology and Models*. Kyiv: Kyiv University, 2014.
- [4] M. Kim, and D. Kang, "Ensemble with neural networks for bankruptcy prediction," *Expert Systems with Applications*, vol. 37, issue 4, pp. 3373–3379, 2010.
- [5] C.-F. Tsau, and J.-W. Wu, "Using neural network ensembles for bankruptcy prediction and credit scoring," www.sciencedirect.com
- [6] L.I. Kuncheva, *Combining Pattern Classifiers: Methods and Algorithms*. New York: Wiley and Sons, 2014.
- [7] L.R. Medsker, *Hybrid Intelligent Systems*. Boston: Springer, 2013.
- [8] L. Zernova, *The creditworthiness of bank's clients: Analysis and assessment*. LAP LAMBERT Academic Publishing, 2016
- [9] N. Siddiqi, *Intelligent Credit Scoring: Building and Implementing Better Credit Risk Scorecards*, 2nd ed.. Wiley, 2017.
- [10] L. Thomas, J. Crook, and D. Edelman, *Credit Scoring and Its Applications*, 2nd Revised ed. SIAM-Society for Industrial & Applied Mathematics, 2017.
- [11] H. Chen, M. Jiang, and X. Wang, "Bayesian Ensemble Assessment for Credit Scoring," 4th International Conference on Industrial Economics System and Industrial Security Engineering (IEIS), 2017.
- [12] S. Dahiya, S.S. Handa, and N.P. Singh, "Credit Scoring Using Ensemble of Various Classifiers on Reduced Feature Set," *Industrija*, vol.43, no.4, pp. 163-172, 2015.
- [13] A. G. Armaki, M. F. Fallah, M. Alborzi, and A. Mohammadzadeh, "A Hybrid Meta-Learner Technique for Credit Scoring of Banks' Customers," *Engineering, Technology & Applied Science Research*, vol. 7, no. 5, pp. 2073-2082, 2017.
- [14] S. H. Van, N. N. Ha, and H. N. Thi Bao, "A hybrid feature selection method for credit scoring," *EAI Endorsed Transactions on Context-aware Systems and Applications*, vol. 4, issue 1, 09 2016 - 03 2017.
- [15] H. Xiao, Z. Xiao, and Y. Wang, "Ensemble classification based on supervised clustering for credit scoring," *Applied Soft Computing*, vol. 43, pp. 73-86, June 2016.
- [16] M. Ala'raj and M. F. Abbod, "A new hybrid ensemble credit scoring model based on classifiers consensus system approach," *Expert Systems with Applications*, vol. 64, pp. 36-55, December 2016.
- [17] Analytics Vidhya / Loan Prediction: Practice Problem // <https://datahack.analyticsvidhya.com/contest/practice-problem-loan-prediction-iii/>
- [18] K. S. Cho, "Ensemble learning with feature selection for Alzheimer's disease prediction," – <http://www.academia.edu/30496678>, 2016.
- [19] B. Himmetoglu, "Stacking models for improved predictions" – <https://burakhimmetoglu.com/2016/12/01/stacking-models-for-improved-predictions/>, 2017

A Hybrid Neuro-Fuzzy Element: a New Structural Node for Evolving Neuro-Fuzzy Systems

Zhengbing Hu

*School of Educational Information Technology
Central China Normal University
Wuhan, China
hzb@mail.ccnu.edu.cn*

Oleksii K. Tyshchenko^{1,2}

*¹Institute for Research and Applications of Fuzzy Modeling, CE
IT4Innovations, University of Ostrava
Ostrava, Czech Republic
²Control Systems Research Laboratory,
Kharkiv National University of Radio Electronics
Kharkiv, Ukraine
lehatish@gmail.com*

Abstract — A modification of the structure for a neuro-fuzzy unit was offered which is generally a hybrid system that combines nonlinear synapses and an activation function to form the hybrid system's output value. The introduced neuro-fuzzy element is specifically an extension of the common neo-fuzzy neuron which is upgraded at the expense of application of an additional (contracting) activation function. A particular robust learning procedure is also considered for this case that makes it possible to reduce errors while processing data containing abnormal observations.

Keywords — *Learning Method; Evolving System; Computational Intelligence; Neuro-Fuzzy Unit; Data Stream Processing; Hybrid Neuro-Fuzzy System; Machine Learning.*

I. INTRODUCTION

Multiple real-world systems and applications generate huge sequences of observations (data flows/streams) that arrive sequentially at a high speed. Data analysis should be brought into play in real time using limited storage and computing capabilities. As far as is known, Data Stream Mining [1-5] deals with extracting knowledge structures from continuous rapid data. Hybrid Computational Intelligence systems [6-10] like neuro-fuzzy systems, common neural networks, and wavelet neuro-fuzzy systems have proliferated extensively for taking decisions on a vast class of challenges [8-10] that come into existence within Data Mining in relation to their universal fitting properties and their aptitude of linguistic interpretation for obtained results. Most of these systems have proven their efficacy in problems of intensively developing Data Stream Mining [3-5], where information to be processed is fed sequentially (an observation by an observation), and a tuning process of the system's parameters should be carried out in an online mode by means of adaptive learning algorithms. First of all, we should mention here such systems where an output signal depends linearly on parameters to be tuned as radial-basis function neural networks [11-20], neuro-fuzzy systems by Takagi-Sugeno-Kang [21-22], hybrid systems that use neo-fuzzy neurons [23-30] as their nodes.

Experts' attention in the area of Computational Intelligence has been attracted to deep neural networks [31-39] recently. These networks considerably transcend conventional shallow neural networks regarding the quality of information processing. At the same time, they are instantiated by a low rate of learning that can be explained by

a necessity of using the error backpropagation method for multiple hidden layers of the network. In view of this, it seems reasonable enough to synthesize/update a structure (a computational unit) and its learning methods to be later used as a part of some more complex computational systems like evolving cascade systems and deep neural networks. A point to be noted here about the novelty is the fact that a new type of membership functions is used in synapses to raise approximating abilities of a hybrid neuro-fuzzy element as well as different types of activation functions are considered for the introduced topology based on the fundamental properties of an issue being solved. The paper is composed in such a manner. Section 2 comprises complementary information concerning a neuro-fuzzy unit and its modification. Section 3 describes a robust learning procedure applied to the system and different activation functions to be used for practical applications. Section 4 embodies experimental results of the presented neuro-fuzzy system. Conclusions are presented in the last section.

II. A MODIFICATION OF THE NEURO-FUZZY UNIT'S STRUCTURE

Ye. Bodyanskiy and S. Popov proposed in broad brushstrokes further changes [40, 41] to a topology of the neo-fuzzy neuron (NFN) [23-25]. The proposed system is intrinsically a hybrid combination of a neuro-fuzzy system (more specifically the neo-fuzzy neuron) and the elementary neuron by McCulloch and Pitts (also known as the MCP neuron). The exploited architecture eliminates shortcomings (which are typical for the neo-fuzzy neuron) at the cost of induction into the system's structure of a tightening activation function that brings some additional nonlinear effect.

That means a nonlinear calculative framework (Fig.1) which embodies nonlinear synapses (Fig.2) succeeded by a summation block and a nonlinear activation function to calculate the system's output.

Said another way, input signals are transformed with the help of synapses into the signals $f_i(x_i)$. These signals are

later joined into the internal activation signal $u = \sum_{i=1}^n f_i(x_i)$.

The neuron's output signal is made up with a nonlinear activation function $y = \psi(u)$.

As a general matter, a NFU output signal executes a mathematical transformation (based on the quadratic criterion) in a certain way

$$y = \psi(u) = \psi\left(\sum_{i=1}^n f_i(x_i)\right) = \psi\left(\sum_{i=1}^n \sum_{j=1}^{h_i} w_{ij} \mu_{ij}(x_i)\right)$$

where $\psi(\bullet)$ notes a nonlinear activation function (either a sigmoid function or a hyperbolic tangent); x_i describes input signals; $\mu_{ij}(x_i)$ marks membership levels; w_{ij} determines synaptic weights; h describes a quantity of fuzzy spans; n stands for a plurality of inputs; y is an output value.

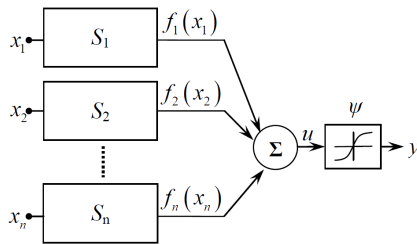


Fig. 1. A structure of the neuro-fuzzy unit

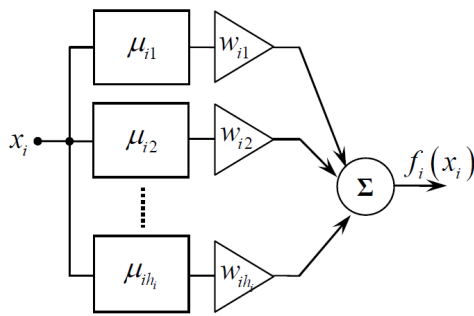


Fig. 2. A scheme of a nonlinear synapse in the neo-fuzzy neuron

On a large scale, triangular membership functions (Fig.3) stand on a distance between the input x_i and centers c_{ij}

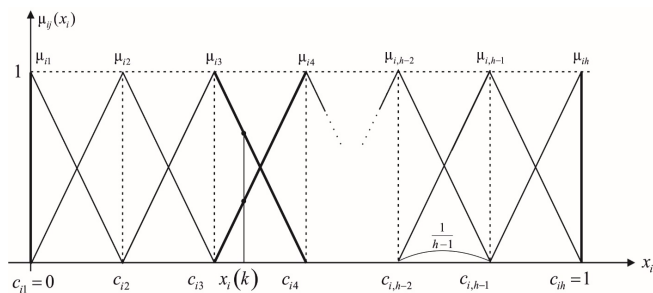


Fig. 3. Triangular membership functions

$$\mu_{i1}(x_i) = (c_{i1} - x_i) / c_{i2},$$

$$\mu_{ij}(x_i) = \begin{cases} (x_i - c_{i,j-1}) / (c_{ij} - c_{i,j-1}), & x_i \in [c_{i,j-1}; c_{ij}] \\ (c_{i,j+1} - x_i) / (c_{i,j+1} - c_{ij}), & x_i \in [c_{ij}; c_{i,j+1}] \\ 0 & \text{otherwise.} \end{cases} \quad (1)$$

$$\mu_{ih}(x_i) = (x_i - c_{i,h-1}) / (1 - c_{i,h-1}),$$

$$c_{i1} = 0, c_{i2} = 1 / (h-1), c_{il} = (l-1) / (h-1), c_{ih} = 1.$$

It's essentially taken all the initial data to be coded in the range [0;1]. It's crucial that this type of constructing membership functions makes automatically provision of the Ruspini (unity) partition

$$\sum_{j=1}^h \mu_{ij}(x_i) = 1 \quad \forall i.$$

Let's hypothesize that a fuzzy interval p is currently active, an output of the nonlinear synapse may be presented in this fashion

$$f_i(x_i) = \sum_{j=1}^h w_{ij} \mu_{ij}(x_i) = w_{ip} \mu_{ip}(x_i) + w_{i,p+1} \mu_{i,p+1}(x_i) =$$

$$= \frac{c_{i,p+1} - x_i}{c_{i,p+1} - c_{ip}} w_{ip} + \frac{x_i - c_{ip}}{c_{i,p+1} - c_{ip}} w_{i,p+1}.$$

Having said stated above, triangular membership constructions (1) are conventionally made use of as activation functions in the neo-fuzzy neuron. It may bring some obstruction for processes' simulation exemplified by differentiable (smooth) functions. The piecewise linear fitting appeared in this case by the neo-fuzzy neuron can account for a diminished accuracy level of the results obtained. A quantity of membership functions could be increased to lessen this negative effect. But finally, it results in enlargement of a number of weight coefficients, and the structure's complexity is growing along with the learning time required. The announced drawback may be avoided by means of cubic spline membership functions to be represented as follows

$$\mu_{ij}(x_i) = \begin{cases} 0.25 \left(2 + 3 \frac{2x_i - c_{ij} - c_{i,j-1}}{c_{ij} - c_{i,j-1}} - \left(\frac{2x_i - c_{ij} - c_{i,j-1}}{c_{ij} - c_{i,j-1}} \right)^3 \right), & x_i \in [c_{i,j-1}; c_{ij}] \\ 0.25 \left(2 - 3 \frac{2x_i - c_{i,j+1} - c_{ij}}{c_{i,j+1} - c_{ij}} + \left(\frac{2x_i - c_{i,j+1} - c_{ij}}{c_{i,j+1} - c_{ij}} \right)^3 \right), & x_i \in [c_{ij}; c_{i,j+1}]. \end{cases}$$

At some time moment, an input signal engages only two adjoining functions (Fig.4) contemporaneously (this case is rather similar to the triangular membership functions). But the provided set of functions doesn't cater to the needs of the Ruspini partition. Conversely, application of the cubic spline functional relations actualizes smooth polynomial fitting as a substitute for piecewise linear approximation and raises the possibility of carrying out the top grade simulation of substantively nonstationary and nonlinear signals.

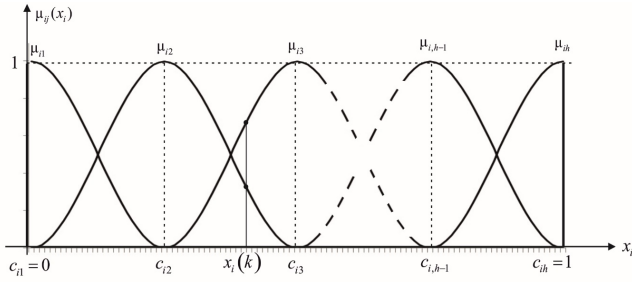


Fig. 4. Cubic spline membership functions

Giving effect to additional nonlinearity (compared to the conventional NFN) at the neuron's output quantity culminates in automatic containment of the element's output amplitude (which is especially relevant for construction of complex multilayer networks). A procedure of the NFE weights' tuning is exploited with reference to the quadratic criterion

$$\begin{aligned}
 E(k) &= \frac{1}{2} (d(k) - y(k))^2 = \frac{1}{2} e^2(k) = \\
 &= \frac{1}{2} (d(k) - \psi(u(k)))^2 = \\
 &= \frac{1}{2} \left(d(k) - \psi \left(\sum_{i=1}^n \sum_{j=1}^{h_i} w_{ij} \mu_{ij}(x_i(k)) \right) \right)^2
 \end{aligned} \quad (2)$$

where k denominates a unit of discrete time; $d(k)$ marks a reference signal; $e(k)$ stands for a learning error;

$$\begin{aligned}
 w_i &= (w_{i1}, w_{i2}, \dots, w_{ih})^T; \\
 \mu_i(x_i(k)) &= (\mu_{i1}(x_i(k)), \dots, \mu_{ih}(x_i(k)))^T.
 \end{aligned}$$

In the interest of minimization of the equation (2), the gradient descent learning procedure should be applied

$$\begin{aligned}
 w(k+1) &= w(k) - \eta(k) \nabla_w E(k) = \\
 &= w(k) + \eta(k) e(k) \frac{\partial e(k)}{\partial u(k)} \nabla_w u(k) = \\
 &= w(k) + \eta(k) e(k) \psi'(u(k)) \mu(x(k))
 \end{aligned} \quad (3)$$

where $\eta(k)$ designates a learning rate.

III. A ROBUST LEARNING PROCEDURE AND ACTIVATION FUNCTIONS USED

Utilizing the neuro-fuzzy element is quite challenging for signal processing. The remarkable thing is that its nonlinear properties can be set up by means of membership functions' parameters inside the nonlinear synapses. From this perspective, outliers may be put out, and an impact of less limitative input terms should get diminished noticeably. Learning methods on the grounds of the quadratic criteria (2) are highly exposed to data distribution's deviations. In the context of various types of irregular observations, learning methods based on the quadratic criteria don't demonstrate high efficiency due to obstacles with the "heavy tail" distribution and massive errors. In these cases, robust estimation methods [42] seem to be the most effective and

appropriate ones [43-44]. The criterion mentioned below is very popular in the theory of robust estimation

$$E^R(k) = \beta \ln \left(\cosh \frac{e(k)}{\beta} \right) \quad (4)$$

where $e(k)$ stands for a learning error; β denotes a scalar parameter magnitude to be chosen commonly in terms of a posteriori knowledge in order to appoint susceptibility for anomalous faults.

According to the initial paper [41], the improved learning procedure for the hybrid neuro-fuzzy element provides an opportunity to reduce processing errors for irregular samples by introducing the robust learning criterion. Through the lens of the learning procedure (3) grounded in this criterion (4) for the neuro-fuzzy element, the tweakage process of the system's weights may be represented in this view

$$\begin{aligned}
 w(k+1) &= w(k) - \eta(k) \nabla_w E^R(k) = \\
 &= w(k) + \eta(k) \beta \tanh \frac{e(k)}{\beta} \psi'(u(k)) \mu(k) = \\
 &= w(k) - \eta(k) \delta^R(k) \mu(k)
 \end{aligned} \quad (5)$$

where

$$\frac{\partial E^R(k)}{\partial e} = \beta \tanh \frac{e(k)}{\beta}; \quad \delta^R(k) = \beta \tanh \frac{e(k)}{\beta} \psi'(u(k)).$$

Besides that, the sigmoid is smooth and dependent on x . The fault is suitable for backpropagation, and weights should be updated subsequently. But anyway, there several issues to be addressed with that. The curve is quite plane beyond the $[-3;3]$ interval which means that once the relationship finds the way in that bracket, its gradients start descending (the gradient is verging to zero, and the network doesn't receive any training in the actual circumstances). An alternative issue that has to do with the logistic function (a sigmoid curve) is that its meanings only gauge in $(0;1)$. This means that the sigmoid curve is not symmetric around the reference point, and the implications gained are positive. But what if there's no need to send permanently to a subsequent neuron the values to be all of the same sign. One of the possible solutions for this case is scaling the sigmoid curve. The tanh function is of the nature of the logistic function and is in sober fact just a scaled version of it. Tanh works in the same fashion compared to the sigmoid relation, but it is symmetric over the initial point and sites from -1 to 1. It principally addresses the challenge of the meanings all being of the same sign. The rest of features are identical to the logistic curve. It is continuous and differentiable at all points. The functional relation is nonlinear and may be applied easily to backpropagating errors. Speaking of the tanh gradient, it's steeper in comparison with the sigmoid function. A selection between sigmoid and tanh essentially is stipulated by the gradient precondition for a problem statement. But there's also the vanishing gradient problem (the tanh graph is plane, and the gradients obtained are close to zero). The softmax function (the normalized exponential function) is some sort of the logistic function, but it's favorable when it comes to handling classification issues. The softmax distribution

would compress output implications for every group between 0 and 1 and divide by the outputs' sum. In order to get a probability distribution of outputs, the softmax function is usually put in requisition to impart probabilities when there is more than one output. It's chiefly advantageous when there is a need to find the most probable occurrence of an output with respect to other ones. Several more words should be said about a sigmoid-weighted linear unit (SiLU)

$$f(x) = \frac{x}{1+e^{-x}} = x\sigma(x)$$

where $\sigma(x)$ signifies the sigmoid function. Its derivative is

$$f'(x) = f(x) + \sigma(x)(1-f(x))$$

and it's generally used as a function approximator for neural networks in reinforcement learning.

The last functional relation to be mentioned is SoftPlus

$$f(x) = \ln(1+e^x).$$

And its derivative is the logistic (sigmoid) function. That's naturally a smooth approximation of a rectified linear unit which is broadly exploited in deep learning, computer vision, and speech recognition. All of these fore mentioned functions could be utilized as activation functions in the hybrid neuro-fuzzy element specifically from the perspective of a task type under consideration and some initial conditions of the problem.

IV. EXPERIMENTS

Theoretical aspects of our research were validated with the help of an experimental study depicting the forecasting challenge of electric loads. An available data sample comprised 6380 values documenting 6 months of electric power consumption in 2012 in Kharkiv region (Ukraine). A number of experiments were conducted to compare performance and prediction results. In our experimental part, we used two learning criteria (the criterion (2) and the robust criterion (5)), a different number of membership functions as well as such activation functions as the hyperbolic tangent and the sigmoid-weighted linear unit (SiLU). The data set was split to training and test data arrays. In our experimental research, for the purpose of simplicity, a quite unsophisticated model embodying only a single NFU was ample. Plots of the data array in Figs.5-6 illustrate apparently footprints of outliers stipulated by peak loads, measuring faults, and other factors. The outliers' fortuitous character is almost unpredictable and results in high prediction errors. It can be seen from Figs.5 and 6 that a prediction quality is growing (RMSE and SMAPE are gradually falling down). It should be noted that in case the outliers' assessed values are used straightforward to manage the learning flow, all in all, that could lead to the model parameters' distortion and consequently to a low rate of a prediction quality. Experimental results demonstrate a dependency between a forecasting accuracy and a number of membership functions (Figs.5-6). There's also a dependence between a forecasting error and an amount of membership functions illustrated in

This scientific work was financially subvented by self-determined research funds of CCNU from the colleges' basic research and operation of MOE (CCNU16A02015).

Fig.7.

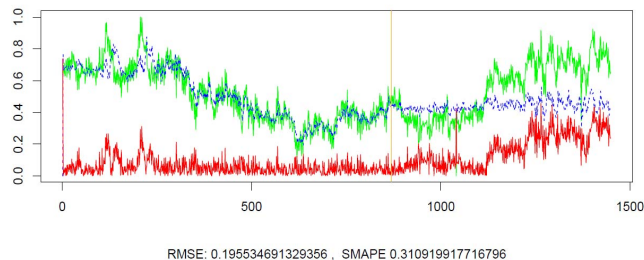


Fig. 5. A forecast performed by the hybrid neuro-fuzzy element (2 membership functions; the tanh activation function)

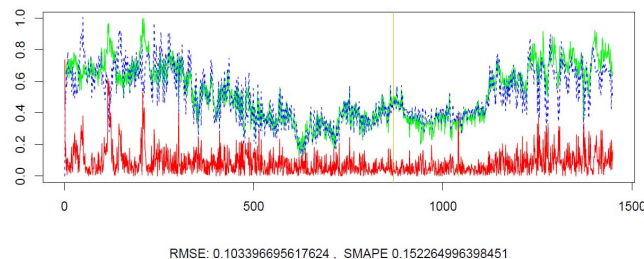


Fig. 6. A forecast performed by the hybrid neuro-fuzzy element (5 membership functions; the SiLU activation function)

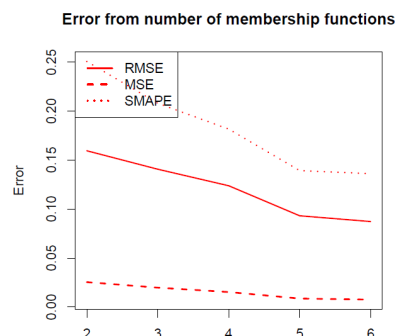


Fig. 7. A forecasting error for the hybrid neuro-fuzzy element depending on a number of membership functions

V. CONCLUSION

The described modification for the hybrid neuro-fuzzy element has been developed as a structural node for more complex computational systems like evolving cascade neuro-fuzzy systems and deep learning systems. Specifications of the membership functions can be set up in a rather straightforward manner to restrict large input values and contract extreme values. A few activation functions were considered for the offered modification in the neuro-fuzzy node in dependence to the nature of a task at hand. It was also recommended to give rise to approximation qualities of the system by applying the cubic splines as the membership functions. Having said that, it should be highlighted that the developed element is quite simple from the actualization point of view and keeps in possession approximating properties and a high processing speed.

ACKNOWLEDGMENT

This scientific work was financially assisted by self-determined research funds of CCNU from the colleges' basic research and operation of MOE (CCNU16A02015).

Oleksii K. Tyshchenko is kindly grateful for the financial assistance of the Visegrad Scholarship Program—EaP #51700967 funded by the International Visegrad Fund (IVF).

REFERENCES

- [1] J. Gama, *Knowledge Discovery from Data Streams*. Boca Raton: Chapman and Hall/CRC, 2010.
- [2] A. Bifet, R. Gavaldà, G. Holmes, and B. Pfahringer, *Machine Learning for Data Streams with Practical Examples in MOA*. The MIT Press, 2018.
- [3] A. Bifet, *Adaptive Stream Mining: Pattern Learning and Mining from Evolving Data Streams*. Amsterdam: IOS Press, 2010.
- [4] Ye.V. Bodyanskiy, O.K. Tyshchenko, and D.S. Kopalani, "An Evolving Connectionist System for Data Stream Fuzzy Clustering and Its Online Learning", *Neurocomputing*, vol. 262, pp.41-56, 2017.
- [5] M. Garofalakis, J. Gehrke, and R. Rastogi (eds.), *Data Stream Management. Processing High-Speed Data Streams*. Berlin Heidelberg: Springer-Verlag, 2016.
- [6] R. Kruse, C. Borgelt, F. Klawonn, C. Moewes, M. Steinbrecher, and P. Held, *Computational Intelligence. A Methodological Introduction*. Berlin: Springer-Verlag, 2013.
- [7] B.M. Wilamowski and J.D. Irwin, *Intelligent Systems*. Boca Raton: CRC Press, 2017.
- [8] J. Kacprzyk and W. Pedrycz (eds.), *Springer Handbook of Computational Intelligence*. Berlin Heidelberg: Springer-Verlag, 2015.
- [9] C.L. Mumford and L.C. Jain, *Computational Intelligence*. Berlin: Springer-Verlag, 2009.
- [10] Ye. Bodyanskiy, I. Pliss, D. Peleshko, Yu. Rashkevych, and O. Vynokurova, "Hybrid Generalized Additive Wavelet-Neuro-Fuzzy-System and its Adaptive Learning", *Dependability Engineering and Complex Systems: Proc. of the Eleventh International Conference on Dependability and Complex Systems DepCoS-RELCOMEX*, Brunow, Poland, pp. 51-61, 2016.
- [11] D. Graupe, *Principles of Artificial Neural Networks (Advanced Series in Circuits and Systems)*. Singapore: World Scientific Publishing Co. Pte. Ltd., 2007.
- [12] K.-L. Du and M.N.S. Swamy, *Neural Networks and Statistical Learning*. London: Springer, 2014.
- [13] K. Suzuki, *Artificial Neural Networks: Architectures and Applications*. NY: InTech, 2013.
- [14] R. Tkachenko and I. Izonin, "Model and Principles for the Implementation of Neural-Like Structures based on Geometric Data Transformations". *Advances in Computer Science for Engineering and Education. ICCSEEA2018. Advances in Intelligent Systems and Computing*, 2018, in press.
- [15] G. Hanrahan, *Artificial Neural Networks in Biological and Environmental Analysis*. NW: CRC Press, 2011.
- [16] Ye. Bodyanskiy, O. Tyshchenko, and A. Deineko, "An Evolving Radial Basis Neural Network with Adaptive Learning of Its Parameters and Architecture", *Automatic Control and Computer Sciences*, Vol. 49, No. 5, pp. 255-260, 2015.
- [17] S. Haykin, *Neural Networks and Learning Machines (3rd Edition)*. NJ: Prentice Hall, 2009.
- [18] R. Tkachenko, P. Tkachenko, I. Izonin, and Y. Tsymbal, "Learning-based image scaling using neural-like structure of geometric transformation paradigm", *Studies in Computational Intelligence*, vol. 730, pp. 537–565, 2018.
- [19] S. Bassis, A. Esposito, F. C. Morabito, *Advances in Neural Networks: Computational and Theoretical Issues*. Springer International Publishing, 2016.
- [20] I. Izonin, R. Tkachenko, D. Peleshko, T. Rak and D. Batyuk, "Learning-based image super-resolution using weight coefficients of synaptic connections", *Proc. Xth International Scientific and Technical Conference "Computer Sciences and Information Technologies" (CSIT)*, Lviv, Ukraine, pp. 25-29, 2015.
- [21] J-S.R. Jang, C.T. Sun and E. Mizutani, *Neuro-Fuzzy and Soft Computing: A Computational Approach to Learning and Machine Intelligence*, New Jersey: Prentice Hall, 1997.
- [22] L.X. Wang, *Adaptive Fuzzy Systems and Control. Design and Stability Analysis*. Upper Saddle River, New Jersey: Prentice Hall, 1994.
- [23] E. Uchino and T. Yamakawa, "Soft computing based signal prediction, restoration and filtering", *Intelligent Hybrid Systems: Fuzzy Logic, Neural Networks and Genetic Algorithms*, Boston: Kluwer Academic Publisher, pp. 331-349, 1997.
- [24] Ye. Bodyanskiy, O. Tyshchenko, and D. Kopalani, "An Extended Neo-Fuzzy Neuron and its Adaptive Learning Algorithm", *IJ. Intelligent Systems and Applications (IJISA)*, Vol.7(2), pp.21-26, 2015.
- [25] T. Miki and T. Yamakawa, "Analog implementation of neo-fuzzy neuron and its on-board learning", *Computational Intelligence and Applications*, Piraeus: WSES Press, pp. 144-149, 1999.
- [26] Ye. Bodyanskiy, O. Tyshchenko, and D. Kopalani, "A hybrid cascade neural network with an optimized pool in each cascade", *Soft Computing. A Fusion of Foundations, Methodologies and Applications (Soft Comput)*, Vol.19, No.12, pp.3445-3454, 2015.
- [27] Zh. Hu, Ye.V. Bodyanskiy, and O.K. Tyshchenko, "A Deep Cascade Neural Network Based on Extended Neo-Fuzzy Neurons and its Adaptive Learning Algorithm", *Proc. of 2017 IEEE First Ukraine Conference on Electrical and Computer Engineering (UKRCON)*, Kyiv, Ukraine, pp.801-805, 2017.
- [28] Zh. Hu, Ye.V. Bodyanskiy, O.K. Tyshchenko, and O.O. Boiko, "An Evolving Cascade System Based on a Set of Neo-Fuzzy Nodes", *International Journal of Intelligent Systems and Applications (IJISA)*, Vol. 8(9), pp.1-7, 2016.
- [29] Zh. Hu, Ye.V. Bodyanskiy, and O.K. Tyshchenko, "A Hybrid Growing ENFN-Based Neuro-Fuzzy System and its Rapid Deep Learning", *Proc. of 12th International Scientific and Technical Conference on Computer Sciences and Information Technologies (CSIT'2017)*, Lviv, Ukraine, pp.514-519, 2017.
- [30] Ye. Bodyanskiy, O. Tyshchenko, and D. Kopalani, "Adaptive learning of an evolving cascade neo-fuzzy system in data stream mining tasks", *Evolving Systems*, Vol.7, No.2, pp.107-116, 2016.
- [31] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016.
- [32] Ye. Bodyanskiy, O. Vynokurova, P. Mulesa, G. Setlak, and I. Pliss, "Fast Learning Algorithm for Deep Evolving GMDH-SVM Neural Network in Data Stream Mining Tasks", *Proc. of 2016 IEEE First International Conference on Data Stream Mining & Processing (DSMP)*, pp. 257-262, 2016.
- [33] A. L. Caterini and D. E. Chang, *Deep Neural Networks in a Mathematical Framework*. Springer, 2018.
- [34] J. Heaton, *Artificial Intelligence for Humans, Volume 3: Deep Learning and Neural Networks*. CreateSpace Independent Publishing Platform, 2015.
- [35] A. Menshaw, *Deep Learning By Example: A hands-on guide to implementing advanced machine learning algorithms and neural networks*. Packt Publishing Limited, 2018.
- [36] M. Fullan, J. Quinn, and J. McEachen, *Deep Learning: Engage the World Change the World*. Corwin, 2017.
- [37] D. Graupe, *Deep Learning Neural Networks: Design and Case Studies*. World Scientific, 2016.
- [38] Y. LeCun, Y. Bengio, G.E. Hinton, "Deep learning", *Nature*, vol. 521, pp. 436–444, 2015.
- [39] J. Schmidhuber, "Deep learning in neural networks: an overview," *Neural Networks*, no. 61, pp. 85–117, 2015.
- [40] Y. Bodyanskiy and S. Popov, "Neuro-fuzzy unit for real-time signal processing," *Proc. IEEE East-West Design & Test Workshop (EWDTW'06)*, Sochi, Russia, pp. 403-406, 2006.
- [41] Ye. Bodyanskiy, S. Popov, and M. Titov, "Robust learning algorithm for networks of neuro-fuzzy units", *Innovations and Advances in Computer Sciences and Engineering*, pp. 343-346, 2010.
- [42] W. J. J. Rey, *Robust Statistical Methods*. Berlin-Heidelberg-New York: Springer, 1978.
- [43] D. S. Chen and R. C. Jain, "A Robust Back Propagation Learning Algorithm for Function Approximation", *IEEE Trans. Neural Networks*, vol. 5, pp. 467-479, 1994.
- [44] A. Cichocki and R. Unbehauen, *Neural Networks for Optimization and Signal Processing*. Stuttgart: Teubner, 1993.

Implementation of Neural Networks with Help of a Data Flow Virtual Machine

Kostyantyn Kharchenko
System Design Department
Institute for Applied System Analysis
National Technical University of
Ukraine "Igor Sikorsky Kyiv
Polytechnic Institute"
Kyiv, Ukraine
k.kharchenko@kpi.ua

Oleksandr Beznosyk
System Design Department
Institute for Applied System Analysis
National Technical University of
Ukraine "Igor Sikorsky Kyiv
Polytechnic Institute"
Kyiv, Ukraine
o.beznosyk@kpi.ua

Valery Romanov
System Design Department
Institute for Applied System Analysis
National Technical University of
Ukraine "Igor Sikorsky Kyiv
Polytechnic Institute"
Kyiv, Ukraine
v.romanov@kpi.ua

Abstract—The main goal of this paper is to show how a neural network can be implemented with help of the data flow management system at a virtual machine. As an example, the three-layer neural network realization has been investigated to solve a simple XOR function with two inputs and one output. For that purpose, a sigmoid command required to make a neuron activation function has been added into the data flow virtual machine. It is presented in the paper that neural networks can be described as data flows with help of the declarative approach on a base of the JSON format.

Keywords—neural networks, data flow virtual machine, JSON, activation functions

I. INTRODUCTION

Using virtual machines in computational systems is well-known for years. Such an approach established itself well as it supports compatibility at the bytecode level and its productivity on various hardware platforms is often comparable to the software development with native code compilers.

In the same time, the data flow computations were developing actively in contrast to the command flow ones. Currently, data flow management systems are widespread in the different areas.

In the previous papers [1, 2, 3, 4], some implementations of the data flow virtual machine (DFVM) as well as a concept of input data representation for a data flow system were described. One of them is a JSON-based format for the data flow virtual machine that is rather simple and effective. So, at the moment a DFVM's input file is a JSON file that is easily understandable and self-descriptive.

Currently, it is possible to describe neural networks as data flows [5, 6, 7]. The aim of the paper presented is to create an example of the test neural network working at the data flow virtual machine. At the same time, the neural network input configuration is described by a static JSON file in contrast to the widespread approach to describe neural networks by means of the codes in Python, C++, Java etc. programming languages.

The possibility and convenience to describe a neural network in the declarative form at the level of neural network's separate signals is under investigation. The problem of the description at the layer's level will be considered separately.

II. EXISTING SOLUTIONS

There are a lot of tools for neural network computations such as, for example, TensorFlow [8], Theano [9], MXNet [10], CNTK [11], Keras [12] (known as symbolic frameworks), Torch [13, 14], Caffé [15] (imperative frameworks).

TensorFlow is a platform-independent open source software library for artificial intelligence and machine learning developed by Google for internal use to build and train neural networks for automatically finding and classifying images and correlations, with a goal to achieve the quality of human perception. It is currently used for researches as well as the development of Google products such as Speech Recognition, Gmail, Photos, Search, Maps. The main API for working with the library is implemented for Python while there are implementations for C++, Haskell, Java and Go too. TensorFlow computations are represented as stateful data flow graphs. Its name comes from so called "tensors" – multidimensional data arrays, on which such neural networks perform the operations. TensorFlow can run on multiple central and graphic processors (with optional CUDA and SYCL extensions for general-purpose computing on graphics processing units). In 2015, it was released under the open Apache 2.0 license.

Theano is a Python numerical computations library. It deals with symbolically specified mathematical computations and optimizes them to produce efficient low-level realization. The calculations in Theano are expressed by the NumPy syntax and compiled for efficient parallel processing on conventional CPUs and GPUs. On September 28, 2017, it was announced that work on the project was discontinued after the release of 1.0, while the minimum support was maintained for one year.

Apache MXNet is a modern open-source fast and scalable deep learning framework with a compact and easy-to-use machine learning API. It is used to train and deploy deep neural networks, and it supports a flexible programming model and multiple languages such as C++, Python, Perl, Scala, Matlab, Wolfram, JavaScript, Go, R. MXNet includes the Gluon interface, which makes it easy for developers to get started with deep learning, for example, in the cloud or in mobile applications. With just a few lines of Gluon code, such features as linear regression, convolutional networks, and recurring LSTMs for object and speech recognition, recommendation, and personalization can be developed. The MXNet library is portable. Also, it is scalable to multiple

graphic processors and computers. MXNet is supported by major public cloud providers such as Amazon and Microsoft as well as a number of world-famous companies and research institutions.

Microsoft Cognitive Toolkit, formerly known as CNTK, is a deep learning framework developed by Microsoft Research. It describes neural networks as a series of computational steps via a directed graph.

Keras is an open neural network library written in Python. It is an add-on for the DeepLearning4j, TensorFlow and Theano frameworks. It is capable of working on top of them. It is aimed at operational work with deep learning networks, while being designed to be compact, modular and expandable. It was planned that Google will support Keras in the main TensorFlow library, but Keras was designed as an interface rather than an end-to-end machine learning system. It represents a high-level, intuitive set of abstractions, which makes the formation of neural networks easy, regardless of the library of scientific computing used at the lower level. This library contains numerous implementations of widely used building blocks of neural networks, such as layers, target and transfer functions, optimizers, and many tools for simplifying the work with images and text.

Torch is an open library for the open source Lua programming language. It provides a lot of algorithms for deep machine learning and scientific computing. The kernel is written in C, the application part is executed on LuaJIT, also it supports parallelization of calculations by means of CUDA and OpenMP.

Caffe is a system for deep learning developed by Yangqing Jia as part of his doctoral work at the University of California at Berkeley. Caffe is open source software distributed under the BSD license. It is written in C++ and supports the Python interface. Its name comes from the reduction of "Convolution Architecture For Feature Extraction". Caffe first ported the MATLAB implementation of fast convolutional neural networks (CNN) to C and C++. Caffe includes numerous algorithms and deep-learning architectures for classifying and clustering image data. CNN, R-CNN (recurrent neural network), LSTM (long short-term memory) and fully connected neural networks are supported. With Caffe, the graphics processor-based acceleration can be used with Nvidia's cuDNN. Caffe supports Python and MATLAB programming environments. Yahoo has integrated Caffe into Apache Spark to distribute deep learning.

Some of the systems mentioned operate with data flows but all of them use an imperative form of the neural network description and, thus, require using programming languages to describe a neural network behavior. In the same time, the approach proposed in this article allows describing a neural network by means of the declarative JSON-like constructions; it seems to be more convenient and simple than using conventional programming languages. Each of the approaches mentioned has its own advantages and disadvantages but the main idea is that the declarative approach would allow working with neural networks for users without knowledge of the programming languages.

III. DATA FLOW VIRTUAL MACHINE FOR NEURAL NETWORK COMPUTATIONS

Let's consider an example often used for an acquaintance almost with any system or library for neural networks.

Let's assume that neural network's coefficients to solve a test XOR example are already known (the problem of neural network training to select coefficients required, for instance, by a gradient descent method is not currently under investigation).

At the current implementation of a neural network, one neuron can accept only two input signals. The XOR function to be implemented on the neural network is presented in Table 1.

TABLE I. XOR FUNCTION

A	0	0	1	1
B	0	1	0	1
Z	0	1	1	0

A. Implementation of Sigmoid Function in DFVM

Let's implement a sigmoid function

$$y = \frac{1}{1 + e^{-x}}$$

in DFVM. The following DFVM JSON file shows an example of using sigmoid; if the input value is 1.0 then the expected output is 0.731059:

```
{
  "comment": "sigmoid y = 1 / (1 + exp(-x))",
  "inputs": [ { "name": "x", "value": 1.0 },
              { "name": "y", "value": 0.0 } ],
  "outputs": [ { "name": "y", "assert":
                [{"equal": 0.731059}] } ],
  "nodes": [ { "double": "x", { "double": "y" } ],
  "commands": [ { "code": "sigmoid", "inputs": ["x"],
                  "outputs": ["y"] } } ]
}
```

This sigmoid function can be used as an activation one.

For a lot of tasks in neural networks, it is needed also to use w_0 value (bias). Let's add it to the sigmoid function:

$$y = \frac{1}{1 + e^{-x + w_0}}$$

This will allow to simplify working with a neural network and to avoid adding one more DFVM component to implement addition (Fig. 1).

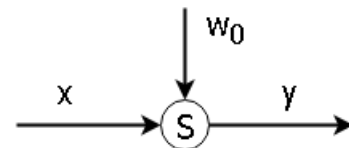


Fig. 1. DFVM implementation of sigmoid function with bias.

So, the DFVM sigmoid function takes two input arguments x and w_0 and looks as follows:

```
{
  "comment": "sigmoid y = 1 / (1 + exp(-x+w0))",
  "inputs": [ { "name": "x", "value": 1.0 },
              { "name": "w0", "value": 1.0 },
              { "name": "y", "value": 0.0 } ],
  "outputs": [ { "name": "y" } ],
}
```

```

"nodes": [ {"double": "x"},
            {"double": "w0"},
            {"double": "y"} ],
"commands": [ {"code": "sigmoid_bias",
               "inputs": ["x", "w0"],
               "outputs": ["y"]} ]
}

```

B. Description of a Simple Neural Network with the Data Flow Paradigm

The following data flow JSON file implements a simple neural network in the data flow virtual machine. The first input layer consists of two neurons, the second (hidden) layer consists of two neurons, and the output layer consists of one neuron (Fig. 2).

This DFVM JSON file represents an implementation of neural network on data flow paradigm for XOR sample shown in Fig. 3.

Each neuron has a w_{ij} coefficient for multiplication of the neuron's input and passes a result to the sigmoid activation

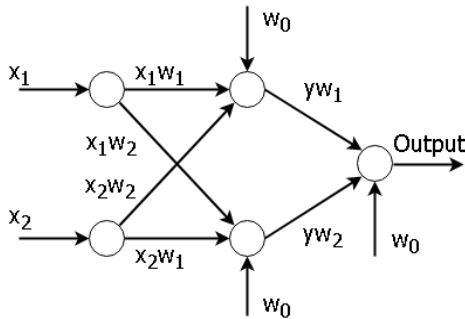


Fig. 2. Simple example of neural network for XOR function.

function. This neural network consists of two inputs $input1$, $input2$ and 6 weight coefficients w_{ij} . All these inputs are of *double* type. This neural network consists of two layers, and

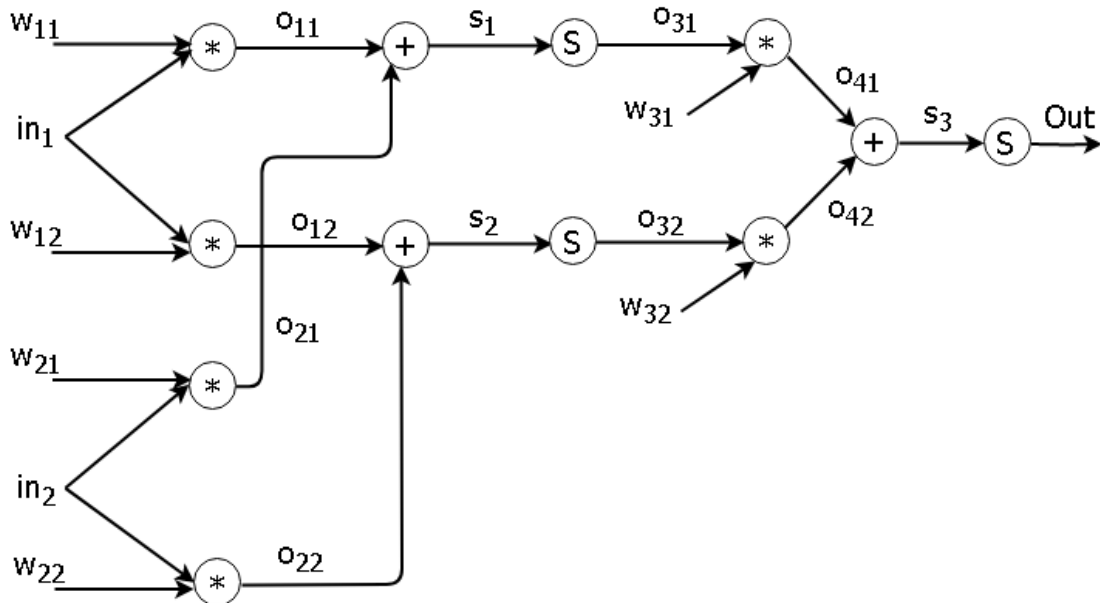


Fig. 3. Neural network on data flow paradigm.

sigmoids are used as activation functions:

```

{
  "comment": "neural network",
  "inputs": [ {"name": "input1", "value": 1.0},
              {"name": "input2", "value": 1.0},
              {"name": "w11", "value": 1.0},
              ...
              {"name": "w32", "value": 1.0} ],
  "outputs": [ {"name": "output"} ],
  "nodes": [ {"double": "input1"},
              {"double": "input2"},
              {"double": "w11"},
              ...
              {"double": "output"} ],
  "commands": [ {"code": "mul", "inputs": ["input1",
                                           "w11"], "outputs": ["o11"] },
                 {"code": "mul", "inputs": ["input2",
                                           "w21"], "outputs": ["o21"] },
                 {"code": "mul", "inputs": ["input1",
                                           "w12"], "outputs": ["o12"] },
                 {"code": "mul", "inputs": ["input2",
                                           "w22"], "outputs": ["o22"] },
                 {"code": "add", "inputs": ["o11",
                                           "o21"], "outputs": ["s1"] },
                 {"code": "add", "inputs": ["o12",
                                           "o22"], "outputs": ["s2"] },
                 {"code": "sigmoid", "inputs": ["s1"],
                                               "outputs": ["o31"] },
                 {"code": "sigmoid", "inputs": ["s2"],
                                               "outputs": ["o32"] },
                 {"code": "mul", "inputs": ["o31",
                                           "w31"], "outputs": ["o41"] },
                 {"code": "mul", "inputs": ["o32",
                                           "w32"], "outputs": ["o42"] },
                 {"code": "add", "inputs": ["o41",
                                           "o42"], "outputs": ["s3"] },
                 {"code": "sigmoid", "inputs": ["s3"],
                                               "outputs": ["output"]} ]
}

```

IV. IMPLEMENTATION DETAILS AND OTHER TYPES OF ACTIVATION FUNCTIONS IN DFVM

The existing data flow virtual machine did not require a lot of modifications, except for adding new commands for activation functions. A component diagram of DFVM is presented in Fig. 4.

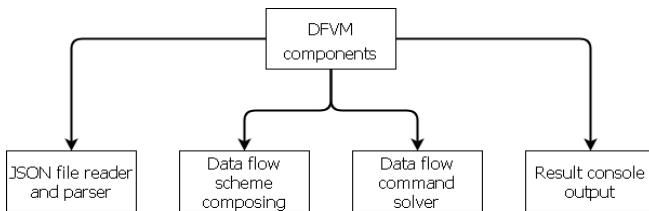


Fig. 4. DFVM component diagram.

At the moment, there are the following activation functions in DFVM:

- relu
- softsign
- sigmoid_bias
- softplus
- sigmoid
- tanh

In C++ code, an activation function implementation takes place in the same way as for usual DFVM commands such as *add*, *sub*, *div*, *mul* for a *double* type. The activation function can take one or more input parameters and return one output parameter. Entire C++ programming functional is available, and the implementation procedure itself is not complex, which makes it possible to add into DFVM new activation functions if needed.

The implementation of sigmoid function in DFVM in C++ is represented as follows:

```
BaseFlow* DoubleFlow::sigmoid()
{
    return new DoubleFlow(1.0/
        (1.0+exp(-this->value)));
}
```

where *BaseFlow* is a superclass of flow, *DoubleFlow* is a class for *double* type, and *this->value* is an input value. This method returns new instance of *DoubleFlow*.

DoubleFlow.h file looks as follows:

```
#pragma once
#include "BaseFlow.h"
#include "DFVMExceptions.h"
#include "IntegerFlow.h"
#include "StringFlow.h"
#include "BooleanFlow.h"
#include <iostream>

class DoubleFlow :public BaseFlow
{
public:
    explicit DoubleFlow(const double val);
    virtual ~DoubleFlow() {}
    void print() const override;
    BaseFlow::Type type() const override;
    ...
    BaseFlow* sigmoid() override;
private:
    double value;
};
```

V. RESEARCH RESULTS AND FUTURE INVESTIGATIONS

An advantage of the neural network description as data flows at the virtual machine is that such an approach does not require writing a program in any high-level language or such object-oriented programming languages as C++ or Python. In a case of the data flow virtual machine, the neural network description is just presented in a declarative form while calculations are being provided by a specialized virtual machine written in C++.

In the next implementations of DFVM a gradient descent

method will be implemented to train neural networks using data flow based description, which will allow training neural networks in the same data flow virtual machine without additional libraries.

A disadvantage of the approach mentioned is that it is required to describe connections with each neuron in the layers. However, if to provide support for matrices in the virtual machine, then it would be possible to describe each layer as a single entity in a JSON file, which will make it possible to present complex large-scale neural networks in some lines.

For high-dimensional neural networks, DFVM will support an array data type for *double* and corresponding description for appropriate operations, such as multiplication, addition, etc. This will allow to describe connection of a data flow with massive data and to lessen number of lines in a JSON file. The next version of DFVM will be presented for bigger examples of practical neural networks with connection of layers as data flows and proper mathematical training methods.

REFERENCES

- [1] K. V. Kharchenko, "Extension of the LLVM virtual machine with parallel instructions to implement a message transfer system," 2012 System analysis and information technology 14th Int. Conf., Kyiv, Ukraine, p. 302, 24 April 2012.
- [2] K. V. Kharchenko, "Dataflow control paradigm and dataflow graphic presentation in SOA," East-European journal for advanced technologies, no. 3/9 (69), pp. 22-29, 2014.
- [3] K. V. Kharchenko, "An Architecture and Test Implementation of Data Flow Virtual Machine," 2016 System analysis and information technology 18th Int. Conf., Kyiv, Ukraine, p. 268, 30 May – 2 June 2016.
- [4] K. Kharchenko, O. Beznosyk and V. Romanov, "A Set of Instructions for Data Flow Virtual Machine," IEEE First Ukraine Conference on Electrical and Computer Engineering (UKRCON 2017), Kyiv, Ukraine, pp. 931-934, 29 May – 2 June 2017.
- [5] B. Lu, B. L. Evans and D. V. Tosic, "Simulation and Synthesis of Artificial Neural Networks Using Dataflow Models in Ptolemy," 4th Seminar on Neural Network Applications in Electrical Engineering NEUREL-97, Belgrade, Serbia, pp. 84-89, Sep. 8-9, 1997.
- [6] M. Bacis, G. Natale, E. Del Sozzo and M. D. Santambrogio, "A pipelined and scalable dataflow implementation of convolutional neural networks on FPGA," 2017 IEEE International Parallel and Distributed Processing Symposium Workshops (IPDPSW), Lake Buena Vista, FL, pp. 90-97, 2017.
- [7] Y. H. Chen, J. Emer and V. Sze, "Using Dataflow to Optimize Energy Efficiency of Deep Neural Network Accelerators," in IEEE Micro, vol. 37, no. 3, pp. 12-21, 2017.
- [8] Jeffrey Dean et al. (2015, November 9). TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems [Online]. Available: <http://download.tensorflow.org/paper/whitepaper2015.pdf>
- [9] Theano GitHub [Online]. Available: <https://github.com/Theano>
- [10] MXNet: A Scalable Deep Learning Framework [Online]. Available: <https://mxnet.apache.org/>
- [11] Microsoft Cognitive Toolkit [Online]. Available: <https://www.microsoft.com/en-us/cognitive-toolkit/>
- [12] Keras Documentation [Online]. Available: <https://keras.io/>
- [13] Torch GitHub [Online]. Available: <https://github.com/torch/torch7>
- [14] Torch. Scientific computing for LuaJIT [Online]. Available: <http://torch.ch/>
- [15] Caffe Deep Learning Framework [Online]. Available: <http://caffe.berkeleyvision.org/>

Self-Diagnosis of the Systems with Intermittently Faulty Units

Viktor Mashkov

Jan Evangelista Purkyně University in Usti nad Labem
Usti nad Labem, Czech Republic
viktor.mashkov@ujep.cz

Jiří Fišer

Jan Evangelista Purkyně University in Usti nad Labem,
Usti nad Labem, Czech Republic
jf@jf.cz

Volodymyr Lytvynenko

Kherson National Technical University
Kherson, Ukraine
immun56@gmail.com

Maria Voronenko

Kherson National Technical University
Kherson, Ukraine
mary_voronenko@i.ua

Abstract—System level diagnosis is an abstraction of high level and, thus, its practical implementation to particular cases of complex systems is the task which requires additional investigations, both theoretical and modeling. Traditionally, system self-diagnosis is used for detecting of permanently faulty nodes. In the paper, we consider the problems of intermittent fault detection and suggest diagnosis procedures which allow distinguishing between different types of intermittent faults. For each type of intermittent faults the diagnosis procedure was developed

Keywords— complex systems; self-diagnosis; intermittent fault; diagnosis procedure Introduction

I. INTRODUCTION

Recent advances in the semiconductor technology have made it possible to design powerful single-chip microprocessors. With the considerable reduction in the cost of microprocessors, it is now feasible to build increasingly sophisticated microprocessor systems with up to thousands of microprocessors.

New opportunities have emerged for developing many-core processors with novel architectures that support better communication among cores and allow for better algorithm parallelization, resulting in a dramatic increase in processor performance [1].

The rapid growth of many-core processors has brought about an increasing demand for high processor reliability and availability.

Implementing many cores on a single die is possible due to shrinking of processing elements. Modern nano-scale technologies make it possible to integrate billions of transistors on a single die. As die size and transistor density grow, the susceptibility of these processors to hardware faults grows as well. Permanent and intermittent hardware faults, caused by defects in silicon or metallization and wear out over time, lead to circuit reliability problems. Due to these circuit reliability problems, dependability becomes one of the major challenges for all future nano-scale technologies; this is why fault-tolerance is becoming an essential property that must be integrated from the very beginning in every chip design.

In order to improve these processor attributes we need to develop new means for identifying and periodically

removing faulty processor cores (by reconfiguration). This can be achieved by masking redundancy, information redundancy which exploits different coding techniques and time redundancy where each computation is repeated on the same hardware. Built-in testing capabilities (so-called built-in-self-test schemes) are often used to improve the efficiency of these schemes. Unfortunately the wide applicability of built-in testing is undermined by the need to have some part of the processor (called the hard-core) operational even in the presence of fault.

Self-diagnosis, which uses the ability of processor cores to test each other, is now actively studied as a promising technique for providing processor cores checking and diagnosis [2], [3], [4].

We assume that the suggested in the paper approach to diagnose intermittent faults can be used not only for many-core processors but also for wide range of complex systems, which employ self-diagnosis.

II. CLASSIFICATION OF INTERMITTENT FAULTS IN THE CONTEXT OF SELF-DIAGNOSIS

Intermittent fault of a system unit can be defined as a fault, which randomly transfers from a latent state to an active state and vice versa. There exist several models, which describe the behavior of an intermittent fault [5], [6]. We use the model proposed in [6]. In this model, the behavior of intermittent fault is expressed by continuous Markov chain, where the time during which an intermittent fault stays in active, respectively passive state, is random value with exponential distribution (see Fig. 1).

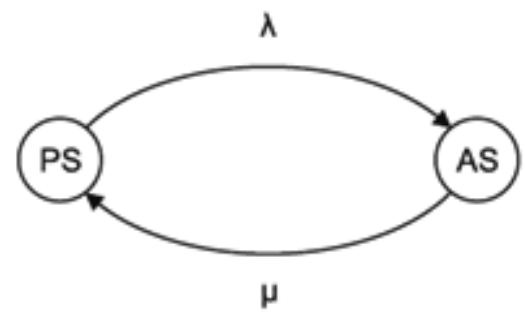


Fig. 1. Model of intermittent fault

In Fig. 1, λ and μ denote the rates of transition from passive to active state and vice versa.

When an intermittent fault is in active state it can cause an error in a system unit and affect the tests related to the erroneous unit.

System level self-diagnosis deals with the mutual testing. In this case, one of the system units performs tests on the other units. Mutual testing can be presented by testing graph. In Fig. 2, example of the testing graph for the system with three units is shown.

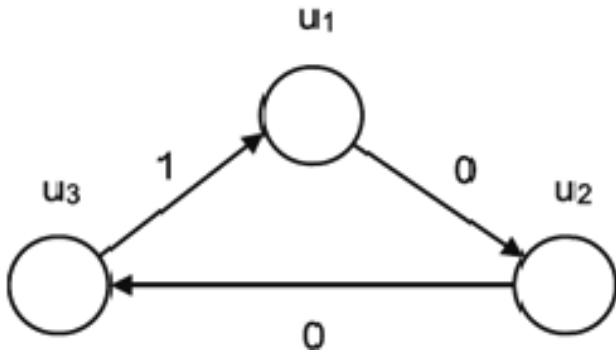


Fig. 2. Testing assignment for system with units

Generally, for providing diagnosis at system level some assumptions are made, such as:

- tests can be performed only in the periods of time when system units do not perform their proper system functions (i.e., when they are in idle state). That is, a system unit is not tested continuously, and, therefore, there exists a probability of not detecting the failed unit;
- even if a unit is failed, a test not always detects this event. It depends on test coverage;
- result of a test is expressed as 0 or 1 depending on the evaluation of the testing unit about the state of the tested unit;
- tests in a system can be performed either according to a predefined testing assignment or randomly.

In this paper, we consider that test coverage is equal to 100%. We also assume that tests among system units are performed according to predefined testing assignment. It means that the total time of testing is known beforehand. Consequently, the periods of time when each unit is involved in tests are also known in advance.

In the given case, it is possible to consider parameters of intermittent fault model (i.e., λ and μ) in relation to the total time of testing, $t_{testing}$. Table I presents possible evaluations of values $1/\lambda$ and $1/\mu$ as compared to the value of $t_{testing}$. Classes of intermittent faults

Cases (classes)	$1/\lambda$	$1/\mu$
case 1	small	small
case 2	small	large
case 3	large	small
case 4	large	large

This comparison bears some resemblance to the techniques based on fuzzy logic. We evaluate the values of $1/\lambda$ and $1/\mu$ as “large” and “small” depending on the ratios of values $1/\lambda$ ($1/\mu$) and $t_{testing}$.

Fig. 3 depicts the cases with different ratios of values $1/\lambda$ ($1/\mu$) and $t_{testing}$.

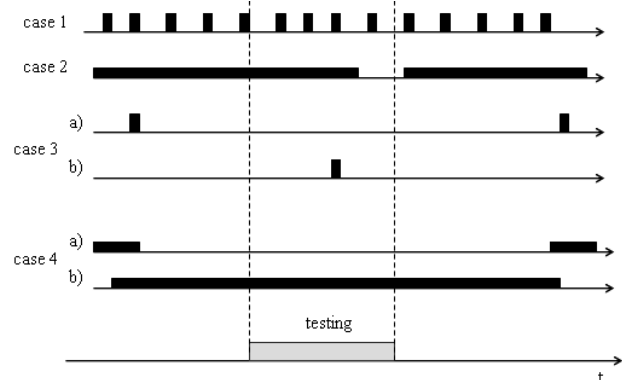


Fig. 3. Different cases of intermittent faults in relation to the testing procedure

As a result of this consideration, it is possible to divide the considered intermittent faults into several classes.

Intermittent faults related to class 1 and class 2 can be detected with high probability during testing procedure. Detection of intermittent faults in case 3 is very improbable (problematic). Probability of detecting such faults is low. As concerns case 4, there exist two options – a) and b) (see Fig. 3). In the given case, probability of intermittent fault detection can be estimated as 0.5 when $1/\lambda \approx 1/\mu$.

Generally, system level self-diagnosis is aimed at detecting permanently faulty system units. Nevertheless, there exist the possibility to detect also intermittently faulty units when intermittent faults are related to classes 1,2 and 4.

III. PROBLEMS WITH DEVELOPING DIAGNOSIS ALGORITHMS WHEN INTERMITTENT FAULTS ARE ALLOWED

Diagnosis is performed on the basis of obtained syndrome. A syndrome is a set of test results. The result of test τ_{ij} is denoted as r_{ij} and can take the values 0 or 1 depending on the fact of how unit u_i evaluates the state of unit u_j .

In the paper, we accept the evaluation proposed by Preparata [7].

$$r_{ij} = \begin{cases} 0 & \text{if units } u_i \text{ and } u_j \text{ are fault-free} \\ 1 & \text{if units } u_i \text{ is fault-free and } u_j \text{ is faulty} \\ X(0,1) & \text{when unit } u_i \text{ is faulty} \end{cases} \quad (1)$$

We also assume that if an intermittent fault is in active state, then unit with this fault behaves as permanently faulty unit.

To explain the problems with diagnosis made on the basis of obtained syndrome, let us consider a simple example. Let system consists of five units. Tests were performed according to predefined schedule. Obtained syndrome is

$Rd = \{u_{12} = 0, u_{23} = 1, u_{24} = 0, u_{34} = X, u_{35} = X, u_{45} = 1, u_{41} = 0, u_{51} = X, u_{52} = X\}$, where $X \in \{0,1\}$.
The results of individual tests are shown in Fig. 4.

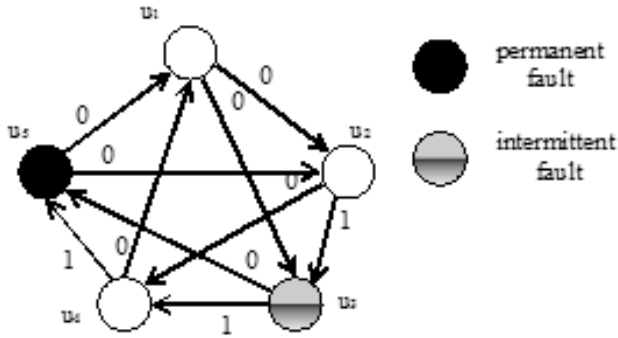


Fig. 4. Example of mutual testing for system with five units

There exist several methods for diagnosis of permanently faulty units [8], [9], [10]. Most of them make the assumption about maximum possible number of faulty units in the system. In [7], there was proved that correct system diagnosis is possible if the total number of faulty units do not exceed the value t , where

$$t = \left\lfloor \frac{N-1}{2} \right\rfloor \quad (2)$$

If we assume, that in the considered example units u_3 and u_5 are permanently faulty, then there should be obtained one of the syndromes that belong to the set R , where

$$R = \{u_{12} = 0, u_{13} = 1, u_{23} = 1, u_{24} = 0, u_{34} = X, u_{35} = X, u_{45} = 1, u_{41} = 0, u_{51} = X, u_{52} = X\}, \text{ where } X \in \{0,1\}.$$

As can be seen, the syndrome Pd does not belong to the set R since any syndrome that belongs to R has r_{13} equal to 1, whereas in the syndrome Rd this result is equal to 0.

It can be explained by the fact that unit u_3 has an intermittent fault which transfers from AS (PS) to PS (AS) during testing procedure. In view of this, most of the algorithms developed for diagnosis of permanently faulty units cannot be directly used for diagnosis intermittently faulty units.

Nevertheless, there was suggested the method [11] based on summary (updated) syndrome, R_Σ . Summary syndrome R_Σ is obtained after performing m rounds of test routine (i.e., m repetition of tests). Summary syndrome R_Σ is computed as

Follows (3):

$$R_\Sigma = \{r_{ij}^*\}, \quad r_{ij}^* = \bigcup_l^l r_{ij}^l, \quad (3)$$

where $r_{ij}^l \in R_l$, R_l - syndrome obtained in l -th round of repetition of test routine.

When summary syndrome R_Σ is a subset of set R_0 (i.e., $R_\Sigma \in R_0$), the algorithms developed for diagnosing permanently faulty units can be also used for considered faulty situation. R_0 is a set of syndromes that can be obtained when only permanently faulty units can take place

in the system. In the given case, the task arises to determine the number of test routine repetitions, l . Neither is determined what should be done if after l repetitions the condition $R_\Sigma \in R_0$ is not true.

To solve these tasks, we suggest the following decision. It is suggested to repeat the test routine several times, l . Concrete number of repetitions of test routine depends on the total number of units in the system N , on the classes of intermittent faults, which are going to be detected, and on the required credibility of diagnosis. If an intermittent fault belongs to class 4, the value of l does not influence the test results. If an intermittent fault belongs to class 3, a unit with such fault behaves either as fault-free or as faulty only during one test. Any next test will show that this unit behaves as fault-free. Thus, two test are sufficient to form r_{ij}^* which make condition $R_\Sigma \in R_0$ true. It also concerns an intermittent fault of class 1. In case 2, a unit with such intermittent fault with high probability will behave as permanently faulty. There is low probability that one of the tests will show that this unit is fault-free. Although, any other test will show that this unit is faulty. Thus, two tests are enough to form R , which satisfies the above mentioned condition.

If after several rounds of tests repetitions the condition $R_\Sigma \in R_0$ is not true, then it is needed to determine a consistent set of units, K_u . Set K_u contains all of the units that, according to the summary syndrome, are diagnosed as fault-free. Units that belong to the set K_u evaluate each other as fault-free. In order to determine the set K_u , it is needed to remove from summary syndrome R_Σ all test results which are equal to 1. Remaining results allow to form a Z-graph.

Z-graph is formed as follows. If r_{ij}^* in R_Σ is equal to 0, then there is an edge between vertices v_i and v_j in Z-graph directed from v_i to v_j .

For Z-graph it is possible to form the matrix M_R . Matrix M_R is square matrix presentation of the subset of R_Σ which has only zero results. If result r_{ij}^* is an element of the resulting subset of R_Σ , then element m_{ij} of matrix M_R has value of 0. Otherwise, element m_{ij} is denoted as dash.

There could be used the diagnosis algorithm presented in [12]. This algorithm is based on the matrix which is similar to matrix M_R and can identify all faulty units (on the condition that total number of faulty units does not exceed the value t). Handling matrices like M_R is also presented in [13], [14]. In the given case, it is needed to calculate the total number of 0 in each column. Then, obtained numbers S_i , $i = 1, \dots, N$, should be compared with value t . If $S_i \geq t$, then unit u_i is diagnosed as fault-free. If condition $S_i \geq t$ is not true for all $i \in \{1, \dots, N\}$ then it is needed to find in Z-graph a simple directed cycle of length $t+1$. Such cycle can be determined from matrix M_R . All units, which are in this cycle, should be identified as fault-free.

Units that are not identified as fault-free are either permanently faulty or have an intermittent fault. It should be

noted that there is low probability of incorrect diagnosis. This probability can be evaluated in relation to different total number of system units, different classes of intermittent faults and different number of test routine repetitions.

IV. CONCLUSIONS

Behavior of intermittent fault can be modeled by continuous Markov chain with two states – Passive (PS) and Active (AS). If an intermittent fault is in PS, unit acts as fault-free. If an intermittent fault is in AS, unit acts as fault (i.e., as if it has a permanent fault). During system level self-diagnosis both permanent and intermittent faults can occur. Each test evaluates the state of a unit either as fault-free or as faulty. In the latter case, it does not discriminate the permanent and intermittent faults. Diagnosis algorithms that deal with the sets of test results can consider (take into account) the testing time and, thus, potentially they could discriminate the permanent and intermittent faults. For this, the testing procedure must be very long. In reality, it is needed to perform the testing procedure in acceptable time for each concrete system. This leads to the situation when it is very difficult to discriminate between permanent and intermittent faults. In view of this, suggested diagnosis algorithm can identify only fault-free system units. All units that are not diagnosed as fault-free should be considered as faulty without further specification.

Suggested diagnosis is developed on the basis of consistent sets of system units. Units that belong to consistent set with high probability can be considered as fault-free. When situation allows to extend the testing time, it is possible to gain higher probability that units of a consistent set are fault-free.

REFERENCES

- [1] V. Mashkov, „Task Allocation among Agents of Restricted Alliance,“ IASTED ISC'2005 conference, Cambridge, MA, USA, pp. 13-18, 2005.
- [2] V. Mashkov, J. Barilla, and P. Simr, „Applying Petri Nets to Modeling of Many-Core Processor Self-Testing when Tests are Performed Randomly,“ Springer, Journal of Testing, pp 25–34, 2013, DOI: 10.1007/s10836-012-5346-8.
- [3] V. A. Mashkov, and O. V. Barabash, „Self-testing of multimodule systems based on optimal check-connection structures,“ Engineering Simulation, vol.13, pp. 479-492, 1996.
- [4] V. A. Mashkov, and O. V. Barabash, „Self-checking of modular systems under random performance of elementary checks,“ Engineering Simulation, vol. 12, pp. 433-445, 1995.
- [5] S. Kamal, and C. V. Page, „Intermittent fault: a model and a detection procedure,“ IEEE Trans. Comput., vol.C-23, no.7, pp.713-719, 1974.
- [6] S. Su, I. Koren, and K. Malaiya, „A continuous-parameter Markov model and detection procedures for intermittent faults,“ IEEE Trans. Comput., vol. C-27, no. 6, pp.567-570,1978.
- [7] T. Preparata, G. Metze, and R. Chien, „On the connection assignment problem of diagnosable system,“ IEEE Trans. on Electronic Computers, vol. EC-16, no. 12, pp. 848-854, 1967.
- [8] G. Sullivan, „A $O(t/\sup 3/+ \text{mod } E \text{ mod })$ fault identification algorithm for diagnosable systems,“ IEEE Trans. Comput., vol. C-37, pp. 388-397, 1988.
- [9] V. Vedeshenkov, „On organization of self-diagnosable digital systems,“ Automation and Computer Engineering, vol. 7, pp. 133-137, 1983.
- [10] H. Fujiwara, and K. Kinoshita, „Some existence theorems for probabilistically diagnosable systems,“ IEEE Trans. on Comp., vol. C-27, no. 4, pp. 297-303, 1981.
- [11] S. Mallela, and G. Masson, „Diagnosable systems for intermittent faults,“ IEEE Trans. Comput., vol.C-27, no.6, pp.560-566. 1978.
- [12] V. Mashkov. Selected problems of system level self-diagnosis. Lviv: Ukrainian Academic Press, 2011.
- [13] S. Babichev, V. Lytvynenko, M. Korobchynskiy, and M. A. Taif, „Objective clustering inductive technology of gene expression sequences features,“ Communications in Computer and Information Science, pp.359-372, 2017.
- [14] S. Babichev, M. A. Taif, V. Lytvynenko, and V. Osypenko, „Critical analysis of gene-expression sequences to create the objective clustering inductive technology,“ IEEE 37th International Conference on Electronics and Nanotechnology, ELNANO, pp. 244-248, 2017

On Intelligent Multiagent Approach to Viral Hepatitis B Epidemic Processes Simulation

Dmytro Chumachenko
Informatics department
National Aerospace University "Kharkiv Aviation Institute"
Kharkiv, Ukraine
dichumachenko@gmail.com

Abstract—Simulation of epidemic processes is actual task, which has high social and economic value. Given research is dedicated to the application of the intellectual multiagent approach to the prediction of the incidence of viral hepatitis B. The structure of the model, agents, environment and rules of agents' interaction has been developed. The multiagent model shows high accuracy.

Keywords—multiagent simulation, epidemic process, viral Hepatitis B, incidence prediction

I. INTRODUCTION

The modern works of many scientists are devoted to the development of intellectual problem-oriented systems and their application to population dynamics. Managing the epidemic process of population dynamics systems is an important part of this area. The most important tool for studying these systems is adequate mathematical models for predicting the spread of the dynamics of the epidemic process. To date, a significant number of such theoretically valid models of population dynamics have been created. They rely on the mathematical apparatus of statistics and probability theory. A common drawback of existing models is the low accuracy of forecasting, as well as its short-term.

One of the main problems of the epidemic process in the systems of population dynamics is the stochastic nature of its behavior. Also, existing models do not take into account the peculiarities of the internal behavior of the population, the assessment of the external environment by objects, the logical behavior of specimens of the population. In analyzing the epidemic process, in contrast to the behavior of population dynamics systems, the researcher is only interested in one "epidemic-recession" cycle, because the further development of dynamics can vary considerably depending on the consequences of epidemic behavior, as well as the external influences taken.

The analysis [1] of existing models and methods for predicting the dynamics of the epidemic process has shown that the most effective is the use of the multiagent approach, it allows to take into account the intellectual components of the simulated system of population dynamics.

Thus, **the aim of the research** is to develop a multiagent model of the epidemic process, which allows the construction of a prognostic morbidity for the selected population.

II. BACKGROUND ON MORBIDITY BY VIRAL HEPATITIS B

Hepatitis B is dangerous viral infection which has global spread. Annually in the world are nearly 50 million people who suffer only acute form of infection. Up to 600 thousand patients with hepatitis B is dying [2]. The prevalence of this virus varies widely in different parts of the world [3]. Ukraine is a country with a middle prevalence.

Hepatitis B can be transmitted only from people infected with the virus, including those with latent infection [4]. The virus can be transmitted through blood transfusions, in violation of the integrity of the skin and mucous membranes (tattooing, acupuncture, use of shared toothbrushes), through sexual contact. In recent years, the increasing significance of sexual transmission of infection. Also, there is a great risk of transmission during pregnancy. Risk groups include people who inject drugs, having disorderly sexual contacts, hemodialysis patients, medical staff, family members of carriers of the virus, patients with chronic diseases of the skin. The virus has an incubation period with a wide range (from 15 to 180 days). Risk of chronic acute hepatitis B inversely proportional to the age at the time of infection: among adults with a normal immune system is not more than 5%, at the age between 1 and 5 years – 30%, for newborns – 90%. For 25% of patients unable to identify the source of infection [5]. Thus, at the present time it is one of the most widespread and dangerous viral infections that cause anxiety for the health of the population and reducing the average life expectancy of people all over the world. Hepatitis B has an exceptionally high infectivity and the frequent occurrence of severe consequences (including death). Also considering that 65–80% of those who infected with hepatitis B virus disease have no external clinical manifestations, the problem remains relevant.

III. DEVELOPMENT OF STRUCTURE OF SIMULATED SYSTEM

The epidemiological model of Viral Hepatitis B is based on the concept of the epidemic process by Gromashevsky [6], according to which the epidemic process exists with the continuous interaction of the three main components - the source of infection, the mechanism of transmission and the susceptible organism.

A. Structure of agents

The most profitable type of agent in the study of epidemic processes is an emotionally-motivated intellectual agent, for the most complete and accurate model of human behavior. Let's consider the agent as a set of properties:

$$a = \langle s, s_t, c, t_a, l \rangle, \quad a \in A, s \in S, c \in C, t_a \in T_a, \quad (1)$$

where s_t is time in state s , A is set of all agents, S is set of different agent's states, C is set of working area's cells, T_a is set of possible agent's types, l is length of life.

The set of agent states is predefined and is constant. Depending on the process being studied, the set can be supplemented by different states, the initial set is:

$$S = \{Susceptible, Exposed, Infected, Convalescent, Recovered, Dead\}. \quad (2)$$

The use of such a set of states is based on the idea of distributing the entire population to subsets, based on their states by epidemic features (the classical model of the SIR type [7]). The proposed set characterizes the model as an analogue of the extended model of the SEIRS type [8].

Fig. 1 shows the transitions between agent states:

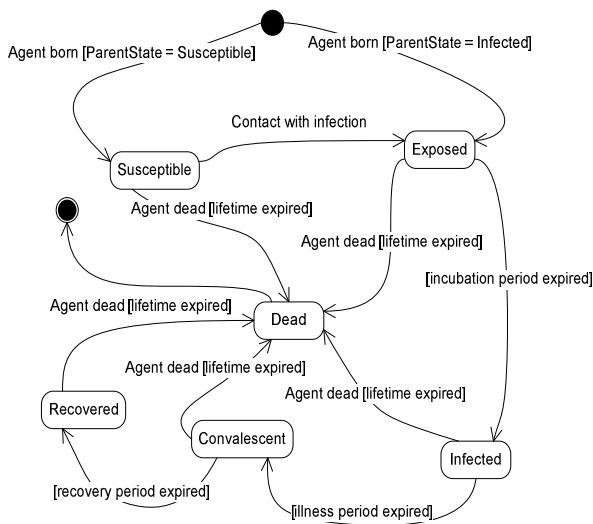


Fig. 1. Agents' states changing

Susceptible – the agent is healthy (may be prone to infection). In this case, a healthy means an agent that is susceptible to the disease of the epidemic process, is modeled.

Exposed – the agent got the disease. This state is an incubation period. During this time, the agent is not yet sick, but already has a chance to transmit the infection.

Infected – agent of the patient. Agents in this condition are the most likely sources of infection for other agents.

Convalescent – the agent recovers. This condition characterizes the period when the clinical symptoms of the disease disappeared, but the agent can still be a carrier of this disease and a source of infection. The presence of this condition is characteristic of certain infectious diseases.

Recovered – the agent recovered (acquired immunity). Agents in this state are no longer subject to the possibility of getting sick.

Dead - the agent is dead from old age or due to illness.

For agents, two types were identified, which can be provisionally called "prudent" and "risky." The characteristics defined for each type are shown in Table I.

TABLE I. DESCRIPTION OF AGENTS' TYPES

Event	Prudent	Risky
Probability of born	80,0 %	20,0 %
Probability of transition to risk zone in susceptible state	1,0 %	10,0 %
Probability of transition to hospital in susceptible state	0,7 %	0,1 %
Probability of transition to risk zone in infected state	0,5 %	5,0 %
Probability of transition to hospital in infected zone	80,0 %	25,0 %
Length of staying in home zone	20 h	15 h
Length of staying in risk zone	2 h	8 h
Length of staying in Hospital	2 h	1 h

Taking into account the modeling features of the epidemic process described above, an internal structure of agents has been developed, which includes the following fields (Fig. 2):

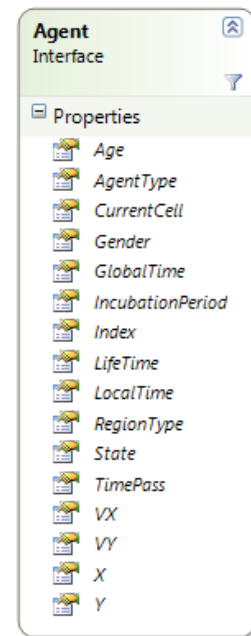


Fig. 2. Internal structure of agent

1. *The index number (Index)*. It is intended for the exact determination of the agent among others.

2. *Agent's local time (LocalTime)*. It is designed to regulate the order of activity of agents, as well as to determine the need for agent processing by the system.

3. *Global Agent Time (GlobalTime)*. Contains the time of the model world, relevant for the agent.

4. *Coordinates of the agent (X, Y)*. Designed to determine the physical location in the model world.

5. *Direction of movement (VX, VY)*. Contains information about the physical direction of the agent in the modeling world.

6. *State of the agent (State)*. Contains the state in which the agent is currently located.

7. *Current cell (CurrentCell)*. Contains characteristics inherent in the current location of the agent.

8. *Current Region (RegionType)*. Lets you get information about the area in which the agent is currently located.

9. *Time spent in the current area (TimePass)*. Contains the moment when the agent has stayed in the current area. At this point in time, the agent moves to another area.

10. *Agent type (AgentType)*.

11. *Age of the agent (Age)*. It is a value that reflects the model age of the agent.

12. *Sex of the agent (Gender)*.

13. *Incubation Period (IncubationPeriod)*. Contains the end of the incubation period. Upon reaching this point in time, the agent becomes ill.

14. *Lifetime (LifeTime)*. Contains the amount of model time assigned to the agent for life. For each agent this value is different. It is also assumed that this value may change under certain conditions (for example, after the transferred illness).

B. Environment

The composition of the workspace leads to the appearance of a set of cells, as conditional abstract objects. It is assumed that one cell can simultaneously include a number of agents, as well as one object-vector of infection (let's call it an instrument). Therefore, the cell can be described as follows:

$$c = \langle z, \tilde{A}, i \rangle, z \in Z, \tilde{A} \subset A, i \in I, \quad (3)$$

where I is set of all instruments,

Z is set of all working areas.

The working area can be described as follows:

$$z = \langle \tilde{C}, t_z \rangle, \tilde{C} \subset C, t_z \in T_z, \quad (4)$$

where T_z is set of possible areas types. It is assumed that, depending on the type of area in which the agent is located, the specific nature of the epidemic process changes.

There were 3 working areas defined for experiments:

- *Home area*. In this area only contacts between agents are allowed.
- *Risk zone*. In this area, in addition to contacts between agents, infection from an infected tool is possible.
- *Hospital*. In this area, partial contact between agents and infection from the instrument is possible. When ingested, the agent is treatable, which reduces the duration of the disease.

Under the tool in this model, there are various objects that can somehow transfer the infection from one person to

another (syringes, scissors, etc.). They can be subjected to various degrees of decontamination. In order to avoid excessive complication of the model, the features described above are reduced to the tool life expectancy. Thus, the tool can be described as a set:

$$i = \langle s, c, l \rangle \quad (5)$$

Thus, the simulation model can be represented as a function:

$$\text{sim}(l_{\text{mean}}, |A|_0, \bar{P}, T), \quad (6)$$

where l_{mean} is average life duration,

$|A|_0$ is the initial power of set of agents,

\bar{P} is the vector of probabilities that are taken into account in the simulation model.

C. Interaction of agents

At its core, intelligent multiagent technologies contain elements of a discrete-event approach [9]. This is manifested in the fact that the system has a timeline for the simulation process. The peculiarity is that on this scale, according to certain rules, on the basis of the general situation in the system and the individual characteristics of individuals, there are events of agents that occur and are processed by the system upon reaching the necessary time moment. Among the events are the events-the intersection of the boundaries of cells forming a stream of events of the first kind. Events of this type are represented by an increasing sequence of instants of time, processed as a transition from one cell to another. Also in the system there are events of interaction with other agents, pulling the branching of the results of the event depending on the individual properties of agents interacting with each other. This creates a stream of events of the second kind. The processing of both types of events and the generation of the following for each agent is a difficult task that involves some technical problems, from the decision of which the adequacy and expediency of using the model directly depends.

If we try to consider the interaction of agents as direct physical contact, then when using the event approach, the agent interactions will be events of the second kind [10]. Adding to the second-order event processing system results in a significant reduction in the simulation speed [11]. A similar situation would be acceptable in the case of modeling physical processes. When modeling a phenomenon such as the epidemic process, it is obvious that participants in the process can interact with each other not only through direct physical contact (for example, airborne diseases) [12]. The processing of such events is rather complicated, which leads to a slowing down of the modeling process [13]. In this paper, it is proposed to simplify the consideration of the moment of infection by establishing the belonging of agents to one cell. This allows you to take into account the possibility of interaction of agents and at the same time significantly reduce the loss of productivity.

The program structure of the cell provides a field containing a list of all agents currently in the cell. To simplify the modeling process, the hit in a single cell of

multiple agents is considered to be their interaction. Interaction of agents is processed by the modeling environment, depending on the area to which the cell belongs, as well as the types of agents interacting.

The contact transmission of the disease from agent to agent is realized as follows. To begin with, the possibility of contact infection is checked. It is believed that this is possible if other agents are also processed in the same cell along with the current agent, processed, as well. Each pair of agents (formed from the current agent, is processed with other agents in the current cell) is compared. If at least one of the agents is the carrier of the disease, it is believed that with a certain probability, there is a sufficient contact between the agents for the infection and a healthy agent becomes infected.

Interaction between agents is handled by the modeling environment, depending on the disease, and is modeled. When considering certain epidemic processes, the logic of handling the interaction of agents can be complicated.

In order to increase the population's detail in terms of its membership in conditional social groups, the types of agents were identified, which are presented as follows:

$$t = \langle \bar{P}_t, \bar{\tau}_t \rangle, \quad (7)$$

where \bar{P}_t is the vector of probabilities that are characteristic for the type of agent, $\bar{\tau}_t$ is the vector of constants of simulation time.

The constructed model admits an extension associated with an increase in the dimensionality of the vectors. In the framework of this task we confine ourselves to the following set:

$$\bar{P}_t = (P_b, P_{hr}, P_{hh}, P_{sr}, P_{sh}), \quad (8)$$

where P_b is probability of agents with a given type born, P_{hr} is probability of transition to risk zone of agent in susceptible state, P_{hh} is probability of transition to Hospital of agent in susceptible state, P_{sr} probability of transition to risk zone of agent in infected state, P_{sh} probability of transition to Hospital of agent in infected state.

$$\bar{\tau}_t = (\tau_h, \tau_r, \tau_m), \quad (9)$$

where τ_h is the duration of the agent with current type spending in home area, τ_r is the duration of the agent with current type spending in risk zone, τ_m is the duration of the agent with current type spending in Hospital.

For interaction between agents, the probability vector \bar{P} :

$$\bar{P} = (P_i, P_r, P_d, P_s, P_a), \quad (10)$$

where P_i is probability of transmission of infection from a sick agent or using an infected tool, P_r is the probability of transmission of infection from the agent is at the stage of the incubation period or at the stage of the decline of the disease

(let's call it "reduced probability of infection"), P_d is the probability that the agent in the hospital will be diagnosed with a disease that is at the stage of the incubation period, P_s is the probability that in the interaction of the two agents will come into contact, P_a is the probability of accidental transmission of infection to a healthy agent from a carrier agent.

In addition, each type causes the agent to have a length of stay in each of the areas. At the end of the stay in a particular area, the agent passes to a different area according to a given probability. The transition is based on the state (CurrentState) and in which area (CurrentRegion) the agent was located, as well as under the influence of the transition probabilities determined by its type.

IV. EXPERIMENTAL INVESTIGATION OF INTELLIGENT MULTIAGENT APPROACH TO SIMULATION OF VIRAL HEPATITIS B EPIDEMIC PROCESS

It is natural to check the model on real statistics about the epidemic process. For the experiments, the incidence of viral hepatitis B and its statistical data for the Kharkiv region from 1994 to 2016 were chosen. Data for conducting experimental studies are provided by the State Institution "Kharkiv Regional Laboratory Center of the Ministry of Health of Ukraine", which monitor the epidemic process and develop and implement preventive and anti-epidemic measures to reduce the incidence rate. The statistical data of the epidemic process of viral hepatitis B included the monthly incidence in the Kharkiv region from 1994 to 2016. The incidence from 1994 to 2009 was divided into the following age groups: children under 2 years, children from 3 to 6 years, children from 7 to 14 years, adolescents from 14 to 18 years, adults from 19 to 30 years, adults from 31 to 40 years, adults from 41 to 50 years, adults from 51 to 60 years, adults over 61 years. Also from 1994 to 2009, the data included the expected pathways of infection: natural, sexual, domestic, vertical, through medical interventions, through drug injections, through manipulations outside treatment and prevention facilities. In turn, data on the parenteral route of infection through medical interventions are divided into injections, blood transfusions, surgical manipulations, gynecological manipulations, blood sampling for laboratory analysis and dental care. On the other hand, all the data were divided into intensive indicators, that is, the real number of patients, and absolute indicators, that is, the incidence of 100,000 people. From 2009 to 2016, statistical data were provided without dividing by age groups and the expected pathways of infection. This is due to changes in the standards of the Ministry of Health of Ukraine on the collection of data on morbidity. Taking into account the structure of the obtained data and the specificity of the spread of the epidemic process, the developed intellectual multiagent model was tuned. Age and social groups were included, it was added that the agents are adults and the interaction does not occur in the "Hospital" area, because the transmission between people is most often performed sexually. To model the epidemic process of viral hepatitis B, the following structure of the population of agents was chosen, the transitions between states are shown in Fig. 3.

The accuracy of the forecast was compared with the statistics on the incidence of viral hepatitis B in the Kharkov region in absolute (Fig. 4) and intensive (Fig. 5) indicators. The abscissa is the year, the ordinate is the indicator. The red

solid line – statistical data, blue intermittent – is calculated by the modeled forecast.

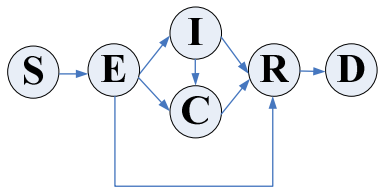


Fig. 3. Transition between states in simulation of Viral Hepatitis B epidemic process

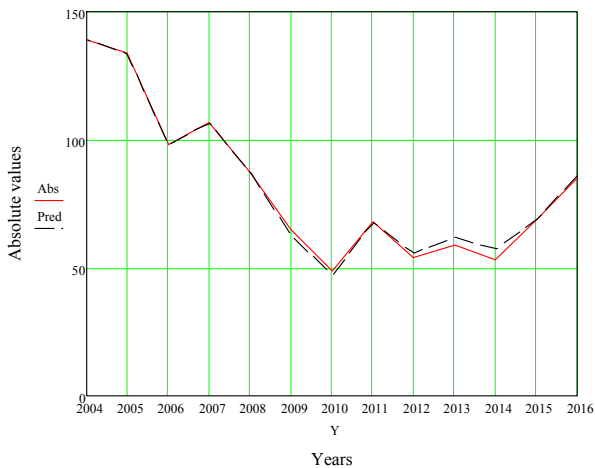


Fig. 4. Comparison of absolute indicators for the incidence of viral hepatitis B in the Kharkov region and the constructed forecast.

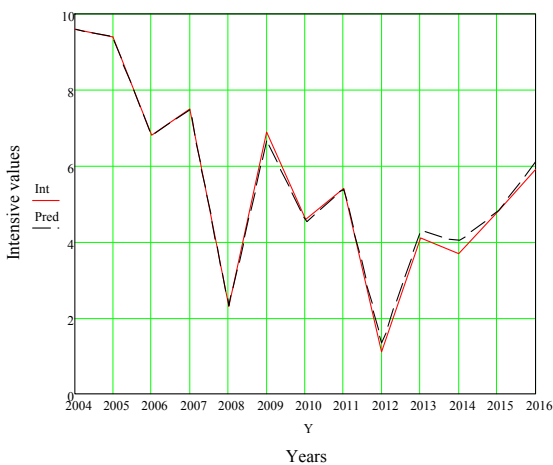


Fig. 5. Comparison of intensive values for the incidence of viral hepatitis B in the Kharkov region and the constructed forecast.

The accuracy of the predicted intensive incidence of viral hepatitis B, calculated using the developed intellectual multiagent model, is shown in Table II.

TABLE II. THE ACCURACY OF THE DEVELOPED MULTIAGENT MODEL

Average absolute error	2,94 %
Root-mean-square error	0,017
Average error	-0,54
Standard error	0,12
Mean deviation	1,95 %

Thus, the accuracy of the constructed forecast is 97.06%.

CONCLUSIONS

A universal intellectual multiagent model of the dynamics of the spread of epidemic processes in population dynamics systems has been developed. The model makes it possible to take into account the types and structure of the population, the features of the spread of the epidemic process are modeled. The described generalized model realizes all the features of the multi-agent approach and is the most universal and susceptible to the type of epidemic process. The advantage of this model construction is the ease in improving and complicating the structure of the simulated system. Without changing already defined modeling patterns, you can add new control parameters, significantly complicate and expand the scope of the subject, and move to higher or lower levels of abstraction. Experimental studies of the universal intellectual multiagent system of epidemic processes have been carried out. The model setting is shown on the example of the incidence of viral hepatitis B, which takes into account the specificity of the epidemic process of viral hepatitis B, the division into age and social groups, the pathways of infection. Calculated with the help of the constructed model, the prognosis was checked on the statistical data on the incidence of viral hepatitis B in the Kharkov region. The accuracy of the forecast is 97.06%.

REFERENCES

- [1] D. Chumachenko, K. Bazilevych and E. Menailov, "Computer simulation of population dynamics epidemic processes," in Mathematical modeling, optimization, information technologies in technical and social-economic systems, 2018, (in press).
- [2] C. Trepo, "Hepatitis B virus infection," in The Lancet, vol. 384, iss. 9959, pp. 2053–2063, 2014.
- [3] L. Rivino, et al, "Hepatitis B virus-specific T cells associate with viral control upon nucleos(t)ide-analogue therapy discontinuation," Journal of Clinical Investigation, vol. 128(2), pp. 668–681, 2018.
- [4] C. Pramoosinsup, "Management of viral hepatitis B," Journal of Gastroenterology and Hepatology, vol. 17, iss. S1, pp. 125–145, 2002.
- [5] J. MacLachlan and B. Cowie, "Hepatitis B Virus Epidemiology," in Cold Spring Harbor Perspectives in Medicine, vol. 5, pp. 1–13, 2015.
- [6] L. Gromashevsky, General epidemiology, 4th ed., Moscow: Medicine, 1965.
- [7] F. Debarre, "SIR models of epidemics" in Modelling course in population and evolutionary biology, Zurich: Institute of Integrative Biology, 2010.
- [8] L. Zhang, L. Yingqiu, R. Qingqing and H. Zhenxiang, "Global dynamics of an SEIRS epidemic model with constant immigration and immunity," in WSEAS transactions on mathematics, vol. 12, iss. 5, pp. 630–640, 2013.
- [9] D. Chumachenko, and S. Yakovlev, "Investigation of agent-based simulation of malicious software," in Econtechmod: an international quarterly journal, vol. 5, iss. 4, Poland: Lublin-Rzeszow, pp. 61–67, 2016.
- [10] D. Chumachenko, and T. Chumachenko, "Agent-based model of the epidemic process of diseases with multiple routes of infection transmission development and evaluation," International Journal of Research Studies in Computer Science and Engineering, vol. 3, iss. 5, India: ARC publications private limited, pp. 61–67, 2016.
- [11] A. Chopra, and P. Munindar, "Agent communication," in Multiagent systems, MIT Press, pp. 9–33, 2011.
- [12] D. Chumachenko, V. Dobriak, M. Mazorchuk, I. Menailov, and K. Bazilevych, "On agent-based approach to influenza and acute respiratory virus infection simulation," 14th International Conference on Advanced Trends in Radioelectronics, Telecommunications and Computer Engineering (TCSET), Lviv-Slavske, Ukraine, pp. 184188, February 20 – 24, 2018.
- [13] S. Russel, P. Norwig, Artificial intelligence. A modern approach. New Jersey: Prentice-Hall, 2003.

Dactyl Alphabet Modeling and Recognition Using Cross Platform Software

Sergii Kondratiuk

*Dept. of Theoretical Cybernetics, faculty of CS and Cybernetics
Taras Shevchenko National University
Kyiv, Ukraine
kondratiuk@univ.net.ua*

Iurii Krak

*Dept. of Theoretical Cybernetics, faculty of CS and Cybernetics
Taras Shevchenko National University
Kyiv, Ukraine
kondratiuk@univ.net.ua, krak@univ.net.ua*

Abstract—The technology, which is implemented with cross platform tools, is proposed for modeling of gesture units of sign language, animation between states of gesture units with a combination of gestures (words). Implemented technology simulates sequence of gestures using virtual spatial hand model and performs recognition of dactyl items from camera input. With the cross platform means technology achieves the ability to run on multiple platforms without re-implementing for each platform.

Keywords— *cross platform; sing language; dactyl modeling; gesture recognition*

I. INTRODUCTION

Communication via gestures is one of the three main means of transmission of information between people, among character (text) and voice (speech) communication. Sign language is usually used by people with hearing disabilities to communicate with each other and with their environment, increasing the number of people who need to know sign language. Note that sign language is universal in the sense that you can send text information via gestures and in case if certain words don't have corresponding gestures (names, cities, areas, etc.), they may be shown via letters one-by-one using dactyl alphabet. Modern hardware is able to collect information fast and almost without restriction, process data both in cloud computing (model, which provides a universal, easy access on demand through the network to the virtual cluster computing resources) [1] and locally on the device, and through data channel processing results are returned to the user. All this is also true for sign language. Signs can be stored and reproduced via a variety of devices and platforms, stationary or mobile, high performance or energy efficient. The actual problem is the reproduction of sign language on all these platforms, for further usage by people with hearing disabilities in particular and everyone in general. Deployment of a single unified technology on various platforms (android, ios, windows, linux, web) without need to port it or to implement it under each platform is a major problem. One way of solving the stated problem of visualization and reproduction of sign language is cross platform software development. Unlike single-platform technologies that operate only on a specific platform under which they were developed, "cross platform software provides the ability to perform on more than one platform with identical (or nearly identical) functionality" [2]. The term "platform" in this context may refer to one of or a combination of several definitions: 1) the type of operating system (such as Microsoft Windows, Mac OS X, Linux, Solaris, Android, iOS); 2) processor type (such as x86, PowerPC, ARM); 3) the type of hardware (e.g., mainframe, workstation, personal computer, mobile device)

[2]. Cross-platform technologies are on a par with the platform independent technologies (those that can operate on any platform, such as Web application) [2] and cross-platform virtual machines (technologies that support individual processes or systems, depending on the level of abstraction at which is virtualization) [3]. In this article the proposed solution of the problem is via cross platform development, taking into account characteristics of different classes of devices (such as hardware, CPU power, amount of memory, presence on the Internet) and setting the number of polygons of the three-dimensional hand model and gesture animation step. Gesture modeling and gesture recognition is performed via cross platform means as a part of proposed communication technology.

II. MATERIALS AND METHODS

Sign modeling is a problem that is considered both independently and as part of the problem of modeling and recognition of gestures and thus as a technology learning and evaluating sign language. One of the systems to display the sign language is American Sing Language Online Dictionary [4], which consists of a video database of words and phrases displayed via sign language. Control via gestures is an actual problem in the development of platform independent human-computer interaction [5]. These developments were involved in a number of commercial agencies [6, 7], but the systems they propose are configured to pre-determined number of gestures, and therefore do not solve the problem of modeling sign language. Also all of them lack functionality of gestures recognitions, thus not allowing to evaluate quality of sign language performed by a user. Creating a model hand is the first step in the task of sign language modeling. In their work [8], authors analyze existing approaches of hand modeling, which are divided into two main groups: spatial and temporal. Former consider the characteristics of different positions for the hand gestures, while latter refer to the description of the dynamics of gestures. Modeling hands in the spatial area can be completed in two and three dimensions. In [9] proposed system by authors is able to simulate sign animations for a given text. As a part of this system a statistical model is used to analyze input text and generative algorithm is used when creating the appropriate simulated kinematics of sign animations. Within the article, the authors have provided ANVIL tools for input text annotation, gesture generator NOVA, and DANCE library developed in [10] is used for gesture animation. The system is built on the Microsoft Windows platform and x86 processor. In [11] authors discuss the modeling of virtual character for spatial reproduction of sign language on the platform of Microsoft Windows. In [12], authors developed a system of signed language training, which consists of two

modules - gesture demonstration module via video and gesture recognition module (required gloves), based on Hidden Markov Model. The training system is based on Microsoft Windows platform and x86 processor. Gesture recognition for mobile platforms is developed in [13], but gesture modeling on mobile devices is not performed.

III. FORMULATION OF PROBLEM

The proposed technology should perform modeling of sign units (morphemes [14]) of sign language, and reproduce animation of gestures structures (words, sentences) via state transitions between shown units using spatial virtual model hand. The proposed technology should perform recognition of sign language based on camera input from the device in order to evaluate sign language performed by user. The technology should be a combined solution for learning sign language via gesture modeling and recognition. Technology should solve the problem of running on existing platforms using cross platform development without implementing the functionality for each platform separately. The effectiveness of the proposed approach is shown in building cross platform technology for modeling and recognition of Ukrainian dactyl alphabet.

IV. PROPOSED APPROACH

To address the modeling of sign language and perform animation of sign structures using spatial virtual model hand the cross platform technology based on cross platform framework Unity3D [15] is proposed. Cross platform framework Unity3D is also used for the user interface, both libraries and technology are implemented with programming language C#. Proposed tools can solve the problem of running the technology on multiple existing platforms. The novelty of the proposed technology is that it is cross platform and has customizable level of polygons for three dimensional hand model and animation step for gesture transitions. This allows to run proposed technology without changes on multiple platforms (different types of processors, operating systems and hardware. Advantage of cross platform technology over technologies developed for a single platform is that there is no need to modify or re-implement the functionality already available for other platforms (porting) [2], which speeds up the process of developing and deploying technologies, and increases the number of potential users. The advantage of cross platform technology over cross platform virtual machine emulators is performance speed and absence of necessity to install additional software (software dependencies).

V. INFOLOGICAL Model

The core of the technology is composition of three cross-platform modules as shown in Figure 1: three dimensional hand model (which is implemented with cross platform framework Unity3D), user interface (implemented also with cross platform framework Unity3D) and gesture recognition module (implemented with cross platform framework Tensorflow [16]). Core functionality is implemented with C# and Python and runs on desktop OS (MacOS, Linux, Windows) and on mobile OS (Android, iOS). Hand model module is cross platform and provides hand model representation for gesture recognition module. Hand renderer receives hand model representation and gesture specifications from gesture storage module, and

provides a high-polygon rendered hand model. Gesture learning module and gesture modification module are implemented with cross platform Unity3D, both taking as input results of hand model renderer. Gesture modification module provides updated gesture specifications and transmits them to gesture storage. Gesture recognition module is proposed to be implemented with Tensorflow framework and receives as input hand model, gesture specifications and input from camera.

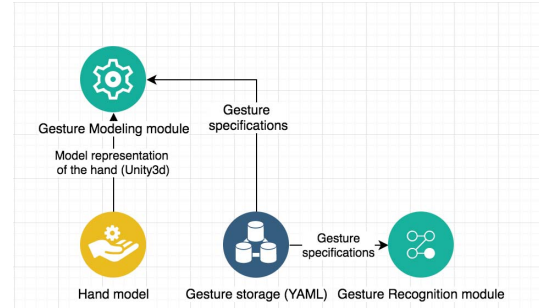


Fig. 1. Infologic model of cross platform gesture communication technology

VI. GESTURE Modeling & Recognition

The hand model which is built in gesture modeling module has 27 bones, 8 of the bones are in wrist, 3 are in the thumb (one metacarpal and 2 phalanx) and 4 metacarpus and 12 phalanges are in other fingers [17]. Each bone is connected to the other through different types of joints. Designing your own cross platform engine for simulating the hand is non-trivial task, thus as the core technology for modeling three-dimensional hand model and gesture animations between morphemes cross platform framework Unity3D was selected.

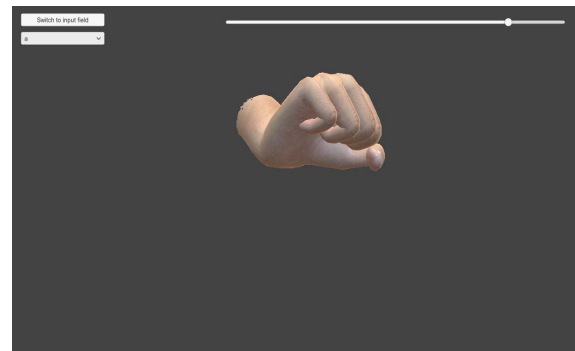


Fig. 2. Gesture modeling under iOS platform.

Unity3D framework is able to effectively reproduce a realistic hand model which consists of more than 70,000 polygons as shown in Fig. 2. Based on the anatomy of the hand within Unity3D hand model was developed with 25 degrees of mobility, four of them located in the metacarpal-carpal joint, to the little finger and thumb to provide movement palm. The thumb has 5 degrees of mobility, middle and index fingers have 4 degrees of mobility (metatarsophalangeal joint with two degrees of mobility, and the distal and proximal interphalangeal joints each have one. Gesture learning and gesture recognition modules, developed with cross platform tools (frameworks based on Python, C++) can be embedded into information and gesture communication cross platform technology. Multiple approaches were considered as an approach for gesture

recognition. Automatic sign language recognition can be approached similarly to speech recognition, with signs being processed similar to phones or words. Most previous work has used approaches based on hidden Markov models (HMMs) [19, 20]. Conventionally, sign language recognition consists of taking an input of video sequences, extracting motion features that reflect sign language linguistic terms, and then using pattern mining techniques or machine learning approaches on the training data. For example, Ong et al. propose a novel method called Sequential Pattern Mining (SPM) that utilizes tree structures to classify signs [21]). Convolutional Neural Networks (CNNs) have shown robust results in image classification and recognition problems, and have been successfully implemented for gesture recognition in recent years. In particular, deep CNNs have been used in researches done in the field of sign language recognition, with input-recognition that utilizes not only pixels of the images. With the use of depth sense cameras, the process is made much easier via developing characteristic depth and motion profiles for each sign language gesture [22]. Multiple existing researches done over various sign languages show that CNNs achieve state-of-the-art accuracy for gesture recognition [23, 24, 25].

Convolutional neural networks have such advantages: no need in hand crafted features of gestures on images; predictive model is able to generalize on users and surrounding not occurring during training; robustness to different scales, lightning conditions and occlusions. Although, selected approach has couple of disadvantages, which may be overcome with a relatively big dataset (1,000 images for each gesture, among more than 10 people of different age, sex, nationality and images taken under different environment conditions and scales): need to collect a rather big and labeled gesture images dataset; black-box approach which is harder to interpret. F1-score of gesture recognition on test dataset of 0.2 fraction of whole dataset is shown at Table 1 (based on number of train samples per class).

TABLE I.
F1-SCORE BASED ON NUMBER OF IMAGE SAMPLES PER CLASS

Image samples	F1-score
100	0.6
200	0.74
500	0.8
1000	0.82

Usage of cross platform neural network framework such as Tensorflow allows to implement gesture recognition as a cross platform module of proposed technology and serve trained recognition model on server or transfer it to the device. Sample dataset chunk shown in Fig. 3.

VII. APPLICATION OF CROSS PLATFORM TOOLS

The hand model which is built in gesture modeling module has 27 bones. Due to selected cross platform implementation tools, the proposed technology solves the problem of executing on multiple platforms without the implementation under each platform separately. Software offered and used in the implementation of information technology is cross-platform and operates unchanged

regardless of operating system (Windows, Linux, Android, iOS), CPU type (x86, arm), and the type of hardware (mobile or stationary device).

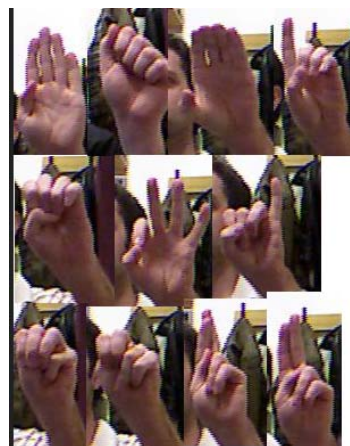


Fig. 3. Sample of dataset

With its cross platform build system Unity3D it is possible to create applications for each platform without porting or changing the original code.

As there are no specific hardware requirements for information technology for modeling sign language, there are objective obstacles for performance speed of older generations devices. To overcome this problem, the following adaptive approach to information technology was proposed as shown on Fig. 4.

Further modules implementation will leverage from existing cross-platform technology. Gesture learning and gesture recognition modules, developed with cross platform technologies (Python, Tensorflow) will be embedded into information and gesture communication cross-platform technology. In case of the mobile app (iOS, Android) or application on the device with a stationary operating system (Windows, Linux), during installation on the device,

information technology analyzes the existing hardware and, depending on its capacity, conducts a series of adjustments: 1) number of polygons of the hand model changes to priority for performance speed; 2) during rotation hand model changes pitch angle at which it rotates, with priority for speed. If the available hardware does not meet the minimum requirements of information technology, the user is given the recommendation to choose “online” mode, in which the calculation is not performed on hardware.

VIII. CONCLUSION

The proposed technology is built with cross platform tools for gesture modeling, gesture transitions animation and gesture recognition. The technology uses virtual spatial model of hand. With the help of cross platform development, the technology solves the problem of execution on the existing multiple platforms without implementing functionality under each platform separately. Thus, it was shown the effectiveness of the technologies built using cross platform tools, for example modeling and recognition elements of dactyl Ukrainian alphabet sign language. Information and gesture communication technology was developed with further scaling capabilities in mind for gestures of other languages alphabets. To implement this idea, the validation mechanism of new gestures to the

common database can be applied. Cross platform information and communication technology and standardized protocol and data format (YAML) allows a range of solutions for remote computing using cloud computing, Web servers, local

servers using a single sign database PostgreSQL [26]. The gesture communication technology can be augmented with other cross platform modules, such as gesture recognition and gesture learning modules.

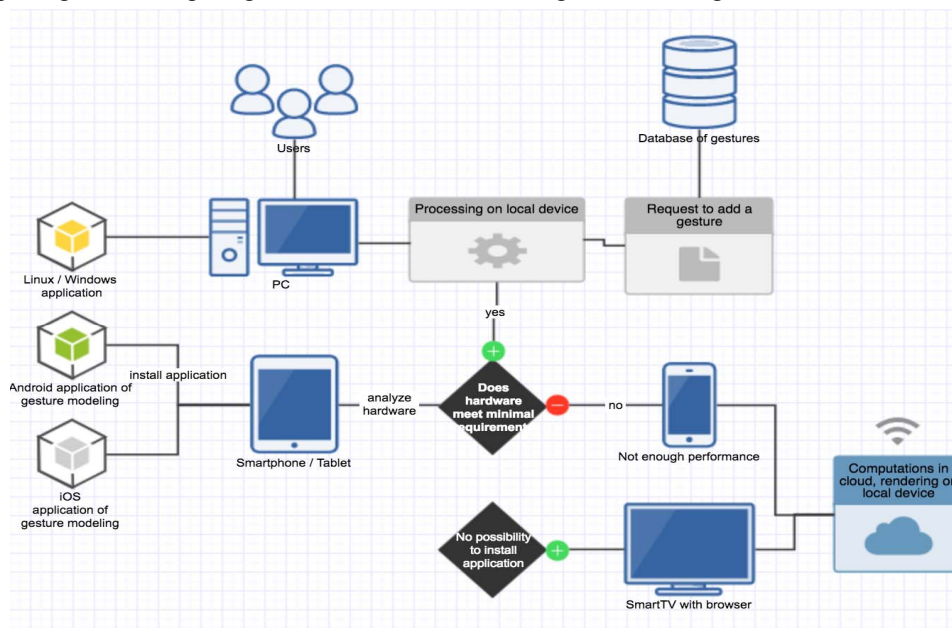


Fig. 4. General scheme of cross-platform and adaptive execution of information technology.

REFERENCES

- [1] P. Mell and T. Grance, "The NIST Definition of Cloud Computing," (Technical report), National Institute of Standards and Technology: U.S. Department of Commerce, pp. 1-7, September 2011. doi:10.6028/NIST.SP.800-145. Special publication 800-145.
- [2] The Linux Information Project, Cross-platform Definition.
- [3] J. Smith, and R. Nair, "The Architecture of Virtual Machines," Computer, vol. 38, no 5, pp. 32–38, 2005.
- [4] ASL Sing language dictionary www.signasl.org/sign/model
- [5] L. A. Graschenko, A. P. Fisun and et., Teoreticheskie i prakticheskie osnovyi cheloveko-kompyuternogo vzaimodeystviya: bazovye ponyatiya cheloveko-kompyuternykh sistem v informatike i informatsionnoy bezopasnosti: Monografiya / Red. A. P. Fisun. Orel: OGU, 2004. Dep. v VINITI 15.10.2004 g. # 1624 — V2004
- [6] Samsung TV Gesture book www.samsung.com /ph/ smarttv/common/guide_book_3p_si/waving.html
- [7] Apple Touchless Gesture System for iDevices http://www.patentlyapple.com/patently-apple/2014/12/apple-invents-a-highly-advanced-air-gesturing-system-for-future-idevices-and-beyond.html
- [8] R. Z. Khan, I. A. Noor, "Comparative study of hand gesture recognition system," Natarajan Meghanathan, et al. (Eds): SIPM, FCST, ITCA, WSE, ACSIT, CS & IT 06, pp. 203–213, 2012.
- [9] M. Neff, M. Kipp, I. Albrecht, and H.-P., "Seidel Gesture Modeling and Animation by Imitation," Technical Report MPI-I-2006-1-005, Max-Planck-Institut Informatik, Saarbrücken, Germany, 2006.
- [10] A. Shapiro, D. Chu, B. Allen, and P. Faloutsos, "A Dynamic Controller Toolkit," Sandbox '07 Proceedings of the 2007 ACM SIGGRAPH symposium on Video games, , San Diego, California, pp. 15-20, August 04 - 05, 2007
- [11] Iu. G. Kryvonos, Yu. V. Krak, Yu. V. Barchukova, and B. A. Trocenko, "Human Hand Motion Parametrization for Dactilemes Modeling," Journal of Automation and Information Sciences, vol. 43, no. 12, pp.1-11, 2011.
- [12] O. Aran, I. Ari, A. Benoit, A. Huerta Carrillo, F.-X. Fanard, P. Campr, L. Akarun, A. Caplier, M. Rombaut and B. Sankur, "Sing language tutoring tool," 2006 eINTERFACE'06, Dubrovnik, Croatia. Final Project Report, July 17 – August 11, pp. 1-11, 2006.
- [13] J.L. Raheja, A.S Sadab and A. Chaudhary, "Android based portable hand sign recognition system," Ed: A. Chaudhary, Recent Trends in Hand Gesture Recognition, GCSR, vol. 3, pp. 1-18, 2015. DOI: 10.15579/gcsr.vol3.ch1,
- [14] W. C. Stokoe, "Sign Language Structure," An Outline of the Visual Communication Systems of the American Deaf., p.61-67, 1960.
- [15] Unity3D framework www.unity3d.com
- [16] Tensorflow framework documentation www.tensorflow.org/api/
- [17] R. Tubiana, J. Thomine, and E. Mackin, Examination of the hand and wrist. 2nd ed. Martin Dunitz, 1996. ISBN: 1853175447/1-85317-544-7. Publisher: Informa Healthcare
- [18] YAML – The Official YAML Web Site www.yaml.org
- [19] O. Koller, J. Forster, and H. Ney, "Continuous sign language recognition: Towards large vocabulary statistical recognition systems handling multiple signers," Computer Vision and Image Understanding, vol. 141, pp. 108–125, 2015.
- [20] P. Dreuw, D. Rybach, T. Deselaers, M. Zahedi, and H. Ney, "Speech recognition techniques for a sign language recognition system," In Interspeech, Antwerp, Belgium, ISCA best student paper award Interspeech 2007, pp. 2513-2516, August 2007.
- [21] Eng-Jon Ong et al. "Sign language recognition using sequential pattern trees," 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2200–2207, 2012.
- [22] A. Agarwal, and M. Thakur, "Sign Language Recognition using Microsoft Kinect," IEEE International Conference on Contemporary Computing, pp. 181-185, 2013.
- [23] L. Pigou, S. Dieleman, P.-J. Kindermans, and B. Schrauwen, "Italian Sign language: Sign Language Recognition Using Convolutional Neural Networks. ELIS, Ghent University, Ghent, Belgium, 2015
- [24] Brandon Garcia, and Sigberto Alarcon Viesca, American Sign language: Real-time American Sign Language Recognition with Convolutional Neural Networks Stanford University Stanford, CA, pp. 225-232, 2015.
- [25] V. Bobic, "Hand gesture recognition using neural network based techniques," School of Electrical Engineering, University of Belgrade, 2016.
- [26] PostgreSQL official web site www.postgresql.org

Intelligent Support for Resource Distribution in Logistic Networks Using Continuous-Domain Genetic Algorithms

Lukasz Wieczorek
Institute of Information Technology
Lodz University of Technology
Łódź, Poland
lukasz.wieczorek.1@edu.p.lodz.pl

Przemysław Ignaciuk, IEEE, Senior Member
Institute of Information Technology
Lodz University of Technology
Łódź, Poland
przemyslaw.ignaciuk@p.lodz.pl

Abstract—The paper addresses the issue of improving the goods distribution efficiency in logistic networks subjected to uncertain demand. The class of networks under consideration encompasses two types of entities – controlled nodes and external sources – forming a mesh interconnection structure. In order to find the optimal operating conditions for the *a priori* unknown, time-varying demand, numerous, computationally involving simulations need to be conducted. In this work, the application of genetic algorithms (GAs) with continuous domain search is proposed to optimize the goods reflow in the network. The objective is to reduce the holding costs while ensuring high customer satisfaction. Using a network state-space model with a centralized inventory management policy, GA automatically adjusts the policy parameters to a given network topology. Extensive tests for different statistical distributions validate the analytical content.

Keywords—inventory management, optimization, genetic algorithms, uncertain demand.

I. INTRODUCTION

Despite the recent financial crisis, overall, the world economy has experienced an increasing growth rate in the last twenty years. New branches of industry and services have emerged and many of the existing ones have expanded even more than throughout the entire previous century [1, 2]. One of such well prospering fields is logistics, which involves planning and managing a cost efficient flow of raw materials, assembly parts, and finished products [3]. Moreover, logistics creates opportunities for the development of other sectors and facilitates their implementation.

Meanwhile, the use of intelligent methods gains in popularity in the applications beyond pure computer science [4–6]. Most companies gather a substantial amount of information from their customers and users. On the other hand, the growth of the related field – Internet of Things – makes new types of sensors and embedded systems commercially available and put into practice [7]. The collected data can be used in many innovative ways, in particular, to obtain more accurate information about the current state of transportation systems, supply chains, and networks. The next step is capturing more viable trends and obtaining predictions of the system future behavior to further boost the performance and alleviate the costs. For instance, the data retrieved from the GPS navigation devices allow one to determine the most appropriate path, or to estimate the

traffic intensity on-the-fly to reduce the congestion and shorten the transfer time to the intended destination. Certain companies, such as Google, or TomTom, in their mobile mapping technology incorporate cameras and lasers mounted on the metering cars to create realistic 3D models of the urban areas for autonomous driving vehicles to be used in future logistic solutions [8].

Nevertheless, the scientific examination of modern distribution networks is a difficult task owing to the analytical intricacies and high mathematical complexity of realistic models. So far, the following types of simplified structures have received the primary attention:

- single-echelon systems [9–11],
- serial multi-echelon chains [12–14],
- star-bus networks [15–17].

With the pace of improvements in the current industry, it is necessary to address the design issues in more complex configurations that would respond well to the growing and more stringent customer expectations. In order to provide high-quality services, various resource management policies [18, 19] and heuristics [20, 21] are being formulated. However, they typically require sophisticated tuning mechanisms to reach a desired efficiency level. In this work, it is shown how to automatize such tuning process with respect to a centralized inventory management policy to be deployed in the logistic networks with arbitrary connectivity structure using genetic algorithms (GAs).

First, the considered strategy is described in the analytical terms. Then, it is adjusted to a given topology using a continuous GA. The optimization process objective is to achieve a balance between the customer satisfaction and the goods holding costs. The customer satisfaction is quantified as the fraction of fulfilled demand requests, whereas the holding costs are related to the number of excess goods stored at the controlled nodes during the distribution process. The implemented continuous-domain GA allows for fast, automatic policy adjustment to the specified requirements. Numerical studies prove the efficiency of GAs in solving optimization issues in non-trivial logistic system configurations.

II. NETWORK MODEL

A. Network Structure and Node Behavior

The structure of the network under consideration comprises N controlled nodes and M resource suppliers. The connections among the nodes form a general, mesh-type topology. The uncertain customer demands are imposed on the controlled nodes at any instant throughout the distribution process execution. The controlled nodes have limited storage capacity and the external resource suppliers are uncapacitated. Although arbitrary connectivity is permitted in the model, there are no separate nodes in the structure (the network is connected), and no node can replenish itself (through a direct loop). The connection between two nodes i and j is characterized by a pair of attributes (DT_{ij}, SC_{ij}) , in which:

- DT_{ij} – is the lead-time delay of a replenishment order coming from node i to j ,
- SC_{ij} – is the contribution of the overall order issued by node i to be obtained from node j .

The operation sequence at a controlled node occurring in each period is illustrated in Fig. 1. At the beginning, the quantity of incoming shipments is registered into the stock. Afterwards, the node processes the customer demand requests and tries to fulfill them. Finally, the node focuses on maintaining the balance throughout the network by satisfying the replenishment orders issued by the other, directly connected nodes.

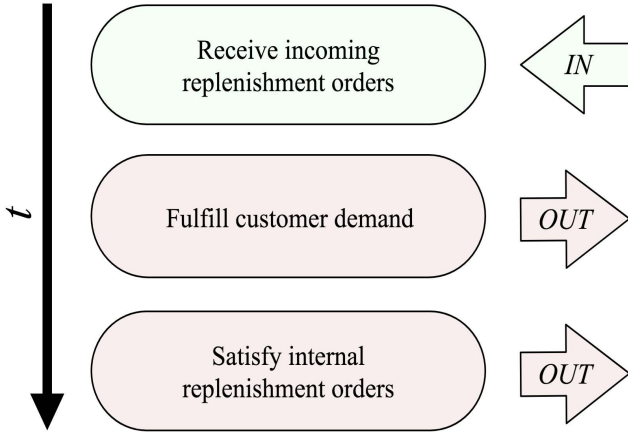


Fig. 1. Node operational sequence.

B. Model of Node Interaction

According to the routine of operations related to handling the flow of goods, the stock balance equation at controlled node i can be expressed through

$$l_i(t+1) = l_i(t) + \Omega_i^I(t) - d_i(t) - \Omega_i^O(t), \quad (1)$$

where:

- $l_i(t)$ – is the on-hand stock level at time t , $t = 0, 1, \dots$
- $\Omega_i^I(t)$ – is the quantity of replenishment orders – incoming shipments – received by node i ,
- $\Omega_i^O(t)$ – is the quantity of replenishment orders sent to the directly connected nodes,

- $d_i(t)$ – is the quantity of satisfied external demands.

According to [22], the incoming replenishment orders from the suppliers of node i can be calculated as

$$\Omega_i^I(t) = \sum_{j=1}^{N+M} \mu_{ji} o_j(t - \gamma_{ji}). \quad (2)$$

Similarly, the amount of shipments sent to the other nodes equals

$$\Omega_i^O(t) = \sum_{s=1}^N \mu_{is} o_s(t - \tau_i^p). \quad (3)$$

In equations (2) and (3):

- μ_{ji} – denotes the supply fraction of the total replenishment order to be acquired from node j ,
- o_i – is the total amount of goods requested from node i ,
- γ_{ji} – represents the lead-time delay at the link between node j and i , $\gamma_{ji} \in \{1, \dots, \Gamma\}$. In detail, $\gamma_{ji} = \tau_j^p + \tau_{ji}$, where τ_j^p is the shipment preparation time at node j and τ_{ji} is the transportation latency from node j to i . All the delays are expressed as non-negative integers.

C. State-Space Description

In order to perform numerical studies of the network behavior in a manageable way, a state-space model will be introduced as vector representation of node interactions. In the adopted framework, the goods distribution process proceeds according to

$$I(t+1) = I(t) + \sum_{\gamma=1}^{\Gamma} \Phi_{\gamma} o(t - \gamma) - d(t), \quad (4)$$

where:

- $I(t)$ – is the vector of on-hand stock levels in period t ,
- $o(t)$ – is the vector of stock replenishment orders in period t ,
- $d(t)$ – is the vector of external demands in period t ,
- Φ_{γ} – denotes the matrix of node interconnections,

$$\Phi_{\gamma} = \begin{bmatrix} \sum_{i \in \Gamma_{i1}=\gamma} \mu_{i1} & x_{12} & \cdots & x_{1N} \\ x_{21} & \sum_{i \in \Gamma_{i2}=\gamma} \mu_{i2} & \cdots & x_{2N} \\ \vdots & \vdots & \ddots & \vdots \\ x_{N1} & x_{N2} & \cdots & \sum_{i \in \Gamma_{iN}=\gamma} \mu_{iN} \end{bmatrix}. \quad (5)$$

The entries on the main diagonal of matrix Φ_{γ} represent the incoming shipments with lead-time delay γ . The off-diagonal entries are set as

$$x_{ik} = \begin{cases} -\mu_{ik}, & \text{if } \tau_i^p = \gamma, \\ 0, & \text{if } \tau_i^p \neq \gamma, \end{cases} \quad (6)$$

where $k \in \{1, \dots, N\}$.

D. Networked Inventory Policy

For efficient performance, the networked inventory policy requires proper selection of target inventory levels (TILs), that should be adjusted for all the controlled nodes with respect to the external demand and network topology. The policy will try to maintain the stock at TIL in each period as the distribution process evolves. The quantity of replenishment orders issued by node i to its suppliers in period t is obtained from

$$o(t) = \Phi^{-1} \left[I^T - I(t) - \sum_{\gamma=1}^{\Gamma} \sum_{k=\gamma}^{\Gamma} \Phi_{\gamma} o(t-k) \right], \quad (7)$$

where:

- I^T – the vector of TILs for the controlled nodes,
- Φ – a gain matrix holding the summary information about the network interconnections and delays,

$$\Phi = \sum_{\gamma=1}^{\Gamma} \Phi_{\gamma}. \quad (8)$$

According to [20], there exists a minimum TIL at the corresponding node beyond which full customer satisfaction is achieved. In order to calculate the TIL vector for the entire network, an estimate of the highest demand is needed. Assuming a persistent demand at each node, such TIL vector may be calculated as

$$I^T = \left(\mathbf{I}_N + \sum_{\gamma=1}^{\Gamma} \gamma \Phi_{\gamma} \right) \Phi^{-1} d_{\max}, \quad (9)$$

where \mathbf{I}_N denotes an identity matrix of size $N \times N$ and d_{\max} is the vector of demand upper estimates.

Reference [22] covers the analytical details of the networked policy operation.

III. OPTIMIZATION PROCESS

The analyzed policy needs to be adapted to a given topology and external factors, in particular the demand. In this paper, the perspectives of using GAs for that purpose are explored. The adaptation process concentrates on the selection of TILs, which the nodes will try to maintain in the course of goods reflow in the network. The multidimensional search space imposes application of a numerical approach to find the optimal TIL vector. The variables on which the network state depends are evaluated through simulations. GA is employed to steer the computations towards the optimal solution. Owing to the continuous search space in the discussed inventory control problem (TILs can assume any value in a given interval), the application of continuous-domain GAs is examined [23]. Since the candidate solutions need not be represented in a binary form, typical for GA applications, in the case analyzed here, it is easier to relate the results directly to the problem variables and speed up computations.

A. Fitness Function

The fitness function allows one to determine how closely a given individual conforms to the problem objectives. It takes a candidate solution as an input argument and returns a number indicating the importance of this individual in the population. One cannot obtain the optimal solution *a priori* based solely on the fitness function equation. This function serves the purpose of comparing the individuals in a particular population and judging which one is better than another. Once the evolution terminates the best candidate solution is selected.

In the considered problem, the objective of the optimization process is to reduce the goods holding costs in the network while maintaining high customer satisfaction. The following fitness function has been chosen to quantify this objective

$$fitness(CS, HC, \varepsilon, \sigma) = CS^{\varepsilon} \left(1 - \frac{HC}{HC_0} \right)^{\sigma}, \quad (10)$$

where:

- CS – is the customer satisfaction level related to the quantity of fulfilled external requests, $CS \in [0, 1]$,
- HC_0 – denotes the fixed initial holding cost established from (9),
- HC – is the holding cost obtained for a given vector of TILs,
- ε – is a coefficient that allows one to numerically emphasize the priority of cost reduction,
- σ – is a coefficient stressing the importance of customer satisfaction.

B. Selection

The selection probability for the recombination purposes is calculated for each individual in the population. It is related to the obtained fitness function value calculated according to (10). The selection process is realized using one of the classical methods – stochastic universal sampling (SUS). Unlike the majority of fundamental methods, e.g., roulette wheel selection, SUS divides a given population into pairs using multiple points. In the considered case, two random selectors in each iteration will determine the pair for the next population. The applied approach allows one to reduce the time of executing the selection operation twice with respect to the methods based on only one selector as the pair of individuals (parents) is chosen in a single iteration. Fig. 2 visualizes the selection operation incorporating the SUS technique. The uniformly distributed random points for selection S_1 and S_2 equal 32% and 83%, respectively.

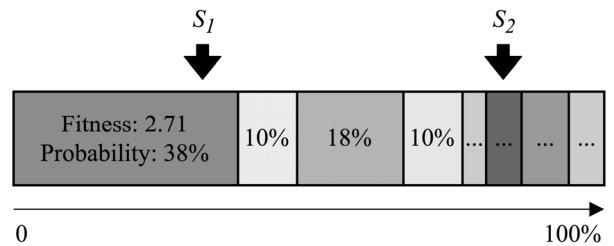


Fig. 2. Stochastic universal sampling.

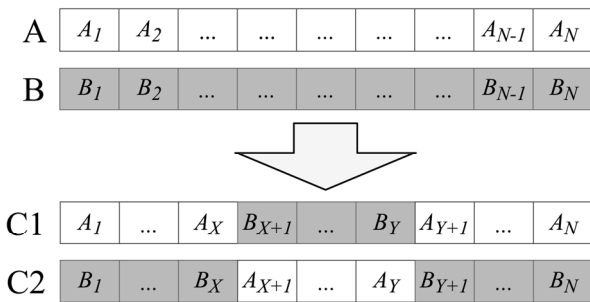


Fig. 3. Multi-point crossover operation.

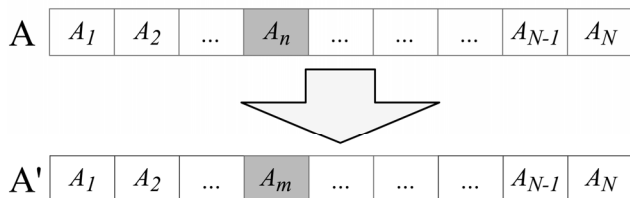


Fig. 4. Uniform mutation operation.

C. Crossover

Recombination is the algorithm step in which each pair of the candidate solutions (parents) creates two new children. As in a biological process, the children inherit traits from their parents. Similarly, in the implemented GA, the TILs in the newly created vectors inherit the traits from the source vectors. In the approach taken in this work, the crossover operation is realized by a multi-point method with two split points. For each pair, two random split points X and Y are determined from the range $[0, N]$. The split points should be different ($X \neq Y$) and are sorted in an ascending order. Once the split points have been established, the crossing operation is performed on the candidate solutions from each pair. Fig. 3 illustrates the crossover operation for individuals A and B using two split points X and Y , where $X < Y$. As a result, two new individuals $C1$ and $C2$ are created.

D. Mutation

The last step of the GA execution as well as the entire evolution process is related to the phenomenon of mutation. This step involves a random gene modification. In the considered continuous-domain problem, the gene modification, i.e., the change of a single element of the TIL vector, can be set only within the boundaries of the search space. Fig. 4 shows the mutation of candidate solution A in which one gene is altered. In biology, mutations occur relatively infrequently. It is assumed here that the mutation probability does not exceed 1–2%.

IV. NUMERICAL TESTS

The properties of considered logistic system and performance of GA are evaluated numerically. The optimization process of adjusting the policy parameters to the given logistic topology is realized using the continuous-domain GA described in section III. In order to support numerical research, a simulation program in the Java language was implemented. It enables one to create logistic topologies, satisfying the connectivity and direction assumptions, and to perform goods distribution simulations.

Fig. 5 depicts the structure of logistic network under consideration. There are two external sources (1–2) and three

controlled nodes (3–5). The pair of attributes above the connection arrows denote the supply contribution and lead-time delay, respectively. In the considered example, node 3 orders 40% of the required goods from node 1 and their delivery takes 2 periods. The goods distribution process is analyzed with two different types of external demand imposed on the network: one based on gamma distribution and another generated using Poisson distribution. Although for illustrative purposes a non-sophisticated topology has been selected, with the granularity of 1 unit, the full search requires exploring the space of over 10^8 combinations.

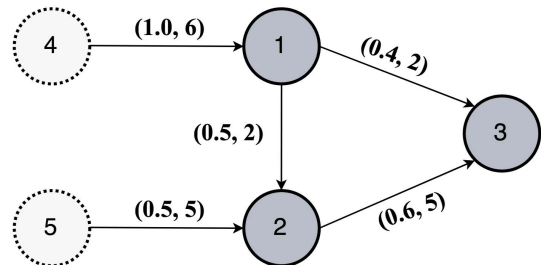


Fig. 5. Network topology.

For GA, the following parameters are assumed:

- the population comprises ten individuals,
- the mutation probability equals 2%,
- the optimization priority coefficients equal $\varepsilon = 40$ and $\sigma = 40$, respectively.

The priority coefficients are chosen so that a significant holding cost reduction is obtained while maintaining near full customer satisfaction. The simulation lasts 50 periods and the initial inventory levels are set equal to the target ones, i.e., $I(0) = I^T$. Two stop criteria: the generation limit of 10^4 and the number of generations without fitness values improvements $3 \cdot 10^3$; are enforced.

A. Results for gamma demand

In the first series of simulations, the external demand is generated using gamma distribution, as frequently applied in the study of inventory control problems [23]. Fig. 6 illustrates the demand requests imposed on the nodes in the investigated network. The distribution is parameterized using two coefficients – shape and scale – which are set equal to 5 and 10, respectively.

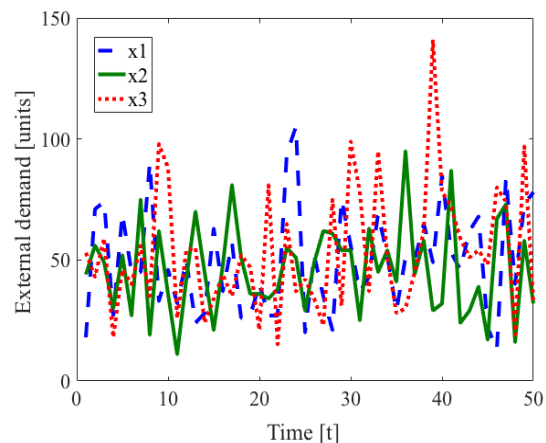


Fig. 6. External demand generated using gamma distribution.

The baseline holding cost for the initial simulation with the external demand fixed to its highest value d_{\max} amounts to $99 \cdot 10^3$. The GA optimization run allows one to keep a full customer satisfaction – 100% fulfilled requests – yet with the costs reduced to $14.5 \cdot 10^3$.

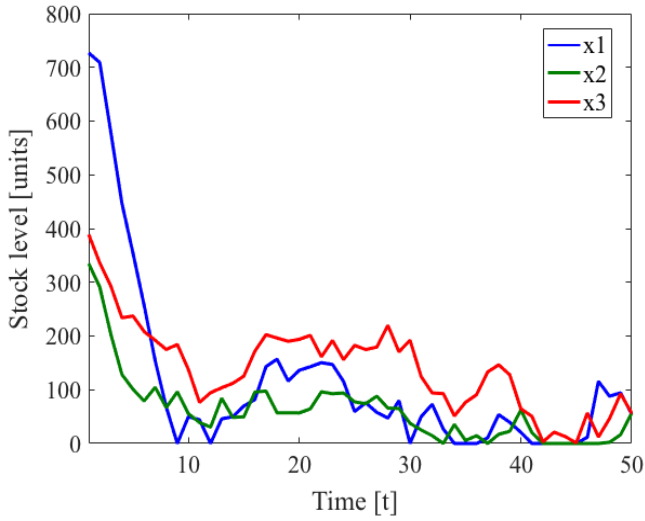


Fig. 7. On-hand stock level after adjusting the networked policy through GA for gamma demand.

Fig. 7 visualizes how the controlled nodes are trying to avoid storing an excessive amount of goods. However, due to the random, *a priori* unknown demand variations (Fig. 6), it is not possible to keep the stock near the zero level all the time.

B. Results for Poisson demand

The values obtained from the Poisson distribution are integers, hence they need not be rounded as required in the study of goods reflow. The distribution generator used in the test was parameterized by $\lambda = 10$ and the resulting evolution is sketched in Fig. 8.

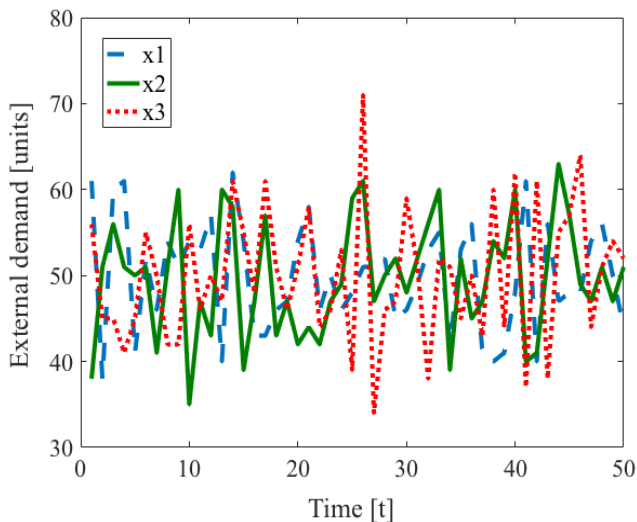


Fig. 8. External demand generated using Poisson distribution.

Similarly, as in the gamma distribution case, the holding cost reduction from $30 \cdot 10^3$ to $4.5 \cdot 10^3$ does not adversely affect the customer satisfaction. Owing to smoother evolution and smaller standard deviation of the assumed

Poisson demand, the holding cost reduction via the action of GA is even more profound than in the gamma distribution case. Fig. 9 depicts how the stock reserves decline for the same initial conditions as in the gamma distribution case.

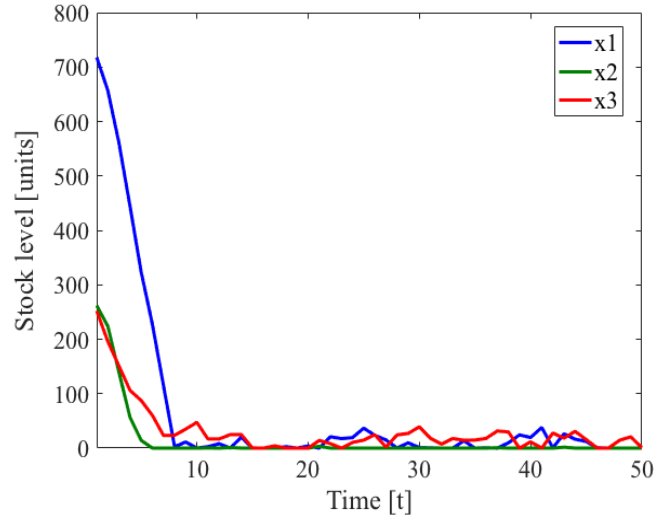


Fig. 9. On-hand stock level after adjusting the networked policy through GA for Poisson demand.

C. Comments

The conducted tests confirm that in both cases IV.A and IV.B, GA successfully adjusts the operation of the examined management policy to the time-varying, uncertain demand. Owing to the smaller standard deviation set for the Poisson distribution the holding cost reduction is more substantial. Fig. 10 presents the fitness function changes in the course of the optimization process. The dashed line maps the best solution established using a full search method. Despite the initially slower convergence rate, GA ultimately steers the network towards the optimal state for less variable demand.

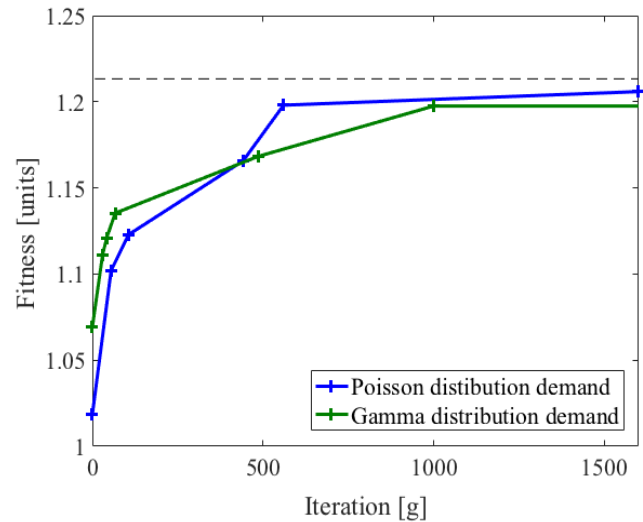


Fig. 10. Progress of the best fitness value improvements.

V. CONCLUSIONS

The paper discusses the application of continuous GAs for optimizing the performance of goods distribution networks with uncertain exogenous demand. The topologies under consideration form a general, mesh-type structure, typically encountered in current logistic systems. The goods distribution process is controlled by a networked inventory policy deployed in a centralized way. Since the external demands are not known *a priori* at the instant of taking the stock replenishment decisions, the policy is adjusted to the given topology and demand type using continuous-domain GA. The implemented GA adapts the target inventory level so that a propitious balance between the holding cost reduction and high customer satisfaction is obtained. The numerical studies, conducted for different network topologies, GA coefficients, and demand distributions, demonstrate the effectiveness of both the considered policy and GAs as tools for artificial intelligence based optimization of modern logistic networks.

REFERENCES

- [1] G. Gereffi and S. Frederick, "The global apparel value chain, trade G. Gereffi and S. Frederick, The global apparel value chain, trade and the crisis: Challenges and opportunities for developing countries. Policy Research Working Papers, no. 5281, 2010.
- [2] T. Berger and C. B. Frey, "Industrial renewal in the 21st century: Evidence from US cities," *Regional Studies*, vol. 50, pp. 1–10, 2015.
- [3] M. Grazia Speranza, "Trends in transportation and logistics," *European Journal of Operational Research*, vol. 264, pp. 830-836, 2018.
- [4] S. Sagioglu and D. Sinanc, "Big data: A review," 2013 International Conference on Collaboration Technologies and Systems (CTS), San Diego, CA, pp. 42-47, 2013.
- [5] M. A. Waller and S. E. Fawcett, "Data science, predictive analytics, and Big data: A revolution that will transform supply chain design and management," *Journal of Business Logistics*, vol. 34, pp. 77-84, 2013.
- [6] G. Wang, A. Gunasekaran, E. W. T. Ngai, and T. Papadopoulos, "Big data analytics in logistics and supply chain management: Certain investigations for research and applications," *International Journal of Production Economics*, vol. 176, pp. 98-110, 2016.
- [7] E. Ahmed, I. Yaqoob, I. A. T. Hashem, I. Khan, A. I. A. Ahmed, M. Imran, and A. V. Vasilakos, "The role of big data analytics in Internet of Things," *Computer Networks*, vol. 129, pp. 459-471, 2017.
- [8] V. Potó, J. Á. Somogyi, T. Lovas, and Á. Barsi, "Laser scanned point clouds to support autonomous vehicles," *Transportation Research Procedia*, vol. 27, pp. 531-537, 2017.
- [9] K. Xu and P. T. Evers, "Managing single echelon inventories through demand aggregation and the feasibility of a correlation matrix," *Computers & Operations Research*, vol. 30, pp. 297-308, 2003.
- [10] P. Ignaciuk and A. Bartoszewicz, "Dead-beat and reaching-law-based sliding-mode control of perishable inventory systems," *Bulletin of the Polish Academy of Sciences-Technical Sciences*, vol. 59, pp. 39-49, 2011.
- [11] P. Ignaciuk and A. Bartoszewicz, "Linear-quadratic optimal control of periodic-review perishable inventory systems," *IEEE Transactions on Control Systems Technology*, vol. 20, pp. 1400-1407, 2012.
- [12] C. A. Garcia, A. Ibeas, and R. Vilanova, "A switched control strategy for inventory control of the supply chain," *Journal of Process Control*, vol. 23, pp. 868-880, 2013.
- [13] H. D. Purnomo, H. M. Wee, and Y. Praharsi, "Two inventory review policies on supply chain configuration problem," *Computers & Industrial Engineering*, vol. 63, pp. 448-455, 2012.
- [14] P. Ignaciuk, "Discrete inventory control in systems with perishable goods – a time-delay system perspective," *IET Control Theory & Applications*, vol. 8, pp. 11-21, 2014.
- [15] C. O. Kim, J. Jun, J. K. Baek, R. L. Smith, and Y. D. Kim, "Adaptive inventory control models for supply chain management," *The International Journal of Advanced Manufacturing Technology*, vol. 26, pp. 1184-1192, 2005.
- [16] P. Ignaciuk, "Nonlinear inventory control with discrete sliding modes in systems with uncertain delay," *IEEE Transactions on Industrial Informatics*, vol. 10, pp. 559-568, 2014.
- [17] L. Sun and Y. Zhou, "A knowledge-based tree-like representation for inventory routing problem in the distribution system of oil products," *Procedia Computer Science*, vol. 112, pp. 1683-1691, 2017.
- [18] J. Poppe, R. J. I. Basten, R. N. Boute, and M. R. Lambrecht, "Numerical study of inventory management under various maintenance policies," *Reliability Engineering & System Safety*, vol. 168, pp. 262-273, 2017.
- [19] P. Garcia-Herreros, A. Agarwal, J. M. Wassick, and I. E. Grossmann, "Optimizing inventory policies in process networks under uncertainty," *Computers & Chemical Engineering*, vol. 92, pp. 256-272, 2016.
- [20] S. Kulkarni, R. Patil, M. Krishnamoorthy, A. Ernst, and A. Ranade, "A new two-stage heuristic for the recreational vehicle scheduling problem," *Computers & Operations Research*, vol. 91, pp. 59-78, 2018.
- [21] P. Ignaciuk and L. Wiecek, "Optimization of mesh-type logistic networks for achieving max service rate under order-up-to inventory policy," *Advances in Intelligent Systems and Computing*, Springer International Publishing, vol. 657, pp. 118-127, 2018.
- [22] P. Ignaciuk, "Dynamic modeling and order-up-to inventory management in logistic networks with positive lead time," 2015 IEEE Int. Conf. Intel. Comp. Com. Proc., Cluj-Napoca, Romania, pp. 507-510, Sep. 2015.
- [23] D. Simon, *Evolutionary Optimization Algorithms*. John Wiley & Sons, 2013.

Segmentation and Parametrization of the Phonocardiogram for the Heart Conditions Classification in Newborns

Ihor Shelevytsky
*Faculty of Information Technologies,
Kryvyi Rih Institute of Economics*
Kriviy Rih, Ukraine
sheleviv@gmail.com

Vlad Golovko
*Faculty of Information Technologies,
Kryvyi Rih Institute of Economics*
Kriviy Rih, Ukraine
golovkovlad@gmail.com

Victorya Shelevytska
*Faculty of Postgraduate Education,
Dnipropetrovsk Medical Academy of Health*
Kriviy Rih, Ukraine
shelevika@gmail.com

Bogdan Semenov
*Institute of Information Diagnostic Systems,
National Aviation University*
Kyiv, Ukraine
SemenovBS@gmail.com

Abstract — Phonocardiographs are analyzed for diagnostics of heart conditions in newborns. The algorithms of allocation of heart tones and selection of stationary periods on phonocardiograms are proposed. Dedicated heartbeats can be parameterized in different ways. The first set of parameters characterizes the shape and the time-amplitude features. The second set of parameters is the coefficients of the frequency-time decomposition of cardiac cycles with spline bases. This approach allows detecting of the Patent ductus arteriosus (PDA) by machine learning methods. Software for phonograms analysis has been developed.

Index Terms — Machine Learning; detection of the Patent ductus arteriosus; algorithms for segmentation of phonocardiograms; parametrization of phonocardiograms; classification of phonocardiograms

I. INTRODUCTION

Echocardiographic ultrasonography is the classical approach to the heart defects diagnosis. However, this is a costly and long-term analysis that cannot cover all newborns. An alternative method is the electronic auscultation with the computer analysis of phonocardiogram (PCG). Studies [1] show that the accuracy of the diagnosis of congenital heart disease (CHD) by this method is more than 95%.

Children are born with the Patent Ductus Arteriosus (PDA), which should be closed by the third day. However Ductus Arteriosus remains open after the third day for the small part of newborns [2]. This case requires detailed echocardiography and observation as it may indicate a heart disease. Auscultation in newborns can be difficult due to the presence of movement sounds, breathing and crying of the baby. So the pre-selection of audio recordings is required to highlight high quality fragments. The basic algorithms for PCG analysis are the allocation of heart tones and heart cycles and the following parametrization for classification using methods of machine learning. In [3] different signal norms are used to allocate tones. However, in practice, the recordings are low quality. Therefore, for acceptable tone allocation, more sophisticated algorithms are required. It is important to take into account a priori information. For analysis and classification of PCG various methods of

frequency-time analysis and parametrization are used [4]. There are two main approaches. The phonocardiogram is analyzed as a normal sound signal. Or phonocardiogram is analyzed by parameters specific for the nature and features of such signals. Determining the parameters with a clear interpretation associated with heart is more interesting for doctors and scientists. This makes it possible to understand the dependence of the sound phenomena on the work of the heart. In the paper both approaches are applied. The set of parameters of tones and intervals between them is determined and the frequency-time decomposition of heart cycles in spline bases is performed. Based on both methods, the classification of phonocardiograms according to patterns is performed to detect Patent Ductus Arteriosus.

II. INPUT DATA

The recording of the phonocardiography is performed sequentially at five standard auscultation points, 5-10 seconds at each. An electronic stethoscope Thinklabs Model ds32a + in sound amplification mode and narrowed listening sector was used for recording. The recording was performed with a Sony-ICD-UX71 digital voice recorder at a frequency of 44.1 kHz in high quality and stored in MP3 format. A typical image of the medium quality phonocardiogram is shown in Fig. 1. The records at the auscultation points are rather heterogeneous and contain sounds of breathing, newborn movements and manipulations with the stethoscope. PCG consists of heart cycles that include the first tone, the interval between the first and second tones, the second tone, the interval between the second and the first tone. First of all for the analysis of PCG it is necessary to allocate tones of the heart. Automatic segmentation of signals at points allows to select fragments of records by intervals between them. However, this is not enough for qualitative analysis. It is difficult to achieve high-quality allocation of tones without intervals and to detect the false ones.

Therefore, the selection of fragments for the analysis of tones depends on the doctor or the operator who performed the recording. The criterion in this case is the visual homogeneity of the recording and the absence of foreign

sounds. The sequence should include more than six cardiac cycles. This is a quite simple operation that requires minimal training and practice.

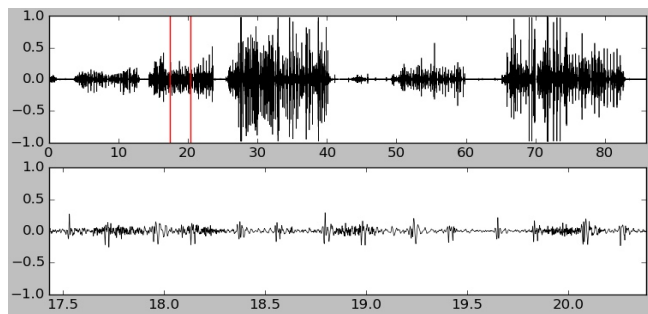


Fig. 1. The typical medium quality neonatal phonocardiogram. Top graph is the recordings at five points. Bottom graph is the selected fragment of the record at the second point.

Therefore, the selection of fragments for the analysis of tones depends on the doctor or the operator who performed the recording. The criterion in this case is the visual homogeneity of the recording and the absence of foreign sounds. The sequence should include more than six cardiac cycles. This is a quite simple operation that requires minimal training and practice.

III. TONE ALLOCATION ALGORITHM

The task of tones allocation requires a compromise between smoothing and sensitivity to the tone boundaries. The task was to achieve subjectively correct allocation of tones for various signal quality. Empirically, the following algorithm for tones allocation was developed.

Algorithm 1: tone seek (x)

1. Decimation of the signal module $|x|$ in 125 times to the sampling frequency of 352.8 Hz.

2. Low-pass filtration of the signal to the frequency of 35.28 Hz.

3. Calculation of the energy of x^2 and normalization with respect to the maximum energy $s = x^2/\max(x^2)$

4. Determination the ranked mean and variance:

to sort $z = \text{sort}(s); \quad m = \text{mean}(z(0.1n:0.9n));$
 $\sigma = \text{std}(z(0.1n:0.9n))$

5. Finding local maxima exceeding the detection threshold: $3m$

6. Finding the boundaries of waves to the right and to the left of the maxima at points crossing the threshold of the lateral boundaries: $m+6\sigma$

7. Analysis of energy waves, for each wave:

if the maxima are within the boundaries of the wave then leave the highest maximum

if the distance between the waves is less than 14 ms then combine them into one wave

if the distance between waves is greater than 60 ms then it is the next wave

8. Wave selection by width:

if the wave is shorter than 17 ms then remove it otherwise leave it as a heart tone

9. Finding the most stable fragment in 5 heart cycles:

for all 5 heart tones windows determine the sum of dispersion of the width of the tones and the intervals between them

define the window with the smallest variance for analysis

10. Determination of the number of the starting tone:

to calculate the average width of even and odd tones

if the average width of even tones is greater than the average width of the odd ones, to shift the original tone to the even one

11. List the tone indexes into the indexes of the initial data (multiply by 125 and take into account the delay of the filtration).

Figure 2 shows an example of the tones allocation with the help of the algorithm, and Fig.3 shows the result of the tones allocation.

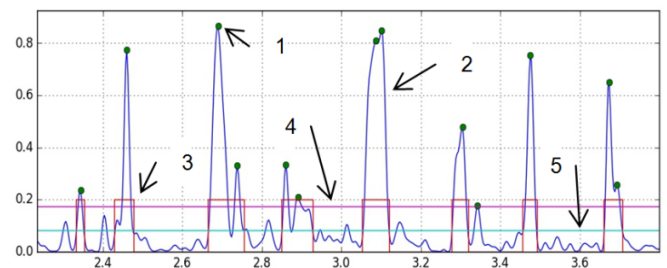


Fig. 2. PCG energy waves and tone allocation scheme. 1 - points of maxima, 2 signal energy, 3 highlighted tones, 4 thresholding, 5 threshold level of side borders of tones.

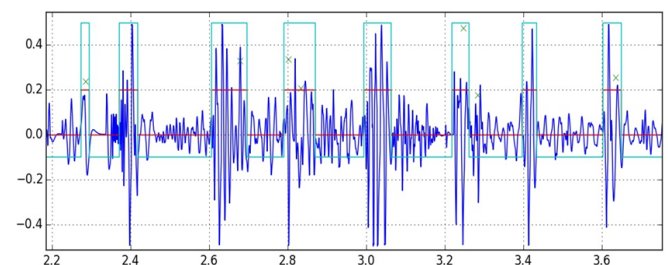


Fig. 3. PCG selected tones along with the signal.

The proposed algorithm confidently highlights the tones in the records of different quality.

IV. PARAMETRIZATION OF TONES AND INTERVALS

The first and the second heart tones have the similar structure, so let us apply the same parameters and recognition algorithms for both of them. Indicators that characterize the shape and features of tones in the time domain are selected as the tone parameters. After the low-frequency filtration by the non-recursive digital filter with a cutoff frequency of 4410 Hz, the search for the special tone points follows. That are maxima, minima, fractures (discontinuity of the first derivative), and zero values. Figure 4 shows the result of the identifying special points.

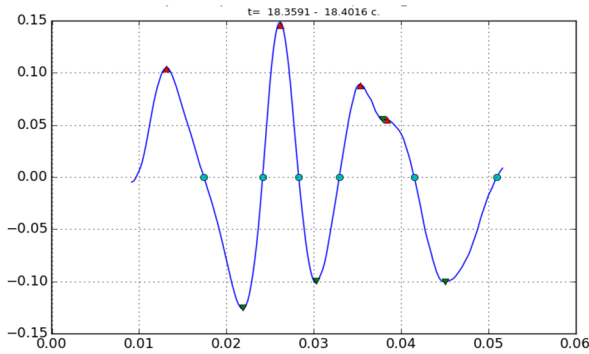


Fig. 4. Heart tones analysis: minima, maxima and zeros

The following tone parameters are calculated based on the selection of the tone and its points.

General parameters: energy – energy of the tone (sum of squares of the signal); width - width of the tone; max_t - position of the maximum of the signal module; max_a - value of the maximum of the signal module; skewnes - the asymmetry of the module maximum (max_t/width)-0.5; n_broken - the number of points in the fracture.

Parameters of the maxima: n_max - number of maxima; t_max - maximum position; a_max - maximum value; mean_t_max - average time position of maxima; std_t_max - standard deviation of the time position of the maxima; mean_dt_max - the average value of the intervals between maxima; std_dt_max - the standard deviation of the distance between the maxima.

Parameters of minima: n_min - number of minima; t_min - minimum position; a_min - minimum value; mean_t_min - average time position of the minima; std_t_min - standard deviation of the time position of the minima; mean_dt_min - the average value of the intervals between the minima; std_dt_min - standard deviation of the distance between the minima.

Parameters of zeros: n_zero - number of zeros; mean_t_zero - mean value of time coordinates of zeros; std_t_zero - standard deviation of time coordinates of zeros.

Signals at intervals between tones are not clear. They have the character of random noise, sometimes with periodic components. Therefore, the parameters for the intervals analyze the shape and intensity of this noise. Therefore the interval is divided into four fragments and the parameters are defined on the interval as a whole and on each of the fragments. Also, the module of the signal values on the interval is approximated by the method of least squares to obtain the coefficients of the polynomial characterizing the shape of the envelope.

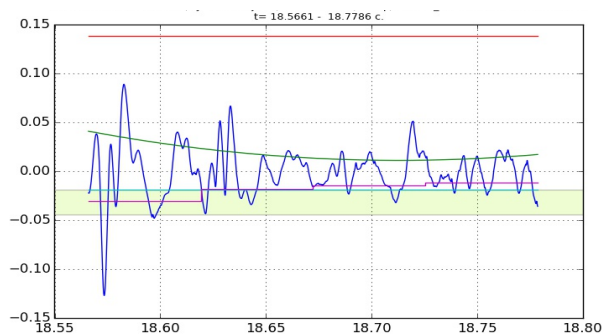


Fig. 5. Analysis of the intervals between the tones.

Parameters of the interval: width - width of the interval; n_zero - number of zeros of the interval; frq_zero - frequency of zeros; energy(m) - signal energy at the interval; energy1, energy2, energy3, energy4 - energy at fragments of the interval; mean - average amplitude; mean_1/4, mean_2/4, mean_3/4, mean_4/4 - average amplitudes at fragments of the interval; std - standard deviation; std_1/4, std_2/4, std_3/4, std_4/4 - standard deviation of the fragments; a2, a1, a0 are the coefficients of the approximation parabola of the amplitude on the interval $a2 * x^2 + a1 * x + a0$.

V. TIME-FREQUENCY ANALYSIS OF HEART CYCLES

Another approach to the analysis of heart cycles is the time-frequency decomposition of the allocated intervals. To perform the decomposition the sampling frequency has been reduced by 25 times to 1764 Hz. Usually for such decomposition the wavelet transform is used. In this paper, application of decomposition in the bases of Hermite splines with the estimation of the coefficients of approximation by the least squares method is proposed. Let's look at the process of the time-frequency decomposition in the operator form. The basic spline consisting of four fragments is $B_{\tau, \vartheta}(t)$, with τ as the offset of basis relative to the initial spline, ϑ as the scale coefficient of the basis relative to the widest one.

The basis can be written as:

$$B_{\tau, \vartheta}(t) = B_0\left(\frac{t - \tau n}{\vartheta}\right), \vartheta \in \mathbb{R}^+, \tau = 0, \pm 1, \pm 2, \dots$$

Let's denote obtained least square estimates of the signal $x(t)$ in the system of basis functions B_{ϑ} as LS. The vector of estimates (decomposition coefficients) in the ϑ scale is:

$$A_{\vartheta} = \text{LS}[x(t), B_{\vartheta}], \quad X_{\vartheta} = \text{IN}[A_{\vartheta}, B_{\vartheta}].$$

Approximation residuals can be denoted as RS, and the process of obtaining the residues of least square approximation as $E_{\vartheta} = \text{RS}[x(t), B_{\vartheta}]$.

In operator form, the decomposition can be written as follows:

$$\begin{aligned} A_{\vartheta_0} &= \text{LS}[x(t), B_{\vartheta_0}], & E_{\vartheta_0} &= \text{RS}[x(t), B_{\vartheta_0}], \\ A_{\vartheta_k} &= \text{LS}[E_{\vartheta_{k-1}}, B_{\vartheta_k}], & E_{\vartheta_k} &= \text{RS}[E_{\vartheta_{k-1}}, B_{\vartheta_k}], \\ k &= \overline{1, K}. \end{aligned}$$

As the result of the decomposition we have the vector of least squares estimates and the vector of approximation residuals at the last stage of the decomposition:

$$\Omega = \{A_{\vartheta_0}, A_{\vartheta_1}, \dots, A_{\vartheta_K}, E\}, \quad \text{де } E = E_{\vartheta_K}.$$

Together with the coefficients of the decomposition we also have the value of the mean square deviations of the coefficients. However, it is difficult to apply the decomposition coefficients directly for the classification task, due to the different duration of cardiac cycles and their components in newborns. This requires the ways to scale up the obtained spectrograms.

VI. RESULTS OF THE ANALYSIS

PDA diagnostic method by the PCG has been developed. In the framework of the HeartTone project, records of phonocardiograms in newborns with further ultrasound heart examination have been performed. For the processing and analysis of phonocardiograms, two versions of the

HeartTone-D program have been developed. The desktop version for detailed research is a tool for researchers. Another, WEB version of the "HeartTone-W" program is designed for the operative work of doctors.

The investigated group consists of 195 healthy newborns. No structural abnormalities in the heart and large vessels were prenatally detected. Newborns were examined from the first to the 5th day of life. The group of interest was newborn babies with an open arterial duct without structural anomalies of the heart. No abnormalities were detected for these children with a traditional (on hearing) auscultation. For the preliminary computer analysis, 27 audiograms of newborn babies with functioning arterial duct and 28 audiograms of newborns with not functioning arterial duct has been selected. In order to diagnose the CHD by the heart phonocardiograms, the binary classification using the support vector machines (SVM) with the Gaussian radial basis function has been used. The methods of machine learning of the scikit-learn package has been used to set up the support vector method. To choose the parameters for classification we rely on the maximum of the t-criteria for the parameters of cycles of phonocardiograms with PDA and without PDA. It turned out that the most statistically significant are the heart tones parameters at the second point of hearing. These are the parameters of the I and II tone - \max_a (maximum amplitude of the I tone), a_{\min} (the minimum amplitude of the II tone), the parameters of the intervals between the tones: $m1_{\text{mean}}$ (average amplitude module of the first interval), $\text{mean}_{4/4}$ (the average amplitude module of the last fragment of the second interval), width (the width of the second interval). The parameters were normalized by bringing the deviations to the range $-1, +1$ from the mean values. The best values of the parameters $\gamma = 8$, $C = 7$ have been found by the grid method. The learning of the algorithm took place on the sample, which included 128 periods of the heart tones at the second point without the PDA and 138 periods with the PDA. To evaluate the accuracy of the method, a nine-fold cross-check test with random sampling of 20 clusters has been used. In summary, the SVC classifier has been obtained with the following statistical characteristics: accuracy (ACC) 87.9% $\sigma = 1.1\%$, sensitivity or true positive rate (TPR) 83.3% $\sigma = 1.6\%$, specificity or true negative rate (TNR): 91.0% $\sigma = 1.3\%$.

The obtained classification results are preliminary, performed for a rough estimation of the possibilities of the proposed method, and require the further refinement on large

samples and comparison with the classification method according to the data of the time-frequency decomposition.

VII. CONCLUSIONS

An algorithm for the allocation of heart tones on low-quality phonocardiograms and a significant variety has been developed. It allows allocating tones for 55 real audiograms in newborns in the automatic mode.

Indicators characterizing the shape of the tones and the intervals between them for interpretation and classification of phonocardiograms have been proposed. Indicators are determined on 5 cardiac cycles at 5 heart tone points.

The time-frequency analysis of heart cycles in spline bases with the estimation of the decomposition coefficients by the method of least squares has been shown. The decomposition coefficients allow visualizing of the spectrogram and reducing the classification problem to the image recognition. However, practical implementation requires experiments with scaling of the decomposition.

The software that implements the described algorithms in Python language for the desktop and WEB has been developed.

The method of detecting the PDA in newborns by parametrization and classification by the method of support reference vectors has been proposed. It shows the accuracy of 87.9% on the given samples.

REFERENCES

- [1] L. S. W. Lai, A. N. Redington, A. J. Reinisch, M. J. Unterberger, and A. J. Schriebl, "Computerized automatic diagnosis of innocent and pathologic murmurs in pediatrics: A pilot study," *Congenital Heart Disease*, vol. 11, no. 5, pp. 386-395, Sep. 2016. [Online]. Available: <http://dx.doi.org/10.1111/chd.12328>
- [2] J. Reese and M. M. Laughon, "The patent ductus arteriosus problem: Infants who still need treatment," *The Journal of Pediatrics*, vol. 167, no. 5, pp. 954-956, Nov. 2015. [Online]. Available: <http://dx.doi.org/10.1016/j.jpeds.2015.08.023>
- [3] V. Nivitha Varghees and K. I. Ramachandran, "Multistage decision-based heart sound delineation method for automated analysis of heart sounds and murmurs," *Healthcare Technology Letters*, vol. 2, no. 6, pp. 156-163, Dec. 2015. [Online]. Available: <http://dx.doi.org/10.1049/htl.2015.0010>
- [4] W. Zhang, J. Han, and S. Deng, "Heart sound classification based on scaled spectrogram and partial least squares regression," *Biomedical Signal Processing and Control*, vol. 32, pp. 20-28, Feb. 2017. [Online]. Available: <http://dx.doi.org/10.1016/j.bspc.2016.10.004>

Subsurface Object Identification by Artificial Neural Networks and Impulse Radiolocation

Oleksandr Dumin
*School of Radio Physics, Biomedical
Electronics and Computer Systems,
V. N. Karazin Kharkiv National University,
Kharkiv, Ukraine
dumin@karazin.ua*

Gennadiy Pochanin
*A.Ya.Usikov Institute for Radiophysics and
Electronics of NAS of Ukraine
Kharkiv, Ukraine
gpp@ire.kharkov.ua*

Oleksandr Prishchenko
*School of Radio Physics, Biomedical
Electronics and Computer Systems,
V. N. Karazin Kharkiv National University,
Kharkiv, Ukraine*

Vadym Plakhtii
*School of Radio Physics, Biomedical
Electronics and Computer Systems,
V. N. Karazin Kharkiv National University,
Kharkiv, Ukraine*

Dmytro Shyrokorad
*Dept. of System Analysis and Computer
Mathematics
Zaporizhzhya National Technical University
Zaporizhzhya, Ukraine
hoveringphoenix@gmail.com*

Abstract—The problem of identification of objects under ground surface is solved by the application of irradiation of the surface by short impulse electromagnetic waves and the use of artificial neural networks (ANN) for the analysis of reflected field characteristics. As input data for ANN the normalized amplitudes of electrical component of the field in determined points of observation in equidistant moments of time are used. As an example of the object for the identification, the metal tube under surface of a ground is considered. The plane electromagnetic wave having Gaussian time dependence is used as an incident field. The influence of a number of hidden layers of ANN on precision of the recognition is investigated.

Keywords—artificial neural network, impulse electromagnetic wave, subsurface radar, object recognition.

I. INTRODUCTION

The detection and recognition of objects localized in the complicated media as soils of different kinds are the actual subjects for a number of applications [1]. Another stage of the development of the area is the usage of the ultrawideband radars [2]. Owing to the very wide range of frequencies of a sensing electromagnetic wave spectrum [3], the radars provide significantly higher precision of a resolution and a depth of penetration in lossy media [4] in comparison with traditional ground penetrating radar (GPR) [5]. The idea of a GPR that radiates the electromagnetic wave without definite carrier frequency was proposed by Cook [6] 60 years ago, but the impetuous development of theory and technique of all components of the radar permits to obtain the predicted characteristics [7] of the impulse GPR in current time only [8].

Except the simplest formulation of the problem for a uniform medium of propagation with known parameters there are a number of tasks concerning the investigation of objects by limited set of sources of a reflected field [9] or the reconstruction of a dielectric profile of inhomogeneous media [10]. The problems require the application of complicated techniques especially for cases of mathematically incorrect statements. As for the pursuit of hidden objects of complex shapes there was proposed the approach based on concept of a presence of individual resonant frequencies for the object response on

electromagnetic wave irradiation that can be used as their own footprints [11]. These frequencies were called “Natural Frequencies”. It was suggested that the objects form a response containing the natural frequencies under irradiation by an electromagnetic wave.

The investigation of reflected wave for the object recognition needs the application of complicated mathematical methods [12] to compensate the lack of input data. It is interesting to use more convenient and quick methods of recognition, for example, the approach built upon principles of information processing realized in cortexes of animals [13]. The understanding of the mechanism of brain unit action [14] permitted to construct artificial neural networks [15] that possess multidimensional function approximation properties of ultimate power [16]. Namely the ANN characteristics are used to solve the problem of dielectric object parameter finding from analysis of reflected electromagnetic fields [17]. The application of ANN significantly simplifies the solving of dielectric parameter recognition task by multifrequency multidimensional backscattering problem solution [18]. As it was mentioned above, the utilization of impulse electromagnetic wave for irradiation must provide researches by a bigger size of information about objects under investigation physically [19]. Moreover, the application of electromagnetic fields with ultra-wideband spectrum expands the possibilities to receive in reflected wave components that corresponds to the natural frequencies of hidden objects [11]. So, the approach based on ANN was used for analysis of dielectric parameters of a layered medium that is a model of human body surface [20-23]. There were studied and compared ANN of different structures. It was shown the stability of the parameter recognition in presence of noise of substantial level and measuring errors. The key hypothesis of the ANN actions was established on the prediction of self-invention of recognition method during training directly from time-dependent signals instead of its Fourier Transform or other preprocessing techniques [24]. As for tasks of the parameter recognition, the works [20-24] are more close to the practical problem of remote road quality surveying, but the results obtained cannot satisfy the parameter precision needed and reached by analytical method based on Hilbert

transform [25]. The purpose of the work is to apply the same approach of direct time-domain signal processing for a subsurface object recognition, for example, described in [8] the problem of land mine finding.

II. STATEMENT OF THE PROBLEM

The source of input signals is the amplitudes of the electrical component of electromagnetic field reflected from ground surface and underground objects. The normally incident plane electromagnetic wave having the Gaussian time dependence with duration 0.6 ns. The reflected field is measured at the height 250 mm under the ground surface that is the convenient height to arrange real receiving antenna system. The number of the points for the measurement is 15 with spatial step along the ground surface 100 mm. There are three models of dielectric characteristics of the ground material: homogeneous substance with permittivity $\epsilon=9$ and conductance $\sigma=0.005$ S/m; the same surrounding substance with trench of depth and width 600 mm filled with another matter having permittivity $\epsilon=12$ or $\epsilon=6$ and the same losses. Totally, it is considered six cases where the ground materials have no inclusions and have the inclusion in form of perfectly conducted tube with radius 100 mm buried into the depth 300 mm and oriented perpendicularly to the line of field probes and in parallel to the trench walls.

Each field probe transmit the signal to ANN in form of 500 values of amplitudes of the electrical field obtained with time step 30 ps that is not very dense in comparison with used in [25]. So, input layer of ANN must contain 7500 elements. The output layer consists of one neuron that shows the presence or the absence of tube. All neurons have the sigmoidal excitation function. One should solve the problem of diffraction of the impulse electromagnetic wave on the metal-dielectric structure in time domain, train the ANNs to recognize the presence of the tube for different number of hidden layers and different number of neurons in them.

III. THE SOLUTION OF THE PROBLEM

The problem of impulse electromagnetic wave scattering is solved directly in time domain by numerical FDTD method [26]. The results of the simulation are presented in Fig. 1-3, where the normalized amplitudes of electrical field are shown for different points in space along OX axis (from -700 mm to 700 mm) and for different moments of time (from 0 to 15 ns). The downward orientation of the time axis is chosen for better representation of underground object influence on amplitudes of reflected field and the object location in space without taking into account electromagnetic field slowdown in dielectric media. Fig. 1 describes the case of homogeneous substance with permittivity $\epsilon=9$ and conductance $\sigma=0.005$ S/m, Fig. 2 and Fig. 3 conform to the same substance with the trench filled with other matter of permittivity $\epsilon=6$ and $\epsilon=12$ correspondingly. The pictures marked by the letter "a" shows the cases without any scatterer under a ground, whereas the letter "b" designates the cases with presence of the metal tube.

All figures contain the lightest area of incident pulse appearing and the darkest region of field reflected from the ground surface and changed their polarity. Other domains include significantly weaker changes of field amplitudes caused by influences of metal surface and permittivity changes supplementary diminished by losses in the media. The reflection of the wave from metal tube generates typical hyperboloid-like shape caused by time delay in reaching the observation point shifted from normal to the surface [2].

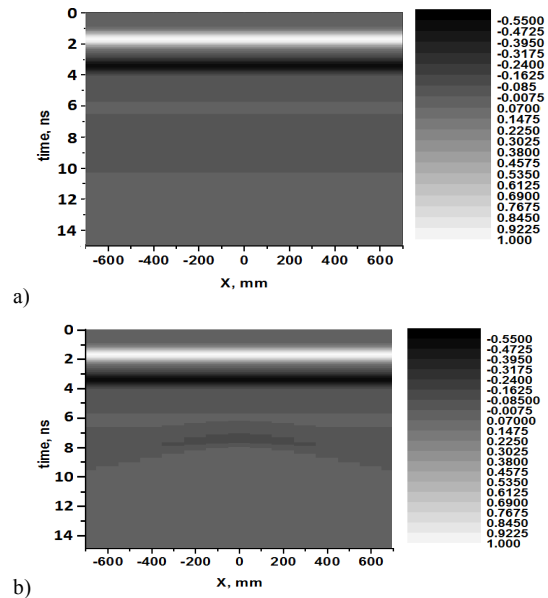


Fig. 1. Time dependence of the normalized amplitudes of electrical field in different points along OX axis calculated for the case of homogeneous substance with permittivity $\epsilon=9$ and conductance $\sigma=0.005$ S/m (a) and the same substance with the metal tube buried at the depth 300 mm (b).

It is interesting to note that the case of trench filled with matter with smaller permittivity (Fig. 2) reminds the impulse field behavior in rectangular waveguide [27] whereas the field distribution for the case of bigger permittivity inside trench (Fig. 3) looks like the pattern for homogeneous ground (Fig. 1). Each of the six time-spatial distributions depicted in Fig. 1-3 forms the training set for our ANN. The purpose of the training is to get the zero level of the ANN output signal for sets presented in Fig. 1-3a and the unit level for sets pictured in Fig. 1-3b. The final result of the learning is checked on verification sets presented in Fig. 4, where the time-spatial distribution of the amplitude of electrical field of reflected wave is depicted for the cases of tube hidden under ground surface for shifted positions relatively to the case pictured in Fig. 1a in 20 mm downward (Fig. 4a), left-hand (Fig. 4b), and right-hand (Fig. 4c). Last two cases are equivalent of 20% error in positioning along OX axis, the first case corresponds to 66 ps error in time, i.e. more than two time steps.

The structures of ANN and results their check on the data corresponded the cases depicted in Fig. 4 are presented in Table 1. It is seen that the case #2 of ANN with two hidden layers governs the worst result for all verification tests whereas the biggest deflection is observed for the case #1. The best fit is supervised in case #4 that can be explained by the biggest informational capacity of the ANN that helps to successfully recognize the presence of the tube.

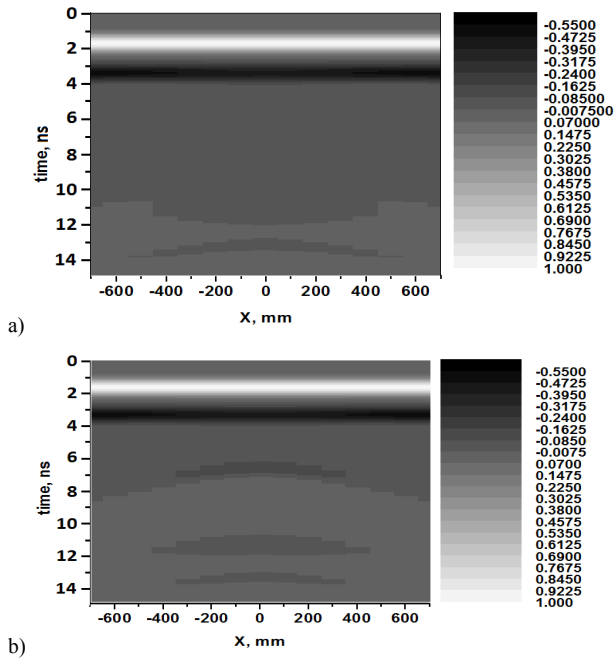


Fig. 2. Time dependence of the normalized amplitudes of electrical field in different points along OX axis calculated for the case of homogeneous substance with permittivity $\epsilon=9$ and conductance $\sigma=0.005$ S/m with trench filled with the matter with permittivity $\epsilon=6$ and the same $\sigma=0.005$ S/m (a) and for the same substances and geometry but with the metal tube buried at the depth 300 mm (b).

It is interesting to illustrate the values of weight coefficients between layers for the best case #4 presented in Table I. The data are depicted in Fig. 5, where the numbers of interconnected neurons are shown on axes, and the brightness of the corresponding dots reflects the values of the weight coefficients

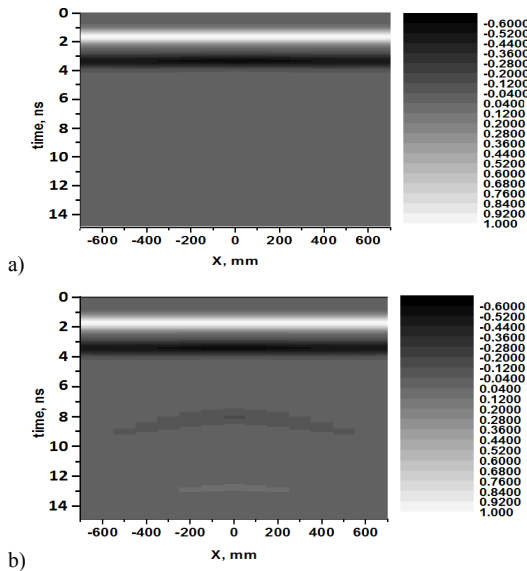


Fig. 3. Time dependence of the normalized amplitudes of electrical field in different points along OX axis calculated for the case of homogeneous substance with permittivity $\epsilon=9$ and conductance $\sigma=0.005$ S/m with trench filled with the matter with permittivity $\epsilon=12$ and the same $\sigma=0.005$ S/m (a) and for the same substances and geometry but with the metal tube buried at the depth 300 mm (b).

Its magnitudes for third and fourth layers are displayed in Fig. 5a, for second and third layers are represented in Fig. 5b, for first and second layers are imaged in Fig. 5c. It is seen that the values have chaotic character in Fig. 5a and Fig. 5b, but the picture in Fig. 5c shows definite periodicity that corresponds to similar algorithm of processing data from each of 15 probes found by ANN during training.

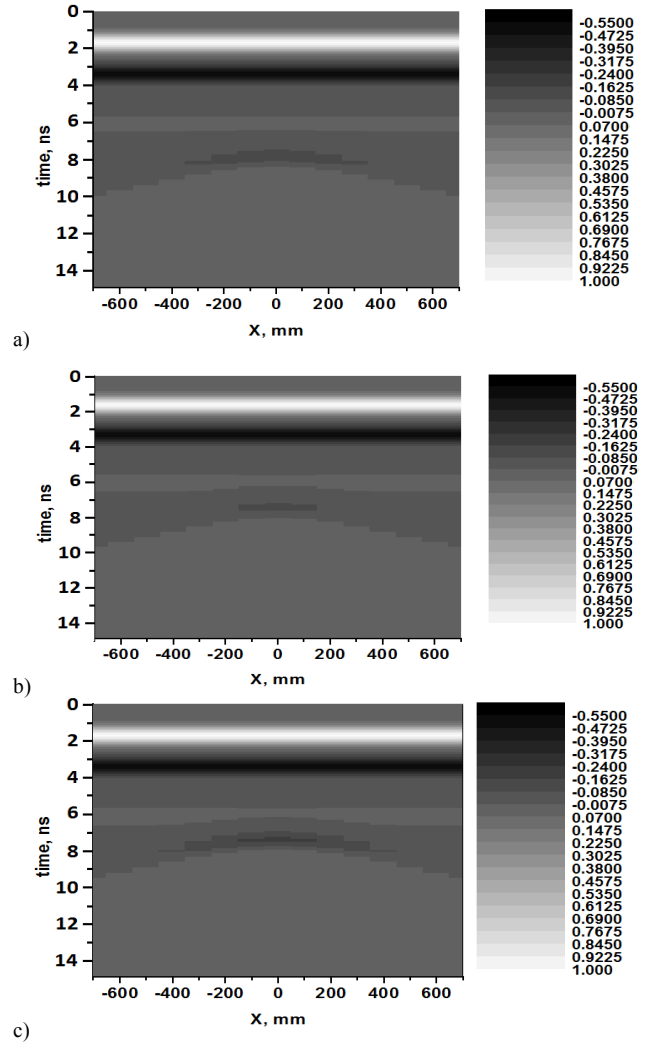


Fig. 4. Time dependence of the normalized amplitudes of electrical field in different points along OX axis calculated for the case of homogeneous substance with permittivity $\epsilon=9$ and conductance $\sigma=0.005$ S/m with the buried metal tube as well as presented in Fig. 1b but for shifted position of tube in 20 mm downward (a), left-hand (b), and right-hand (c).

IV. CONCLUSION

It is shown that ANN can effectively find objects whose presence have a distributed influence on data acquired by the impulse electromagnetic field irradiation of a ground. It is seen that a bigger number of hidden layers of ANN permits to improve the object recognition quality. The occurrence of errors and a low contrast of significant part of the training input data do not hamper the creation of successful methods of object recognition by ANN. The process of ANN training has created a necessary algorithm of signal processing and recognized the samples of the data from different probes blindly preferring the usage of similar manipulation with input data of different probes.

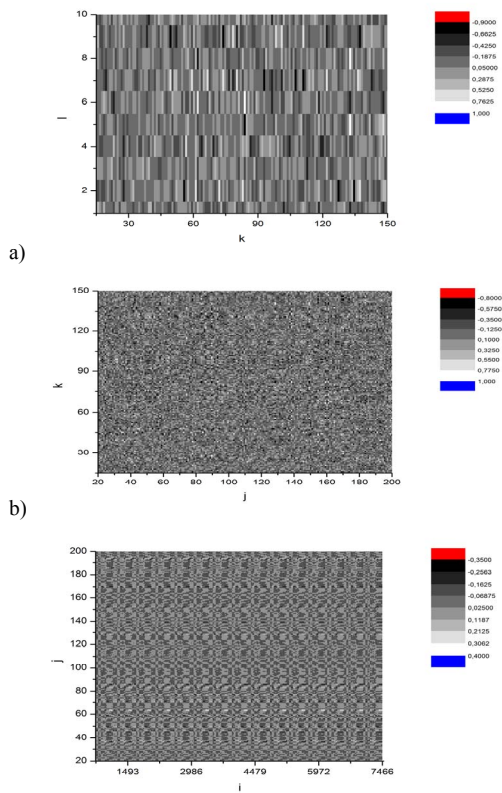


Fig. 4. The values of weight coefficients between k and l neurons of third and fourth layers (a), j and k neurons of second and third layers (b), i and j neurons of first and second layers (c).

TABLE I. ANN STRUCTURES AND RESULTS OF THEIR VERIFICATIONS

ANN number	ANN characteristics			
	Structure, number of neurons in layers	Output signal for case in Fig. 4a	Output signal for case in Fig. 4b	Output signal for case in Fig. 4c
1	7500-100-50-25-1	0.9943	1.2001	1.3316
2.	7500-100-50-1	0.9809	0.9569	0.9493
3.	7500-200-100-50-1	1.0085	1.0227	0.9471
4.	7500-200-150-10-1	1.0065	1.0068	1.0099

REFERENCES

[1] A. S. Turk, K. A. Hocaoglu, and A. A. Vertiy, *Subsurface Sensing, Ho-boken*: Wiley, 2011.

[2] J. D. Taylor, *Ultrawideband radar: applications and design*. Boca Raton, London, NewYork: CRC Press, 2012.

[3] H. F. Harmuth, R. N. Boules, and M. G. M. Hussain, *Electromagnetic signals: reflection, focusing, distortion, and their practical applications*. NewYork: Kluwer Academic, Plenum Publishers, 1999.

[4] I. Immoreev, S. Samkov, and Teh-HoTao, "Short-Distance Ultra-Wideband Radars," *IEEE Aerospaceand Electronic Systems Magazine*, vol. 20, no. 6, pp. 9–14, 2005.

[5] D. J. Daniels, *Ground penetrating radar*, 2nd ed. London: IEEE, 2004.

[6] J. C. Cook, "Proposed monocycle-pulse very high frequency radar for airborne ice and snow measurement," *Trans. AIEE Commun. Electron.*, no. 79, pp. 588–594, 1960.

[7] H. Harmuth, *Nonsinusoidal waves for radar and radiocommunications*. New York: Academic Press, 1981.

[8] G. Pochanin, S. Masalov, I. Pochanina, L. Capineri, P. Falorni, and T. Bechtel, "Modern Trends in Development and Application of the

UWB Radar Systems," 8th International Conference on Ultrawideband and Ultrashort Impulse Signals, Odessa, Ukraine, pp. 7–11, 5-11 September 2016,

[9] O. O. Drobakhin, A. V. Doronin, and V. V. Grigor'ev, "3-probe microwave measuring instrument of vibration of mechanical objects with non-plane surface," 7th Intern. Conf. on Antenna Theory and Techniques, Lviv, Ukraine, pp. 277–279, 2009.

[10] S. Alexin, O. Drobakhin, and V. Tkachenko, "Reconstruction of permittivity profile for stratified dielectric material: Gel'fand-Levitan and Newton-Kantorovich methods," XII Int. Conf. on Math. Meth. in Electrom. Theory (MMET), Odesa, Ukraine, pp. 141–143, 2008.

[11] C. E. Baum, "Direct Construction of a Ksi-Pulse from Natural Frequencies and the Evaluation of the Late-Time Residuals," Interaction Note 519, May 1996, pp. 349-360, in G. Heyman et al (eds), *Ultra-Wideband, Short-Pulse Electromagnetics 4*, Kluwer Academic/Plenum Publishers, 1999.

[12] M. V. Andreev, and O. O. Drobakhin, "Feature of Prony's Method Application for Natural Frequencies Estimation from the Frequency Response," 8th International Conference on Ultrawideband and Ultrashort Impulse Signals, Odessa, Ukraine, pp. 18-20, 5-11 September 2016.

[13] D. H. Hubel and T. N. Wiesel, "Receptive fields, binocular interaction, and functional architecture in the cat's visual cortex," *Journal of Physiology, London*, vol. 160, pp. 106–154, 1962.

[14] D. Hebb, *Organization of behaviour*. New York, J. Wiley, 1949.

[15] S. Haykin, *Neural Networks*, 2nd ed. New Jersey: Prentice-Hall, 1999.

[16] R. Callan, *The essence of neural networks*. New York : Prentice Hall Europe, 1999.

[17] O. Drobakhin, and A. Doronin, "Estimation of thickness of subsurface air layer by neuron network technology application to reflected microwave signal," XII Int. Conf. on MMET, Odesa, Ukraine, pp. 150-152, 2008.

[18] O. O. Drobakhin, and A. V. Doronin, "Neural network application for dielectric structure parameter determination by multifrequency methods," Third International Conference of Ultrawideband and ultrashort impulse signals, Sevastopol, Ukraine, pp. 358–360, 2006.

[19] L. A. Varyanitsa-Roshchupkina, and G. P. Pochanin, "Video Pulse Electromagnetic Wave Diffraction on Subsurface Objects," *Telecommunications and Radio Engineering*, vol. 66, no. 5, pp. 391-414, 2007.

[20] D. Shyrokorad, O. Dumin, and O. Dumina, "Time domain analysis of reflected impulse fields by artificial neural network," IV Conf. on UWBUSIS, Sevastopol', Ukraine, pp. 124-126, 2008.

[21] O. Dumin, O. Dumina, and D. Shyrokorad, "Time domain analysis of fields reflected from model of human body surface using artificial neural network," in Proc. EuCAP, Berlin, pp. 235-238, 2009.

[22] D. Shyrokorad, O. Dumin, O. Dumina, V. Katrich, and V. Chebotarev "Approximating properties of artificial neural network in time domain for the analysis of electromagnetic fields reflected from model of human body surface," Proc. MSMW, Kharkiv, Ukraine, pp. 1-3, 2010.

[23] D. Shyrokorad, O. Dumin, O. Dumina, and V. Katrich, "Analysis of transient fields reflected from model of human body surface using convolutional neural network," Proc. MMET, Kyiv, pp. 1-4, 2010.

[24] O. Dumin, S. Khmara, and D. Shyrokorad, "Artificial neural networks in time domain electromagnetics," Proc. of 11th International Conference on Antenna Theory and Techniques (ICATT–2017), Kyiv, Ukraine, pp. 118-121, 2017.

[25] G. P. Pochanin, V. P. Ruban, P. V. Kholod, O. A. Shuba, I. Ye. Pochanina, A. G. Batrakova, S. N. Urdzik, D. O. Batrakov, and D. V. Golovin, "Advances in ground penetrating radars for road surveying," Ultrawideband and Ultrashort Impulse Signals, Kharkiv, Ukraine, pp. 13-18, 15-19 September 2014.

[26] A. Taflove, and S. Hagness, *Computational Electrodynamics: The Finite-Difference Time-Domain Method*, 3rd ed. Boston, London: Artech House, 2005.

[27] V. A. Katrich., A. N. Dumin, and O. A. Dumina, "Radiation of transient fields from the open end of rectangular waveguide," IV International Conf. on Antenna Theory and Techniques (ICATT–2003), Sevastopol, Ukraine, pp. 583–586, 2003.

Neural-Like Means for Data Streams Encryption and Decryption in Real Time

Ivan Tsmots
Lviv Polytechnic National University
Lviv, Ukraine
ivan.tsmots@gmail.com

Oleksa Skorokhoda
Lviv Polytechnic National University
Lviv, Ukraine
oskorokhoda@lp.edu.ua

Viktor Khavalko
Lviv Polytechnic National University
Lviv, Ukraine
khavalkov@gmail.com

Yurii Tsymbal
Lviv Polytechnic National University
Lviv, Ukraine
yurij.tsymbal@gmail.com

Taras Tesluyk
Lviv Polytechnic National University
Lviv, Ukraine
taras.tesluyk@gmail.com

Abstract — The “model of successive geometric transformations” paradigm has been adapted for the implementation of parallel-streaming neural network encryption-decryption of data in real time. A model and structure of a parallel-streaming neural-like element for the mode have been developed.

Keywords—intensive data stream; neural networks; geometric transformations model

I. INTRODUCTION

The latest information technologies are becoming global in the modern world. Their development and development of communications provides ever-widening opportunities for access to information resources and the transfer of large amounts of data for unlimited distances. In the context of the intensive development of the market for information products and services, information has become a full-fledged product that has its own consumer properties and cost characteristics.

The widespread introduction of information technology makes a relevant problem for the protection of the transmission of information using cryptographic methods that provide encryption of the ready-to-transmit information. The encrypted information is transmitted by a communication channel to an authorized user, who, after receiving it, performs decryption using a reverse transformation. Cryptographic transformations are carried out using special algorithms. In order to encrypt and decrypt real-time data streams, it is suggested to use neural-like network algorithms, the key in which is the network architecture, weighting factors, and masking codes.

Real-time encryption and decryption of intensive streams can be ensured through VLSI-implementation of the corresponding algorithms. For the synthesis of neural-like elements and neural-like real-time networks, it is necessary to develop new parallel-stream neural elements and networks that provide spatiotemporal parallelization of encryption and decryption algorithms. To implement such algorithms programmable logic integrated circuits (FPGAs) can be used. The main advantages of FPGA are low cost, affordability, high performance, reliability and availability of a variety of well-developed and efficient software tools for automated design. A significant feature of the latest generation of

FPGAs is the possibility of reconfiguration of such circuits during operation.

Therefore, development of hardware neural-like networks for encryption-decryption of intense data streams in real time is the actual problem.

II. ANALYSIS OF PUBLICATIONS

Analysis of works [1-15] provided features highlighting of real-time neurocomputers tasks and architectures. Real-time encryption-decryption tasks are characterized by high intensity, continuity of incoming data streams and increased requirements for key lifetime.

The works [3-9] analyze the existing neural networks and the means of their implementation. The analysis showed that the vast majority of neural networks are implemented by software. Such neural networks have relatively low performance and do not provide real-time for the processing of intense data streams.

From the analysis of works [1-13] it is evident that in order to ensure the high performance of neural networks in real time, it is necessary to use hardware implementation and a modern element base. A prerequisite for the construction of efficient neurocomputer architectures is a usage of three basic principles: the parallelism of processing; programmability of the structure; regularity (uniformity) of the structure. This approach is provided by a combination of principles of conveyance, vector and matrix software and hardware organization of computing on the basis of the latest elemental base technologies. These three principles correspond to the problem of adequately mapping the spatial-temporal algorithmic structure of computational processes into the architecture of parallel computers [1].

Several directions in the creation of a new approach to computing processes in neurocomputers were identified in works [1-15]: high level of paralleling at all levels of data processing; pipeline principle of data processing; organization of high-level internal language with hardware support; increasing the regularity of hardware implementations, for example, using the systolic structures; use of multicast access storage environment; the transition to a universal matrix-algebraic system instead of the algebra of

real numbers; hardware implementation of basic operations in computational procedures.

The main components of the hardware neural network are artificial neural elements. There are different models of an artificial neuron. Choosing an artificial neuron model depends on the requirements of specific applications. In [10], the models and VLSI structures of the parallel-vertical type formal neuron are reviewed, which differ in the way of receiving and processing of input data and weight coefficients - using multiplexing of tires, combining the processes of data receipt and processing, and with table formations of macro-partial results. The disadvantage of these models and VLSI structures of the formal neuron is relatively low performance.

It follows from the analysis that the synthesis of the hardware neural network for real-time data encryption and decryption tasks requires the development of a new model and new VLSI structures of the neural element, which should be focused on the processing of intensive data streams.

One of the promising directions for the construction of high-performance neural network means is the application of the “model of successive geometric transformations” (MSGT) paradigm proposed and developed by R. Tkachenko [16-17].

An alternative approach to solving data encryption tasks using algebraic transformations is considered in [18-19].

The purpose of this paper is to adapt the model of successive geometric transformations to the tasks of data encryption and decryption, development of a model and a structure of a parallel-streaming neural-like element and synthesis on its base of a parallel-streaming neural-like network for data streams encryption and decryption.

III. MAIN PART

A. The model of successive geometric transformations

The basis of this paradigm is the non-iterative approach to the teaching of a neural-like network, which involves the direct calculation of weight coefficients during the gradual reduction of the dimension of the space of incoming multidimensional data on neurons in the hidden layer [16]. In this case, the representation of incoming multidimensional data in the new orthogonal basis is used, executed on the basis of the non-iterative greedy algorithm of the most distant point.

The hardware implementation of such neural-like networks using VLSI structures can be greatly simplified if the input, output and weighting coefficients of the MSGT network are presented in a fixed-point format. For this, a preliminary scaling of the input data is anticipated.

When choosing the structure of a neural-like network for real-time encryption-decryption of data streams, it is proposed to use the architecture of an auto-associative network with one hidden layer [17] (Fig. 1). Successive geometric transformations are depicted by lateral connections between neurons in a hidden layer.

This network structure is quite versatile and can be used to solve various tasks that involve converting incoming data and its subsequent recovering: encoding input data for

compression, block symmetric encryption, overlaying of digital watermarks (on image) and steganography. Encryption and decryption modules, respectively, form and use a key that forms the parameters of the trained neural network. To improve the cryptographic stability of the hidden layer output, the structure of the network can be supplemented with a block of encryption using a one-time notepad (masking by XOR).

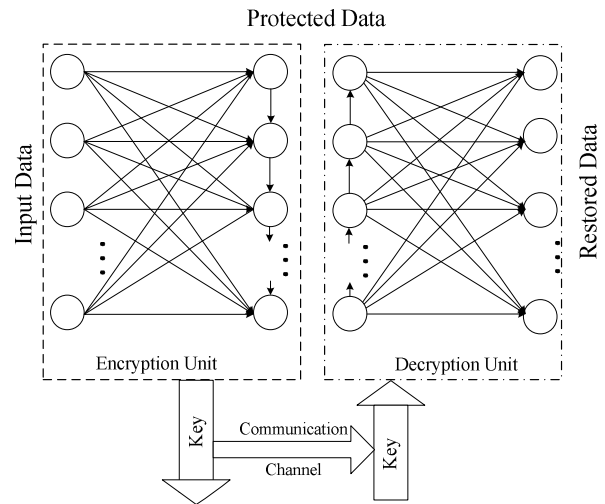


Fig. 1. The graph of a neural-like network for encryption-decryption of data based on a model of successive geometric transformations

With the number of neurons in the hidden layer equal to the number of input and output neurons, it is possible to restore without loss secure data on the network output that will be obtained at the outputs of the hidden layer neurons. Reverse consecutive geometric transformations between hidden and output neurons are used to restore data on network outputs.

It is possible to serially connect several encryption blocks and subsequent appropriate decryption units to create a cascading multilayer network.

Performance of the encryption module can be described as follows:

$$h = F(X, key), \quad (1)$$

where X, h – signals on the input and output of the module, respectively; $key = \{N, W, mask\}$ - a key that consists of a given number of neurons in the input (hidden, output) layer of the network N , the matrix of weight coefficients W , and the mask for a one-time notebook $mask$, F - the function that specifies straight consequent geometric transformations.

Then the functioning of the decryption module is described as:

$$\tilde{X} = \bar{F}(h, key), \quad (2)$$

where h – the signal at the input of the module (protected data), \tilde{X} – the signal at the output of the module (recovered data), \overline{F} – the function that specifies reverse consecutive geometric transformations.

The lifetime of a key key then depends on:

- the number of neurons N ;
- the bit-width n of the mask;
- the number of serially connected encryption and decryption blocks in the cascade network k .

A. The model of a parallel-streaming neural-like element

The main components on the basis of which parallel-streaming neural systems are synthesized are neural elements. The peculiarity of the work of neural-like systems is that they are focused on solving specific problems using the “model of consecutive geometric transformations” paradigm. In a parallel-streaming neural-like element, which is realized on the basis of such a paradigm, weight coefficients W_j are pre-calculated and do not change or change very rarely in the process of operation. In the general case, the neural element makes a transformation in accordance with the formula:

$$Y = f\left(\sum_{j=1}^N W_j X_j\right), \quad (3)$$

where Y – the output signal of the neural element, f – the activation function, N – the number of inputs.

From the formula (3) it follows that the processing of data in neural elements is reduced to calculations of scalar product and activation functions f . During calculation of the scalar product in a parallel-streaming neural-like element, weighted coefficients W_j are pre-computed and stored in memory, and the input data arrives at the same time on all inputs as the parallel binary code.

In parallel-streaming neural-like systems, which are oriented on VLSI implementation, it is expedient to use table-algorithmic methods and the basis of elementary operations for their implementation. The basic operation of a parallel-streaming neural-like element is the operation of calculating the scalar product, which for VLSI-implementation must be provided on the basis of elementary operations. To calculate the scalar product on the basis of elementary operations, it is expedient to use multiplication algorithms with the direct formation of partial products, since they are regular and well-structured. The most common of these are algorithms of multiplication with the analysis of a one bit-slice. The multiplication of the numbers given in the binary complement code, with the analysis of a single digit of the multiplier, is written as follows:

$$\begin{aligned} C_j &= W_j X_j = \sum_{i=0}^{n-1} (-1)^{2^i} 2^{-i} W_j x_i \\ &= \sum_{i=0}^{n-1} (-1)^{2^i} 2^{-i} P_{ji} \end{aligned}, \quad (4)$$

where n – the bit-length of the multiplier; x_i – the value of the i -th bit of the multiplier; P_{ji} – the i -th partial product.

Using the multiplication algorithm (4) we develop a parallel-flow algorithm for scalar products calculation. The algorithm for calculating a scalar product using the multiplication algorithm (4) is written as follows:

$$\begin{aligned} Z &= \sum_{j=1}^N W_j X_j = \sum_{j=1}^N \sum_{i=0}^{n-1} (-1)^{2^i} 2^{-i} P_{ji} \\ &= \sum_{i=0}^{n-1} (-1)^{2^i} 2^{-i} \sum_{j=1}^N P_{ji} \end{aligned} \quad (5)$$

If in the formula (5) the sum of i -th partial products $\sum_{j=1}^N P_{ji}$ replace by the i -th macro-partial product P_{Mi} then we get

$$Z = \sum_{j=1}^N W_j X_j = \sum_{i=1}^n (-1)^{2^i} 2^{-i} P_{Mi} \quad (6)$$

In neural-like elements, weighted coefficients W_j are pre-calculated, that is, they can be considered as constants. In this case, when calculating the scalar product by formula (6), we can calculate the table of macro-partial products P_{Mi} in advance. The calculation of the values of the table of macro-partial products P_{Mi} is carried out according to the following formula:

$$P_{Mi} = \begin{cases} 0, & \text{if } x_{1i} = x_{2i} = x_{3i} = \dots = x_{Ni} = 0 \\ W_1, & \text{if } x_{1i} = 1, x_{2i} = x_{3i} = \dots = x_{Ni} = 0 \\ W_2, & \text{if } x_{1i} = 0, x_{2i} = 1, x_{3i} = \dots = x_{Ni} = 0 \\ W_1 + W_2, & \text{if } x_{1i} = 1, x_{2i} = 1, x_{3i} = \dots = x_{Ni} = 0 \\ \vdots \\ W_2 + \dots + W_N, & \dots \\ \text{if } x_{1i} = 0, x_{2i} = x_{3i} = \dots = x_{Ni} = 1 \\ W_1 + W_2 + \dots + W_N, & \\ \text{if } x_{1i} = x_{2i} = x_{3i} = \dots = x_{Ni} = 1 \end{cases} \quad (7)$$

The volume of the table of macro-partial products P_{Mi} is determined by the formula $Q = 2^N$.

In the parallel-streaming scalar product calculation by the formulas (6) and (7), the spatiotemporal parallelization of the computation process is used. The peculiarity of the parallel-streaming implementation of scalar product calculations is the use of n identical processing elements (steps) and pipelining of the process. For such implementation, it is advisable to use the algorithm to calculate the scalar product starting with the analysis of least significant bits, which will ensure the use of $(n + \log_2 N)$ -bit adders in all processor elements. In this case, in each i -th processor element the following operation will be performed:

$$Z_i = 2^{-1} Z_{i-1} + P_{Mn-(i-1)}, \quad (8)$$

where $Z_0 = 0$.

Analytically, the model of a parallel-streaming neural-like element can be written as follows:

$$Y = f_a \left(\begin{array}{l} f_{3n1}(f_{TP_{Mn}}(f_{\Sigma_1}(f_{Z_1R1}(\dots(f_{3ni}(f_{TP_{M(n-i-1)}} \\ (f_{\Sigma_i}(f_{Z_iR1}(\dots(f_{3ni}(f_{TP_{M1}}(f_{\Sigma_n})))))))))) \end{array} \right) \quad (9)$$

where Y – output of the neural-like element; f_a – activation function; f_{3n1} – buffer memorization of incoming n bit data, $f_{TP_{Mn}}$ – tabular formation of macro-partial products for n -th bits, f_{Σ_1} – calculation of the first partial result $Z_1 = 2^{-1} Z_0 + P_{Mn} = P_{Mn}$, f_{R1Z_1} – shift to the right for one level of the first partial result $R1Z_1$, f_{3ni} – buffer memorization of $(n-i-1)$ -bit operands, $f_{TP_{M(n-i-1)}}$ – tabular formation of macro-partial products for $(n-i-1)$ -th bits of operands, f_{Σ_i} – calculation of i -th partial result $Z_i = 2^{-1} Z_{i-1} + P_{M(n-i-1)}$, f_{R1Z_i} – shift to the right for one level of i -th partial result $R1Z_i$, f_{3nn} – buffer memorization of 1st bits of operands, $f_{TP_{M1}}$ – tabular formation of macro-partial products for 1st bits of operands, f_{Σ_n} – calculations of n -th partial result $Z_n = 2^{-1} Z_{n-1} + P_{M1}$.

The structure of the model of a parallel-streaming neural-like element, which implements the expression (9), is given in Fig. 2

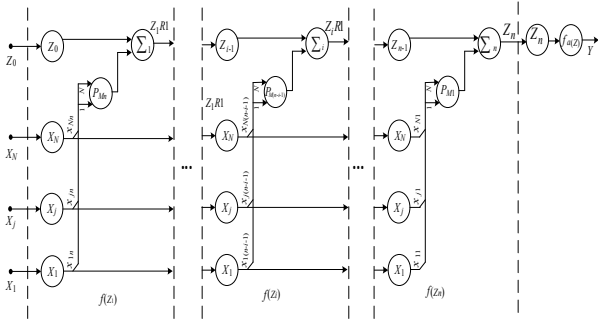


Fig. 2. Model of parallel-streaming neural-like element

The main components of this model are: buffer pipeline memory, memory of macro-partial products and $(n + \log_2 N)$ -bits adders. The peculiarity of this model is the combination of simultaneous processing of n input data arrays, which provides high performance.

A. Structure of a parallel-streaming neural-like element

The development of VLSI structures of a parallel-stream neural-like element for real-time neural networks synthesis with high efficiency of equipment use is proposed to carry out on the basis of an integrated approach based on the

capabilities of a modern element base, covering the methods, algorithms and structures of hardware of neural networks, taking into account the requirements of specific applications. For a complete use of the advantages of modern VLSI technology, the following principles are proposed on the basis of a parallel-streaming neural-like element:

- use of the basis of elementary arithmetic operations;
- preliminary calculation of weighting factors;
- tabular formation of macro products and activation functions;
- pipeline and spatial parallelism;
- homogeneity and modularity of the structure.

In developing the structure of a parallel-streaming neural-like element, we will use its model (Fig. 2). To switch from a model to a structure of a parallel-streaming neural-like element it is necessary to display all components of the model on hardware. In such a neural-like element, the components of the calculation of the activation function and macro-partial products are implemented using memory tables, buffer memorization - using registers, and the calculation of partial results - on the adders.

The structure of the parallel-streaming neural-like element is shown in Fig. 3, where PU – processing unit, Rg – register, RAM – random access memory, Add – adder, P_{Mi} –input of macro-partial products, C_1, C_2 – first and second control inputs, X_1, \dots, X_n – n data inputs, Wr_1, Wr_2 – first and second input for writing into RAM.

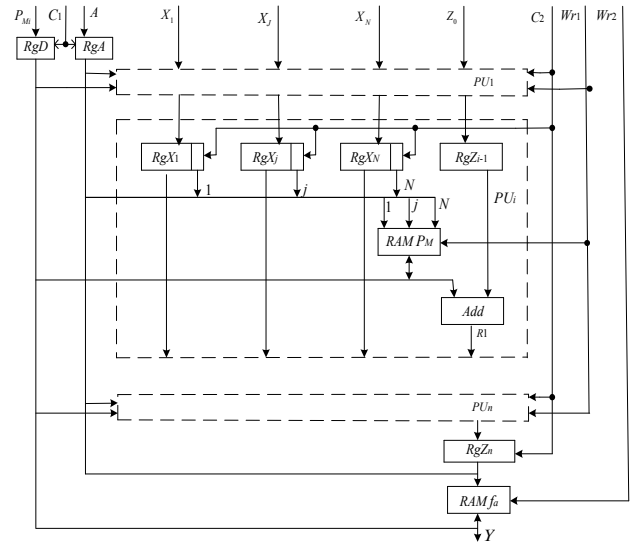


Fig. 3. Structure of a parallel-streaming neural-like element

The calculation of the scalar product in this device requires a preliminary calculation of the weight coefficients W_j and on their basis the formation of the 2^N macro-partial products P_{Mi} according to formula (7), which must be written in the RAM P_M in all PUs. Before beginning the recording of macro products P_{Mi} by signal 0 from the input C_2 outputs of registers $Rg_{X1} - Rg_{XN}$ in all PUs are switched in the third state and RAM_{P_M} in all PUs are switched to the recording mode using the signal 0 from the input Wr_1 . By signaling 1 from input C_1 outputs of registers Rg_D and Rg_A

are set to transmit data and addresses to the corresponding inputs of RAM P_M . Recording of macro-partial products P_{Mi} in the RAM P_M of all PUs is carried out simultaneously starting from the zero address, which is set at the outputs of the registry RgA . To write the k -th macro-partial product in the RAM P_M ($k = 1, \dots, 2^N$), it must be written in the register RgD , and in the RgA – the value of $k-1$.

The calculation of the activation function in a neural-like element is carried out tabularly. The values of the activation function are recorded in RAM f_a using the registers RgD and RgA . To write data in RAM f_a it must be switched to recording mode using the signal 0 from the input Wr_2 . After the data is recorded in the RAM f_a and RAM P_M , they are switched to read mode, and registers RgD and RgA are set to the third state.

The characteristic of the developed structure of the parallel-streaming neural-like element (Fig. 3) is its implementation on n PU of the same type, which operate according to the pipeline principle and provide the time parallelization of the algorithm of work in such a way that the results of the work of i -th PU are input data for $(i+1)$ -th PU. The velocity of a parallel-streaming neural-like element is determined by the tact of the pipeline's operation, which is determined by the following formula:

$$T_{\varepsilon} = \tau_{P\gamma} + \tau_{PAM} + \tau_{A\delta\delta}, \quad (10)$$

where t_{Rg} – time delay of the register; t_{RAM} – time to read data from memory; t_{Add} – time of addition on the adder.

D. Basic structure of parallel-stream neural network for encryption-decryption of data

The basic structure of the parallel-stream neural network for data encryption-decryption in real time will be synthesized on the basis of the developed parallel-streaming neural-like element (Fig. 3). The purpose of synthesis of such a neural-like network is to obtain a modular and regular structure oriented on VLSI technology.

The initial information for the synthesis of a real-time neural-like network is:

- algorithms of training and operation of the neural network;
- graph representation of the neural network;
- the number of input N and neurons K ;
- the intensity of incoming data;
- requirements for the interface;
- bit-length of input data, weighting factors, and accuracy of calculations;
- technical and economic requirements and restrictions.

When synthesizing a neural-like network it is necessary to ensure its functioning in real time with minimal hardware costs. The transition from graph representation to the hardware structure of a neural-like network formally reduces to minimizing hardware costs while providing a real-time mode.

The structure of the neural-like network for data encryption is shown in Fig. 4, where PU – processing unit, Rg – register, RAM – random access memory, Add – adder, Sub – subtractor, C_1, C_2, C_3 – first, second and third control inputs, InD – data input, $OutY$ – output of the result of encryption.

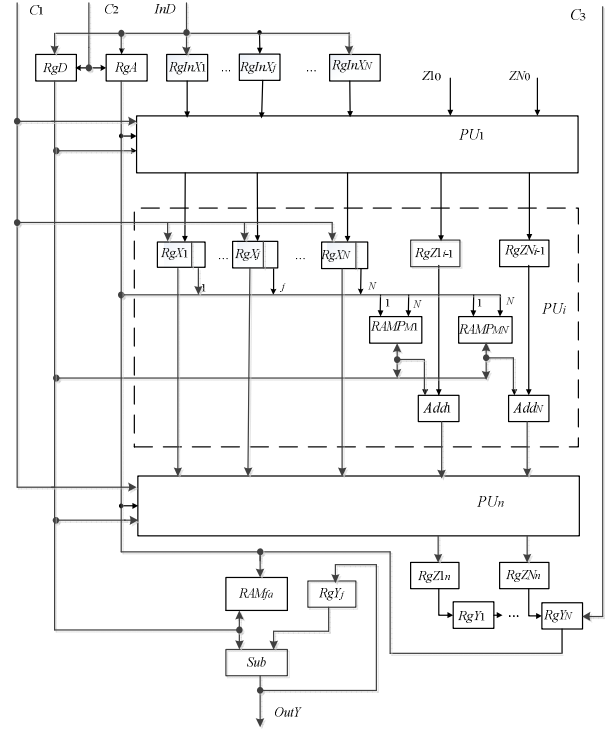


Fig. 4. Structure of a parallel-stream neural network for data streams encryption

Encryption of data streams using a parallel-stream neural network requires a preliminary calculation of weight coefficients W_j , the formation of macro-partial products P_M and their simultaneous recording in all RAM P_M . The peculiarity of the developed structure of the parallel-streaming neural-like encryption network is that the data move consecutively one after another and with the help of inputs $RgIn_1, \dots, RgIn_N$ is converted into a parallel stream of data entering the input of the first PU_1 . Parallel-streaming neural network is implemented based on n identical PU_n , which operates according to the pipeline principle. The operating time of the pipeline of such a network is equal to the tact of the neural-like element (10). In each cycle of work, calculated scalar products are written in registers RgZ_1, \dots, RgZ_N , and after them in registers RgY_1, \dots, RgY_N with the help of which parallel-sequential transformation of the receipt of scalar products is performed. At the output of the subtractor Sub the stream of encrypted data is formed.

IV. CONCLUSIONS

The “model of successive geometric transformations” paradigm has been adapted for the implementation of parallel-streaming neural network encryption-decryption of data in real time.

A model and structure of a parallel-streaming neural-like element have been developed, which provides spatiotemporal parallelization of the process of calculation and its implementation on the basis of n processing units of the same type.

For the VLSI-implementation of parallel-streaming neural-like element and network, a table-algorithmic method of calculation on the basis of elementary operations has been used.

Following principles are proposed for implementation of a parallel-streaming neural-like element: use of the basis of elementary arithmetic operations; preliminary calculation of weighting factors; tabular formation of macro products and activation functions; pipeline and spatial parallelism; homogeneity and modularity of the structure.

The velocity of a parallel-streaming neural-like network for data encrypting-decrypting is determined by the tact of the pipeline's operation, which is determined as the sum of the times of data delay on the register, reading data from memory and adding of two numbers.

REFERENCES

- [1] A. V. Palagin, and Yu. S. Yakovlev, System integration of computer equipment. Vinnytsia: UNIVERSUM-Vinnytsia, 2005. (in Russian)
- [2] V. P. Gribachev, "Element base of hardware implementations of neural networks," in Components and technologies, no. 8, 2006. (in Russian)
- [3] S. Haykin, Neural networks and learning machines, 3rd ed. New York: Prentice Hall, 2009.
- [4] Ye. V. Bodyanskiy and O. G. Rudenko, Artificial neural networks: architectures, learning, applications. Kharkiv: TELETEH, 2004. (in Russian)
- [5] W. S. McCulloch, and W. Pitts, "A logical calculus of the ideas immanent in nervous activity" in The Bulletin of Mathematical Biophysics, vol. 5, iss. 4, pp. 115–133, 1943.
- [6] ADALINE (Adaptive linear) [Electronic Resource]: <http://www.cs.utsa.edu/~bylander/cs4793/learnsc32.pdf>
- [7] K. Fukushima, "Cognitron: A self-organizing multilayered neural network" in Biological cybernetics, vol. 20, iss. 3-4, pp. 121–136, 1975.
- [8] J. J. Hopfield, "Neural networks and physical systems with emergent collective computational abilities" in Proceedings of the national academy of sciences, vol. 79, iss. 8, pp. 2554–2558, 1982.
- [9] J. Cao, and J. Liang, "Boundedness and stability for Cohen–Grossberg neural network with time-varying delays," in Journal of Mathematical Analysis and Applications, vol. 296, iss. 2, pp. 665–685, 2004.
- [10] Yu. M. Rashkevich, R. O. Tkachenko, I. G. Tsmots, and D. D. Peleshko, Non-linear methods, algorithms and structures for processing of signals and images in real time: monograph. Lviv: Lviv Polytechnic Publishing House, 2014. (in Ukrainian)
- [11] I. G. Tsmots, O. V. Skorokhoda, and B. I. Balych, "Model and VLSI structures of the parallel-vertical type formal neuron using bus multiplexing," in Modeling and Information Technologies, Digest of Scientific Papers of the G.E. Puhov Institute of Modeling Problems in the Energy, Lviv, vol. 67, pp. 160-166, 2013. (in Ukrainian)
- [12] I. G. Tsmots, O. V. Skorokhoda, and V. B. Krasovskii, "Models and VLSI structures of a parallel-vertical type formal neuron combining the processes of data incoming and processing," in Modeling and Information Technologies, Digest of Scientific Papers of the G.E. Puhov Institute of Modeling Problems in the Energy, Lviv, vol. 70, pp. 137-145, 2013. (in Ukrainian)
- [13] I. G. Tsmots, O. V. Skorokhoda, and B. I. Balych, "Model and VLSI structure of a parallel-vertical type formal neuron with tabular macro-partial results," in Modeling and Information Technologies, Digest of Scientific Papers of the G.E. Puhov Institute of Modeling Problems in the Energy, Lviv, vol. 73, pp. 133-138, 2014. (in Ukrainian)
- [14] I. G. Tsmots, O. V. Skorokhoda, and V. M. Tesliuk, A device for calculating scalar product. Patent № 101922 Ukraine, G06F 7/38. Bul. no. 9, 2013. (in Ukrainian)
- [15] I. Tsmots, O. Skorokhoda, V. Rabyk, and I. Ignatyev, "Basic vertical-parallel real time neural network components," XIIIth International Scientific and Technical Conference "Computer Sciences and Information Technologies" (CSIT), Lviv, pp. 344–347, 2017.
- [16] I. Izonin, R. Tkachenko, D. Peleshko, T. Rak, and D. Batyuk, "Learning-based image super-resolution using weight coefficients of synaptic connections," Xth International Scientific and Technical Conference "Computer Sciences and Information Technologies" (CSIT), Lviv, pp. 25-29, 2015.
- [17] Y. Tsymbal, and R. Tkachenko, "A digital watermarking scheme based on autoassociative neural networks of the geometric transformations model," 2016 IEEE First International Conference on Data Stream Mining & Processing (DSMP), Lviv, pp. 231-234, 2016.
- [18] M. Nazarkevych, R. Oliiamyk, H. Nazarkevych, O. Kramarenko, and I. Onyshchenko, "The method of encryption based on Ateb-functions," 2016 IEEE First International Conference on Data Stream Mining & Processing (DSMP), Lviv, pp. 129-133, 2016.
- [19] I. Dronyuk., M. Nazarkevych, and Z. Poplavska, "Gabor filters generalization based on ateb-functions for information security," in Advances in Intelligent Systems and Computing, vol. 659, pp. 195-206, 2018

Selection the “Saturated” Block from Interval System of Linear Algebraic Equations for Recurrent Laryngeal Nerve Identification

Mykola Dyvak

*Faculty of Computer Information Technologies
Ternopil National Economic University
mdy@tneu.edu.ua*

Iryna Oliynyk

*Faculty of Computer Information Technologies
Ternopil National Economic University
ois@tneu.edu.ua*

Andriy Pukas

*Faculty of Computer Information Technologies
Ternopil National Economic University
apu@tneu.edu.ua*

Andriy Melnyk

*Faculty of Computer Information Technologies
Ternopil National Economic University
melnyk.andriy@gmail.com*

Abstract—The task of design a “saturated” experiment for measuring the characteristics of tissues in surgical wound in order to identify the recurrent laryngeal nerve (RLN) during operation on the neck organs considered in this paper. In this task, the method of selection a “saturated” block from an interval system of linear algebraic equations (ISLAE) is used, which allows to reduce the duration of surgical operation by decreasing the number of points for stimulation the surgical wound tissues to detect the RLN location and reduce the risk of its damage.

Keywords—neck surgery, recurrent laryngeal nerve, design of experiment, interval analysis, interval model.

I. INTRODUCTION

RLN monitoring is very important procedure during the neck surgery. For these purposes, special neuro monitors are used. They work based on the principle of surgical wound tissues stimulation and estimation of results of such stimulation [1-4]. However, these methods intended solely for the RLN monitoring.

The methods of RLN identification considered in papers [5, 6]. In particular, in the paper [5] the task of visualizing the RLN location based on evaluation the amplitude of signal as response to its stimulation by alternating current was considered. In paper [6] the method of constructing the difference schemes as a model for RLN location identifying based on interval analysis of response to stimulation the tissue in surgical wound by alternating current was considered.

It should be noted that the informative parameter in both methods used maximum amplitude of the signal as response to stimulation of tissues in surgical wounds, and as basis for determining the RLN damaging area assigned an interval model of distribution on surgical wound surface the maximum amplitudes of information signals as responses to stimulation the tissues in surgical wounds. However, both methods require creation the uniform mesh on surgical wound for tissues stimulation, which substantially increases the time of surgical operation.

However, in [7, 8], the methods of design of “saturated” experiments, aimed at providing guaranteed prognostic properties of models and requiring a minimum amount of

measurements, are considered. These methods based on selection a “saturated” block from the interval system of linear algebraic equations, which used for calculating the parameters of interval model of distribution on the surface of surgical wound of the maximum amplitudes of signals as responses on stimulation the surgical wound tissues. In our case, application of such methods will significantly shorten the surgical operation duration by reducing the time for RLN detection.

Therefore, the task of design of “saturated” experiments and interval data analysis for RLN identification is actual. Solving this task will accelerate rendering of RLN location, reduce the risk of its damage during operations on the neck organs and decrease the overall duration of the surgery.

II. TASK STATEMENT

The stimulation of surgical wound tissues during the neck surgery based on electrophysiological method allows identifying the type of tissue with the purpose of RLN identification.

In the papers [9], a method for RLN identification among tissues of a surgical wound is given. This method is based on stimulation the surgical wound tissues by alternating current with an active intensity from 0.5 to 2 mA and on registering the results of stimulation by sound sensor located above the vocal cords.

In respiratory tube 1 that inserted into larynx 2, the sound sensor 3 implemented and positioned above vocal cords 4.

Probe 5 is connected to stimulation block 7. It functions as a current generator controlled by the single-board computer 8. Surgical wound tissues are stimulated by the block 7 via probe 5. As a result, vocal cords 4 are stretched.

Flow of air that passes through patient’s larynx, is modulated by stretched vocal cords. The result is registered by sound sensor 3. Obtained signal is amplified and processed by single-board computer 8.

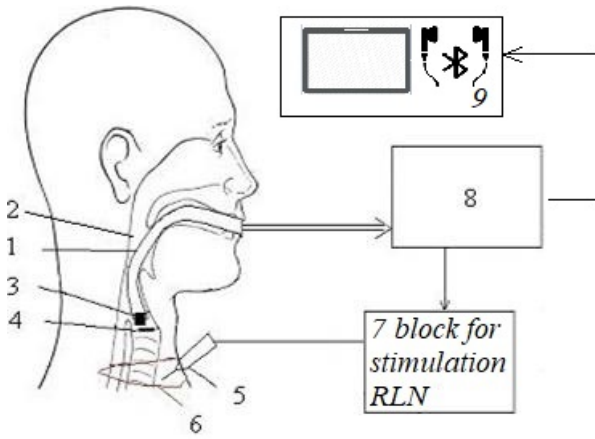
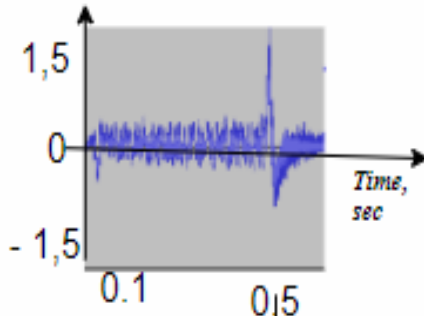


Fig. 1. Method of RLN identification among tissues in surgical wound.

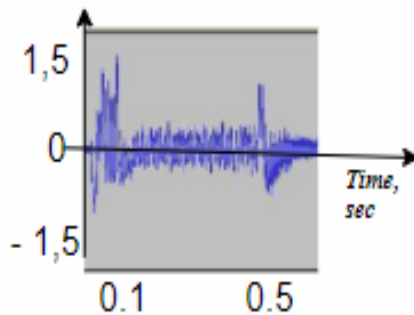
For processing of obtained signal, special software is installed on a single-board computer. The main functions of the software are:

- segmentation the information signal based on analysis of its amplitude;
- calculation the maximum amplitude of signal;
- classification the tissues in surgical wound at the points of stimulation [9].

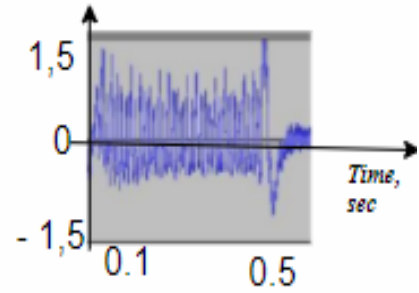
On Fig. 2 the fragments of amplified information signal which registered by sound sensor are shown.



a)



b)



c)

Fig. 2. Result of RLN stimulation by alternating current with frequency 300 Hz.

On Fig 2 a) it can be seen the result of stimulation the muscle tissue on distance more than 1 cm to RLN with minimal amplitude of information as reaction on stimulation. On Fig. 2 b) shown result of stimulation the muscle tissue on distance not more 3 mm. As we can see the signal as reaction on stimulation has higher amplitude then the signal as reaction on stimulation in previous case. Lastly on Fig. 2 c) shown result of RLN stimulation. As we can see the reaction has the highest amplitude.

Above mentioned research give possibility to assume that this characteristic can be used for RLN identification.

Let's represent the resulting set of points as follows:

$$[z(x_{1i}, x_{2i})], i=1, \dots, I, \quad (1)$$

where $[z(x_{1i}, x_{2i})]$ is an interval estimation of maximal amplitude of information signal fragment; x_{1i}, x_{2i} are increments of coordinate values on x_1 and x_2 axes relative to some initially given point. Interval estimation of amplitude $[z(x_{1i}, x_{2i})]$ caused by that for equal x_1, x_2 can be obtained different values of maximum amplitude $[z(x_{1i}, x_{2i})]$ of information signal.

Moreover some error of finding the point with coordinates x_1, x_2 exist. Let's denote by $i_o, o=1, \dots, O$ the points indexes in the vicinity of the point with x_1, x_2 coordinates. Lower and upper values of interval of maximal amplitude estimations of information signal are obtaining by Eqs.:

$$z^-_{i_o} = \min\{z_{i_o}, o=1, \dots, O\}; z^+_{i_o} = \max\{z_{i_o}, o=1, \dots, O\}.$$

Let's consider the mathematical model for RLN identification in kind of algebraic equation described in [5]:

$$\hat{z}(\vec{x}) = \beta_1 \cdot \varphi_1(\vec{x}) + \dots + \beta_m \cdot \varphi_m(\vec{x}), \quad (2)$$

where $\vec{\beta} = (\beta_1, \dots, \beta_m)^T$ is the vector of unknown parameters; $\vec{\varphi}^T(\vec{x}) = (\varphi_1(\vec{x}), \dots, \varphi_m(\vec{x}))^T$ is the vector of known basic functions; $\vec{x} = (x_1, x_2)$ is vector of stimulation point coordinates; $\hat{z}(\vec{x})$ is predicted value of maximal amplitude of information signal in point with coordinates (x_1, x_2) . Further, the model (2) will be called an interval model (IM).

Based on the requirements ensuring accuracy of the model within the accuracy of the experiment, the setting of IM (1) will realize with the using of such criterion [10, 11]:

$$[\hat{z}_i^-; \hat{z}_i^+] \subset [z_i^-; z_i^+], \forall i = 1, \dots, I. \quad (3)$$

By substituting to the expression (3), the recurrent expression (1) instead of the interval estimates $[\hat{z}_i^-; \hat{z}_i^+]$ together with the defined initial interval values of each we receive the following ISLAE [12]:

$$\begin{cases} z_1^- \leq b_1 \varphi_1(\bar{x}_1) + \dots + b_m \varphi_m(\bar{x}_1) \leq z_1^+; \\ \vdots \\ z_i^- \leq b_1 \varphi_1(\bar{x}_i) + \dots + b_m \varphi_m(\bar{x}_i) \leq z_i^+; \\ \vdots \\ z_N^- \leq b_1 \varphi_1(\bar{x}_N) + \dots + b_m \varphi_m(\bar{x}_N) \leq z_N^+; \end{cases} \quad (4)$$

where $\bar{b} = (b_1, \dots, b_m)^T$ is the vector of parameters $\bar{\beta} = (\beta_1, \dots, \beta_m)^T$ estimation.

III. METHOD OF SELECTION THE "SATURATED" BLOCK FROM ISLAE

The method of directed selection the "saturated" block built in accordance with the procedure of I_G -optimal design of experiment. The essence of this method consist in selecting a "saturated" block; calculating the corridor of interval models; analyzing the prognostic properties of these models, which influence on designing the way of next "saturated" block selection.

If structure of mathematical model of a static system is determined by the Eq. (2) with unknown parameters and interval data are given then ISLAE is created in kind (4).

Let's select from ISLAE arbitrarily a "saturated" block, calculate its area of solutions and construct a prediction corridor of interval models:

$$[\hat{z}(\bar{x})] = [\bar{\varphi}^T(\bar{x}) \cdot \bar{b} - \frac{1}{2} \cdot \Delta_{z(\bar{x})}; \bar{\varphi}^T(\bar{x}) \cdot \bar{b} + \frac{1}{2} \cdot \Delta_{z(\bar{x})}]. \quad (5)$$

Then by analogy with procedure of sequential I_G -optimal design of experiment from $\bar{x}_i, i = 1, \dots, N$ points for which ISLAE (4) created, will calculate the vector \bar{x}^{\max} which has maximal prediction error:

$$\begin{aligned} \bar{x}^{\max} &= \arg \max_{\bar{x}_i, i=1, \dots, N} \left\{ 2 \cdot \sum_{j=1}^m |\alpha_j(\bar{x}_i) \cdot \Delta_j|, \bar{x}_i, i=1, \dots, N \right\}, \\ \bar{\alpha}^T(\bar{x}_i) &= \bar{\varphi}^T(\bar{x}_i) \cdot F_m^{-1} \end{aligned} \quad (6)$$

It is significant that procedure (6) is simple as executed on finite set of points $\bar{x}_i, i = 1, \dots, N$. The vector obtained by Eq. (6) is a vector of input variables. This vector defines a certain interval equation in ISLAE (4). In accordance with procedure of sequential I_G -optimal design of experiment at this point, it is necessary to carry out the next measurement.

In paper [7] it is proved that if vector \bar{x}^{\max} coincides with vector of values the input variables in one from interval equations in the "saturated" block in ISLAE, then it specifies a point with a minimum value of prediction error. Hence, it is advisable to replace one interval equation in current "saturated" block by interval equation from ISLAE with vector of values the input variables \bar{x}^{\max} determined by Eq. (6). Thus, by analogy with the procedure of sequential I_G -optimal design, we "simulate" the procedure of additional measurement at a point \bar{x}^{\max} with the maximum error of prediction by the interval model. We will perform this procedure for each interval equation in "saturated" block. We obtain p ($p = 1, \dots, m$) new "saturated" blocks.

As a result, for each of the m "saturated" blocks we obtain m values of maximum errors for corresponding interval models:

$$\begin{aligned} \Delta_{\max}^p &= \max_{x_i, i=1, \dots, N} \left\{ 2 \cdot \sum_{j=1}^m |\alpha_{jp}(\bar{x}_i) \cdot \Delta_j| \right\}, \\ \bar{\alpha}_p^T(\bar{x}_i) &= \bar{\varphi}^T(x_i) \cdot F_m^{-1}(p), p = 1, \dots, m, \end{aligned} \quad (7)$$

where p is index, which means number of "saturated" block, $F_m(p)$ is matrix of base functions values for p block, $\alpha_{jp}(\bar{x}_i)$ is i -th component of vector $\bar{\alpha}$, that is calculated for p -th "saturated" block. Obviously, in order to choose the optimal "saturated" block in this step, it is enough to choose from m "saturated" blocks the one that provides the lowest value of sequence (7):

$$F_m^{opt} = \arg \min_{p=1, \dots, m} \left\{ \Delta_{\max}^p, p = 1, \dots, m \right\}. \quad (8)$$

We get \bar{x}^{\max} that is the vector for which the maximum prediction error for the interval model is reached. The scope of the parameters of this model is calculated from the "saturated" block chosen in the above-described method. Then the iterations continue until such a "saturated" block is obtained, the replacement of which equations does not lead to a decrease in the maximum prediction error by interval models. Localization method of solutions ISLAE allows to obtain explicitly guaranteed ellipsoidal estimation of ISLAE solutions [13]:

$$Q_m(k+1) = \left\{ \bar{b} \in R^m \mid (\bar{b} - \bar{b}(k+1))^T \cdot F^T \cdot E^{-2}(k+1) \cdot F \cdot (\bar{b} - \bar{b}(k+1)) = 1 \right\} \quad (9)$$

where $\bar{b}(k+1) = F_m^{-1} \cdot ((y_1^+(k+1) - y_1^-(k+1)), \dots, (y_m^+(k+1) - y_m^-(k+1)))^T$ is a vector that specifies the center of the ellipsoid; $E(k+1) = \text{diag}(y_1^+(k+1) - y_1^-(k+1), \dots, y_1^-(k+1) - y_1^+(k+1), \dots, y_m^+(k+1) - y_m^-(k+1), \dots, y_m^-(k+1) - y_m^+(k+1))$ is diagonal matrix of the differences of limits of intervals from the Eq. (9).

The corridor for interval models, which are defined by their predictive properties, in this case will look like this:

$$[\hat{y}(\bar{x})]_{\bar{b} \in Q_m} = [\bar{\varphi}^T(\bar{x}) \cdot \bar{b} - \frac{1}{2} \cdot \Delta_{\hat{y}(\bar{x})} \Big|_{\bar{b} \in Q_m}; \bar{\varphi}^T(\bar{x}) \cdot \bar{b} + \frac{1}{2} \cdot \Delta_{\hat{y}(\bar{x})} \Big|_{\bar{b} \in Q_m}] \quad (10)$$

where $\Delta_{\hat{y}(\bar{x})} \Big|_{\bar{b} \in Q_m}$ – the error of prediction (the width of the corridor), which is calculated by expression:

$$\Delta_{y(\bar{x})} \Big|_{\bar{b} \in Q_m} = \sqrt{\bar{\varphi}^T(\bar{x}) \cdot (F_m^T \cdot E^{-2} \cdot F_m)^{-1} \cdot \bar{\varphi}(\bar{x})} \quad (11)$$

IV. EXAMPLE OF APPLICATION THE PROPOSED METHOD FOR RLN IDENTIFICATION

Consider the example of constructing a model of distribution on the surface of a surgical wound the maximum signal amplitudes as reaction to stimulation of surgical wound tissues. The structure of this mathematical model, obtained from the work [5], has the following form:

$$\bar{z}(\bar{x}) = b_0 + b_1 \cdot \sin^2(x_1 \cdot x_2 \cdot \frac{\pi}{36}) + b_2 \cdot x_2 + b_3 \cdot (x_2^2)$$

A fragment of data obtained during the surgical operation on thyroid gland is given in Table 1 in [5]. We apply the method of selection the “saturated” block from ISLAE to find the most informative points. Using the described above method, interval model was obtained in form:

$$\begin{aligned} [\bar{z}(\bar{x})] &= [\bar{b}_0 + \bar{b}_1 \cdot \sin^2(x_1 \cdot x_2 \cdot \frac{\pi}{36}) + \bar{b}_2 \cdot x_2 + \bar{b}_3 \cdot x_2^2 - \Delta_{z(\bar{x})} \Big|_{\bar{b} \in Q_m}; \\ & \bar{b}_0 + \bar{b}_1 \cdot \sin^2(x_1 \cdot x_2 \cdot \frac{\pi}{36}) + \bar{b}_2 \cdot x_2 + \bar{b}_3 \cdot x_2^2 + \Delta_{z(\bar{x})} \Big|_{\bar{b} \in Q_m}] \end{aligned}$$

where $\bar{b} = (10.744; 36.81; 7.82; -1.06)^T$ is the vector of coordinates estimations of the ellipsoid Q_m center [13].

Prediction error $\Delta_{y(\bar{x})} \Big|_{\bar{b} \in Q_m}$ is represented for this case in such form:

$$\begin{aligned} \Delta_{z(\bar{x})} \Big|_{\bar{b} \in Q_m} &= \\ &= \sqrt{\begin{pmatrix} 1 \\ \sin^2(x_1 x_2 \frac{\pi}{36}) \\ x_2 \\ x_2^2 \end{pmatrix} \cdot (F_m^T \cdot \tilde{E}^{-2} \cdot F_m)^{-1} \cdot \begin{pmatrix} 1 \\ \sin^2(x_1 x_2 \frac{\pi}{36}) \\ x_2 \\ x_2^2 \end{pmatrix}^T}, \end{aligned}$$

where

$$F_m = \begin{pmatrix} 1 & 1 & 6 & 36 \\ 1 & 0.25 & 5 & 25 \\ 1 & 0.179 & 1 & 1 \\ 1 & 1 & 3 & 9 \end{pmatrix}$$

is the matrix of the basic function values of “saturated” optimized block;

$$\tilde{E} = \begin{pmatrix} 9.84375 & 0 & 0 & 0 \\ 0 & 5.6875 & 0 & 0 \\ 0 & 0 & 4.2131 & 0 \\ 0 & 0 & 0 & 10.7546 \end{pmatrix}$$

is modified diagonal matrix.

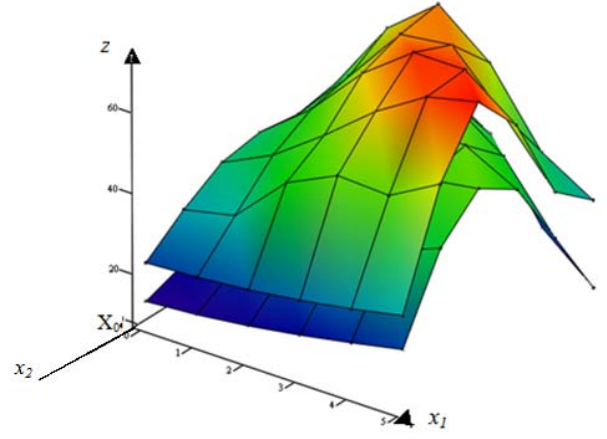


Fig. 3. Graph of distribution the maximal amplitude of information signal relatively to x_0 .

As we see on Fig. 3, instead of applying 36 points of stimulation of surgical wound tissues, it is enough to select 4 most informative points with coordinates [5; 1], [6; 3], [6, 5], [3; 6], with greater accuracy of estimating the maximum amplitude of the information signal.

V. CONCLUSIONS

The method of design of experiment based on new procedure of “saturated” block selection and its application to the task of measuring the characteristics of tissues in surgical wound in order to RLN identification during operation on the neck organs considered in this paper.

As distinct from existing method, in which for constructing the model of distribution the maximum amplitude of information signal it is necessary to use a sterile mesh, in case of application the method with selection the “saturated” block it is enough to choose m base points with relative to some point x_0 on surgical wound. The proposed method reduces the amount of stimulations of surgical wound tissues from m^2 to m , which significantly reduces the time spent on RLN identification, does not require the use of a sterile mesh, and thus reduces the time spent on surgical operation integrally.

ACKNOWLEDGMENT

This research was supported by National Grant of Ministry of Education and Science of Ukraine “Mathematical tools and software for classification of tissues in surgical wound during surgery on the neck organs” (0117U000410).

REFERENCES

- [1] M. C. D. Poveda, G. Dionigi, A. Sitges-Serra, M. Barczynski, P. Angelos, H. Dralle, E. Phelan and G. Randolph, “Intraoperative Monitoring of the Recurrent Laryngeal Nerve during Thyroidectomy: A Standardized Approach (Part 2),” *World Journal of Endocrine Surgery*, vol. 4, no. 1, pp. 33-40, 2012.
- [2] V. K. Dhillon, and R. P. Tufano, “The pros and cons to real-time nerve monitoring during recurrent laryngeal nerve dissection: an analysis of the data from a series of thyroidectomy patients,” *Gland Surgery*, vol. 6, no. 6, pp. 608-610, 2017.
- [3] H. Y. Kim, X. Liu, C. W. Wu, Y. J. Chai, and G. Dionigi, “Future Directions of Neural Monitoring in Thyroid Surgery,” *Journal of Endocrine Surgery*, vol. 17, no. 3, pp. 96-103, 2017.
- [4] W. E. Davis, J. L. Rea, and J. Templer, “Recurrent laryngeal nerve localization using a microlaryngeal electrode,” *Otolaryngology – Head and Neck Surgery*, vol. 87, no. 3, pp. 330-333, 1979.

- [5] M. Dyvak, O. Kozak, and A. Pukas, "Interval model for identification of laryngeal nerves," *Przeegląd Elektrotechniczny*, vol. 86, no. 1, pp. 139-140, 2010.
- [6] N. Porplytsya, and M. Dyvak, "Interval difference operator for the task of identification recurrent laryngeal nerve," 16th International Conference On Computational Problems of Electrical Engineering (CPEE), pp. 156-158, 2015.
- [7] M. Dyvak, and I. Oliynyk, "Estimation Method for a Set of Solutions to Interval System of Linear Algebraic Equations with Optimized "Saturated Block" Selection Procedure," *Computational Problems of Electrical Engineering*, Lviv, vol. 7, no. 1, pp. 17-24, 2017.
- [8] C. F. J. Wu, and M. S. Hamada, *Experiments: Planning, Analysis and Optimization*. Wiley, 2009.
- [9] M. Dyvak, N. Kasatkina, A. Pukas, and N. Padletska, "Spectral analysis the information signal in the task of identification the recurrent laryngeal nerve in thyroid surgery," *Przeegląd Elektrotechniczny*, vol. 89, no. 6, pp. 275-277, 2013.
- [10] Götz Alefeld, and Jürgen Herzberger, *Introduction to interval computations (Computer Science and Applied Mathematics)*. Academic Press, Inc. [Harcourt Brace Jovanovich, Publishers], New York, 1983.
- [11] S. P. Shary, "Algebraic Approach to the Interval Linear Static Identification, Tolerance, and Control Problems, or One More Application of Kaucher Arithmetic," *Reliable Computing*, vol. 2(1), pp. 3-33, 1996.
- [12] E. Walter, and L. Pronzato, *Identification of parametric model from experimental data*. London, Berlin, Heidelberg, New York, Paris, Tokyo: Springer, 1997.
- [13] A. Kurzhanski and I. Valyi, *Ellipsoidal Calculus for Estimation and Control*. Birkhauser, Berlin, 1997.

Optimized Concise Implementation of Batcher's Odd-Even Sorting

Paweł Tarasiuk
Institute of Information Technology
Lodz University of Technology
Lodz, Poland
pawel.tarasiuk@p.lodz.pl

Mykhaylo Yatsymirskyy
Institute of Information Technology
Lodz University of Technology
Lodz, Poland
mykhaylo.yatsymirskyy@p.lodz.pl

Abstract—The odd-even sort algorithm designed by Batcher [1] is a divide-and-conquer parallel sorting algorithm with $O(\log^2(n))$ delay time [2]. It is described in the literature [3] as very practical due to the potentially easy implementation, which is a notable advantage over AKS sorting networks [4]. However, the most basic literature on that topic contains either theoretical background without any implementations usable with the modern compilers [2] or C-like code snippets that are apparently erroneous [3]. In this paper, we propose an alternative to the code from [3] which improves compatibility with the C++ standard [5], fixes the bug that affected the computational complexity, and yields even further practical improvement to the execution time. All the specified enhancements are achieved without increasing the structural complexity of the method, so the proposed code remains as concise as the original.

Index Terms—algorithm design and analysis, computational complexity, data sorting, parallel algorithms, performance analysis

I. INTRODUCTION

The theory of sorting networks is based on a very basic observation – the efficient solutions to data sorting are likely to contain oblivious sequences of comparisons, where the order of specific operations does not affect the final result. This fact is especially easily justified when the subsets of processed elements associated to each comparison in the sequence are disjoint. This instantly leads to the conclusion, that such operations can be – to the extent which is technically possible with the considered hardware – performed at the same time. In the theoretical approach to the parallel sorting algorithms, the delay can be identified as the number of sorting network units – where each unit describes a set of oblivious comparisons.

The reliable theoretical introduction to the sorting networks can be found in the popular literature on algorithms and data structures such as [2], [3]. Specific algorithms for sorting networks with relatively small numbers of units were described by Batcher [1], and are known as odd-even sorting network and the bitonic sorting network. The first approach – odd-even sorting – is valued for the simplicity and potential for a concise implementation [3]. More complex ideas, such as AKS networks [4] or modern approach to the bitonic sorting [6], have the disadvantage of greater structural complexity. This is especially important in the modern applications such as the high-speed FPGA hardware [7], [8].

Another advantage of the sorting networks proposed by Batcher is the ability to sort any natural numbers of elements. Some other methods are limited to 2^k numbers of elements, k being a natural number. This is another reason why Batcher's networks can be suitable for the modern research [9].

II. DESCRIPTION OF THE ALGORITHM

The odd-even sorting algorithm is based on the merging networks proposed in [1]. Each merging networks consists of: merging the elements with odd indexes, merging the elements with even indexes – the corresponding units of both even- and odd-merge can be executed simultaneously, because the processed sets of elements are disjoint. The final stage is a single unit that consists of comparisons between each odd element (in terms of 0-indexing) and the consequent one. The Figure 1 presents the whole sorting network, where steps 4.1 - 4.4 constitute a merging network for two groups of 8 elements. The recursive reasoning described above explains the construction of 4.1 - 4.3 stages, which are equivalent to the merging network for two groups of 4 elements (either top or bottom half of 3.1 - 3.3) performed on both even and odd elements. The correctness proof for this merging procedure based on the Bouricius's theorem [10] is provided in [2].

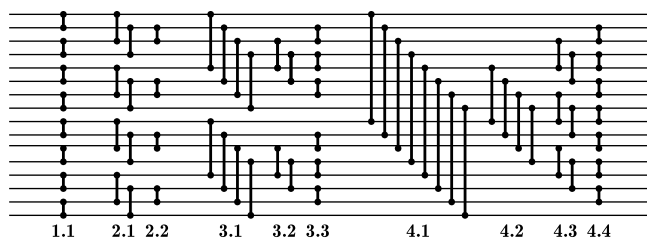


Fig. 1. The sorting network used for 16 elements. The simplicity of Batcher's odd-even sorting makes it especially easy to design arbitrarily large networks – each stage s from 1. to 4. consists of merging networks of size 2^{s-1} . The whole network consists of 10 units.

Sorting of 2^r -sized array consists of r merge stages. After each s -th stage, all the aligned chunks of 2^s elements are sorted internally – for $s = r$ it means sorting the whole array. This is a direct equivalent to the bottom-up construction of the mergesort algorithm [11].

The sorting of array with $n = 2^r$ elements takes $r = \lg n$ stages, where each r -th stage consists of r units. This yields the total number of $(\lg n \cdot (\lg n + 1))/2$ units, which belongs to the $O(\lg^2 n)$ class of parallel computational complexity.

The method described above does not concern array sizes n that are not whole powers of number 2. This problem, however, can be approached by using the network designed for $n' = 2^{\lceil \lg n \rceil}$ elements. Since $\lceil \lg n \rceil < \lg n + 1$, we conclude that $n' < 2n$, so the asymptotic complexity remains unchanged. Elements beyond the first n should be considered as equal to $+\infty$ – or, more practically – omitted alongside with associated comparisons, as any comparison with $+\infty$ leaves the first n elements unchanged. As it is presented in Figure 2, the reduced network contains some redundant comparisons, such as the last comparisons from groups 4.2 and 4.3. The network proposed for such n is not optimal, but remains efficient. In the practical implementation it is enough to make sure that no out-of-bounds memory access is performed on the processed array.

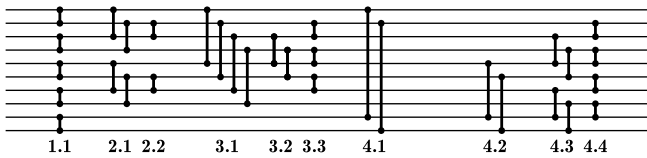


Fig. 2. The sorting networks for numbers of elements other than 2^s for some natural s can be designed by omitting comparisons that would end up out of the array bounds. This figure contains an example for 10 elements. Some of the comparisons are redundant, but the asymptotic complexity and simple code structure can be preserved.

III. OPTIMIZED IMPLEMENTATION

Listing 1

C++ CODE EQUIVALENT TO THE SNIPPET FROM [3]

```

1 for(int p = 1; p < n; p *= 2)
2   for(int k = p; k > 0; k /= 2)
3     for(int j = k % p; j + k < n; j += (k+k))
4       for(int i = 0; i < n-j-k; i++)
5         if((j+i)/(p+p) == (j+i+k)/(p+p))
6           if(A[j+i] > A[j+i+k])
7             std::swap(A[j+i], A[j+i+k]);

```

Let A be a container of size n , such as `std::vector` from the C++ standard. The code snippet taken almost directly from [3] is presented in Listing 1 – the only difference is replacing the obscure `complexx` call with conditional instruction and corresponding `std::swap` (lines 6-7). Should the single instruction for sorting a pair of elements exist (such as `CMPXCHG` [12]), the compiler can apply it to the presented code automatically, since it results in equivalent observable behavior [5].

The concept of this concise code is reasonable: the p -element groups of increasing size are being merged in the consequent stages. Each stage involves some comparisons of elements with distance k between the indexes (for $k = p, p/2, \dots, 1$), in order decreasing with k – groups of such comparisons are labeled in the Figure 1. In each group, we

can consider multiple merging networks that are supposed to work concurrently (actually we should get $n/(2p)$ merging networks in each stage. Let j be a starting index of the first comparison from the first merging network in the current group – which is 0 for the first group in the stage and k for every other group ($k \text{ MOD } p$ is a correct formula, as $k = p$ for the first group and $k < p$ for the others ones). For each merging network, some comparisons between $A[j+i]$ and $A[j+i+k]$ should be performed, starting from $i = 0$. As $A[j+i+k]$ should be a valid element, there is an obvious limitation $i < n-j-k$. What is more, no comparisons between different merging networks (each operating on $2p$ elements) should be performed, so if the network index for $j+i$ is different than for $j+i+k$, the comparison can be omitted (line 5).

While the whole description presented above is reasonable and the code from Listing 1 yields the correct result, the computational complexity of this implementation does not meet the expectations. The limitation $i < n-j-k$ describes some upper bound for i , but is overly loose. Instead of performing steps for the merging network corresponding with the j variable, the loop from line 4 iterates through all the consequent merging networks as well. This does not affect parallel complexity in the case of infinite concurrency, but some operations would be executed multiple times, which can cause a thread safety problem. With this code, the number of comparisons in each stage is not limited with $O(n \lg n)$ anymore, but with $O(n^2 \lg n)$. Thus, the whole non-concurrent sorting algorithm implemented like this can be estimated as $O(n^2 \lg^2 n)$.

Listing 2

CORRECT CODE WITH NO NECESSARY CONDITIONS

```

1 for(int p = 1; p < n; p *= 2)
2   for(int k = p; k > 0; k /= 2)
3     for(int j = k % p; j + k < 2*p; j += 2*k)
4       for(int i = 0; i < k; i++)
5         for(int m = i + j; m < n - k; m += 2*p)
6           if(A[m] > A[m+k])
7             swap(A[m], A[m+k]);

```

In order to fix the computational complexity, code from Listing 2 can be proposed. Some notations such as $p += p$ or $k + k$ were replaced by more natural notation, which includes multiplying by literal 2. Multiplication and division by a constant is optimized by the compiler anyways in the best architecture-related way that yields the same observable behavior. While the possible machine instructions for those tasks include bit shift operations, there is no need to denote them literally.

The most important improvement provided with this code is correctly calculated number of comparisons for each j , which is simply k (line 4). What is more, when the index of the first element taken into comparison is denoted as m , it becomes easier to jump through the merging networks without a specialized conditional statement. Iteration by j variable can be limited to merging network-related $2p - k$ elements, while increasing m by $2p$ till the end of array makes sure that all

the merging networks are included. What is more, using $n - k$ as a boundary for m ensures that no memory leaks will occur for any n (which is sufficient for the algorithm correctness).

This implementation performs $O(n \lg^2 n)$ comparisons in $O(\lg^2 n)$ units.

Listing 3
SIMPLE FIX WITH THE DECREASING LOOP

```

1 for(int p = 1; p < n; p *= 2)
2   for(int k = p; k > 0; k /= 2)
3     for(int j = k % p; j + k < n; j += 2*k)
4       for(int i = std::min(k, n-j-k); i--;)
5         if((j+i)/(2*p) == (j+i+k)/(2*p))
6           if(A[j+i] > A[j+i+k])
7             std::swap(A[j+i], A[j+i+k]);

```

Alternatively, the condition related to the different merging networks and the original meaning of j variable from Listing 1 can be preserved. The code presented in Listing 3 is possibly to the code from Listing 1, while limiting the number of iterations over i to no more than k . An important aspect is to keep the correctness for all n values intact, which makes the construction of the line 4 slightly tricky (the limit should be either k for complexity or $n - j - k$ for correctness). The proper solution can be achieved by using two conditions in the loop, or choosing the correct limit in every step. Instead of putting such a code in the loop stop condition, however, loop initialization can be sufficient in the case of a decreasing loop. The code from Listing 3 has all that properties, but it can be optimized even further.

Listing 4
THE PROPOSED OPTIMIZED CODE

```

1 for(int p = 1; p < n; p *= 2)
2   for(int k = p; k > 0; k /= 2)
3     for(int j = k & (p - 1); j + k < n; j += 2*k)
4       if((j | (2*p - 1)) == ((j+k) | (2*p - 1)))
5         for(int i = std::min(k, n-j-k); i--;)
6           if(A[j+i] > A[j+i+k])
7             std::swap(A[j+i], A[j+i+k]);

```

In the code from Listing 3, the test of a proper merging network is denoted in line 5. However, it can be observed that the comparison result is fixed for each j and never changes with i value. Since both j and $2p$ are divisible by k and i is lower than k , we conclude that $(j + i) \text{ DIV } (2p)$ is equal to $j \text{ DIV } (2p)$. As $j + k$ is divisible by k as well, the same remark goes to $(j + k + i) \text{ DIV } (2p)$ and $(j + k) \text{ DIV } (2p)$. Variable i can be ignored in that comparison.

In addition, as p is a variable, compiler cannot be easily aware that it will always be a whole power of 2 in the runtime. In the result, replacing the p -related division operations with bitwise operations can reduce the execution time even further. That is why in the Listing 4 the notation $k \text{ BITAND } (p - 1)$ is used instead of $k \text{ MOD } p$, and the comparison of integer division results by $2p$ is transformed to the form visible in line 4. Instead of ignoring the lowest $\lg p + 1$ bits (as the division is equivalent to the bit shift), all the bits are replaced with ones. Instead of a division or introducing an additional variable to perform the bit shift, the bitwise alternative operation is performed.

IV. EXPERIMENTAL RESULTS

In order to test the performance of the considered implementations, multiple tests on uniform random integer arrays were performed. Different array sizes n from 16 to 2^{22} were considered. All the 2^{2k} numbers from the considered range were selected from testing, and additional middle points were selected from the distribution defined by the reverse function of $n \lg^2 n$ – the actual values were calculated using the bisection method.

For each array size, three data sets with random integers from the uniform distribution were created. Since the scripts for the test data generation were implemented in Python programming language, the Mersenne twister algorithm [13] was used for pseudo-random number generation. For each data set, all the appropriate algorithms (Listing 1 for tests up to $n = 10\,000$ and Listings 2, 3 and 4 for all tests) were used, and the test with middle value of test times geometric mean for all the algorithms was selected. The collected data describes the “middle tests” in terms of average value and standard derivation of execution time resulting from 10 function calls. The testing scripts performed the correctness check as well, which was passed by all the algorithms in all the performed tests. Whenever it was possible, the test results were summed up by a formula corresponding with the theoretical computational complexity. The parameters were adjusted using the Marquardt-Levenberg algorithm for nonlinear least-squares that is included in Gnuplot graphing utility.

TABLE I
MEASURED TIME FOR RANDOM INTEGER ARRAYS: SMALL TESTS

n	L1		L2		L3		L4	
	t [μs]	σ	t	σ	t	σ	t	σ
256	1 020	60	10	2	26	1	8	1
601	6 900	200	30	2	90	2	33	1
1 024	19 900	200	50	1	160	10	60	10
1 920	77 700	500	120	10	350	10	130	10
2 954	206 000	2 000	217	2	630	20	222	1
4 096	394 000	5 000	300	10	850	10	310	10
6 775	1 163 000	2 000	570	20	1 650	50	560	30
9 752	2 610 000	10 000	1 040	20	2 650	60	890	20

Table I shows the results for small tests, with times displayed in microseconds. The difference between code from Listing 1 and the other algorithms is apparent – for $n = 9752$ the execution time exceeds 2.6s, which is above 1000 times slower than the other considered algorithms. Figure 3 describes the above-quadratic complexity of code from Listing 1. The complexity can be approximately described as $c_1 n^2 \lg^2 n$, where $c_1 = (1\,565 \pm 5) \cdot 10^{-13}$ [s].

More interesting experiments were possible to conduct on codes with proper asymptotic complexity, i.e. Listings 2, 3 and 4. The most expensive tests are described in Table II. The tests describe the cases of n greater than 3 200 000. Since all the considered methods yielded times greater than half a second and standard derivations greater than 5 ms, all the data in Table II was displayed in milliseconds.

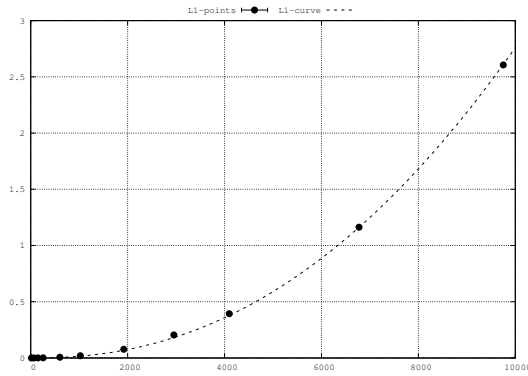


Fig. 3. Performance of algorithm from Listing 1 for different values of n . The experimental points fit the line $(1.565 \cdot 10^{-13}) \cdot n^2 \lg^2 n$.

TABLE II
MEASURED TIME FOR RANDOM INTEGER ARRAYS: BIG TESTS

n	L2		L3		L4	
	t [ms]	σ	t [ms]	σ	t [ms]	σ
3 251 674	2 410	10	2 080	40	670	5
3 384 012	2 570	15	2 130	25	710	20
3 517 161	2 730	25	2 210	10	730	10
3 651 097	3 430	50	2 320	40	730	10
3 785 800	3 500	10	2 380	10	760	10
3 921 248	3 900	100	2 460	20	780	10
4 057 423	4 000	100	2 570	25	840	30
4 194 304	4 130	30	2 650	10	850	5

In the case of Listing 3, the experimental points fit the line $c_3 \times n \lg^2 n$ where $c_3 = (1344 \pm 4) \cdot 10^{-12}$ [s]. Similar formula can be provided for Listing 4, where the line can be described as $c_4 \times n \lg^2 n$ where $c_4 = (432 \pm 2) \cdot 10^{-12}$ [s]. This means that the code from Listing 4 is about three times faster than the code from Listing 3. Code from Listing 2 behaves in much less deterministic and monotonous way, which makes the errors too high to actually fit the formula. However, the plot from Figure 4 and the results from Table II strongly indicate that this code is slower than the two other considered ones for arrays larger

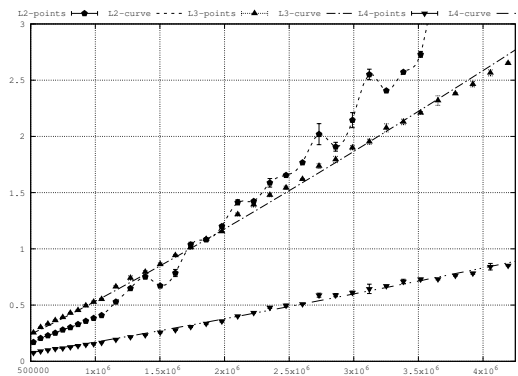


Fig. 4. Plots of experimental points and according curves for Listings 2, 3 and 4. Results for Listing 2 are difficult to fit – cspline-based line was displayed instead. The experimental points for Listing 3 fit the line $(1.344 \cdot 10^{-12}) \cdot n \lg^2 n$, and points for Listing 4 fit the line $(432 \cdot 10^{-12}) \cdot n \lg^2 n$.

than 2 000 000 elements.

V. CONCLUSIONS AND FUTURE WORK

The Batcher’s odd-even sorting algorithm – while based on the simple concept of mergesort – has a great potential for theoretical analysis and technical works on the software implementation. The actual applications are related with hardware that performs the task with actual concurrency, which includes modern works on using this algorithm in FPGA hardware [7], [8]. Another direction of the further research can be achieved with easily accessible hardware as well – the proposed code can be adjusted for OpenMP environment or executed on GPU hardware. While the obvious importance of asymptotic complexity was demonstrated on the code from Listing 1, the differences between other proposed implementations are significant. Similar analysis of technical details can be performed for the other architectures as well.

REFERENCES

- [1] K. E. Batcher, “Sorting networks and their applications,” in *Proceedings of the April 30–May 2, 1968, Spring Joint Computer Conference*, ser. AFIPS ’68 (Spring). New York, NY, USA: ACM, 1968, pp. 307–314.
- [2] D. E. Knuth, *The Art of Computer Programming Volumes 1-3 Boxed Set*, 2nd ed. Boston, MA, USA: Addison-Wesley Longman Publishing Co., Inc., 1998.
- [3] R. Sedgewick, *Algorithms in C++*. Boston, MA, USA: Addison-Wesley Longman Publishing Co., Inc., 1992.
- [4] M. Ajtai, J. Komlós, and E. Szemerédi, “Sorting in $c \log n$ parallel steps,” *Combinatorica*, vol. 3, no. 1, pp. 1–19, Jan. 1983.
- [5] ISO/IEC, “ISO International Standard ISO/IEC 14882:2017(E) – Programming Language C++,” 2017. [Online]. Available: <https://isocpp.org/std/the-standard>
- [6] J.-D. Lee and K. E. Batcher, “Minimizing communication of a recirculating bitonic sorting network,” in *Proceedings of the 1996 ICPP Workshop on Challenges for Parallel Processing*, vol. 1, Aug 1996, pp. 251–254 vol.1.
- [7] Y. Jun, L. Na, D. Jun, G. Yixiong, and T. Zuoxia, “A research of high-speed batcher’s odd-even merging network,” in *2010 International Conference on E-Health Networking Digital Ecosystems and Technologies (EDT)*, vol. 1, April 2010, pp. 77–80.
- [8] R. V. W. Putra, “Vlsi design of parallel sorter based on modified pem algorithm and batcher’s odd-even mergesort,” in *International Conference on ICT for Smart Society*, June 2013, pp. 1–5.
- [9] M. Ouyang, “Sorting sixteen numbers,” in *2015 IEEE High Performance Extreme Computing Conference (HPEC)*, Sept 2015, pp. 1–6.
- [10] W. G. Bouricius and J. M. Keller, “Simulation of human problem-solving,” in *Papers Presented at the March 3-5, 1959, Western Joint Computer Conference*, ser. IRE-AIEE-ACM ’59 (Western). New York, NY, USA: ACM, 1959, pp. 116–119.
- [11] J. Katajainen and J. L. Träff, “A meticulous analysis of mergesort programs,” in *Algorithms and Complexity*, G. Bongiovanni, D. P. Bovet, and G. Di Battista, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 1997, pp. 217–228.
- [12] Intel Corporation, *Intel® 64 and IA-32 Architectures Software Developer’s Manual*, September 2016, no. 325383-060US.
- [13] M. Matsumoto and T. Nishimura, “Mersenne twister: A 623-dimensionally equidistributed uniform pseudo-random number generator,” *ACM Trans. Model. Comput. Simul.*, vol. 8, no. 1, pp. 3–30, Jan. 1998.

Topic #4

Machine Vision and Pattern Recognition

Core Generator of Hypotheses for Real-Time Flame Detecting

Dmytro Peleshko,
IT Step University,
Lviv, Ukraine
dpeleshko@gmail.com

Orysia Voloshyn
Lviv National Polytechnic University
Lviv, Ukraine
vop_ippt@ukr.net

Oleksii Maksymiv
*Department of Information Security
Management*
Lviv State University of Life Safety
Lviv, Ukraine
aleks.maksymiv@gmail.com

Taras Rak,
IT Step University,
Lviv, Ukraine
rak.taras74@gmail.com

Bohdan Morklyanyk
Lviv National Polytechnic University
Lviv, Ukraine
Morklyanykb@gmail.com

Abstract—The flame is a visually unstable and constantly changeable process, which causes considerable difficulties for its detection in the video streams. Although the modern architecture of convolutional neural networks can show high accuracy, their integration into real-time systems is problematic, because they require a large amount of computing resources. To reduce the number of these resources, it is proposed to select possible regions of interest (ROI), which are based on the developed generator of hypotheses. Compared to existing flame detection algorithms, the developed generator of hypotheses allows you to work with the minimum of computing resources and has a high degree of classification completeness due to improved methods of color segmentation and moving objects detection.

Keywords—computer vision, machine learning, flame detection, color segmentation

I. INTRODUCTION

Fire is one of the most dangerous types of emergencies because in the time gap the process of its localization and liquidation is usually more complicated, requiring more and more time and human resources. This, in its turn, makes it necessary to reveal signs of a fire at an early stage and to notify the relevant agencies and units, in order to make effective decisions on minimizing possible losses.

All this can be done by using automatic fire detectors that can detect signs of fire at its early stage. However, according to statistics provided by the Ukrainian Research Institute of Civil Protection, in 2014, 7.9% of the total number of fire protection systems was corrupted, 12.7% of their technical resources worn out and are a subject to replacement, and only 79,8 % from working plants are having scheduled maintenance. In general, these indicators continue to decrease, as the problem of high-quality technical maintenance of fire automatics systems requires a significant amount of both time and material resources.

This is unacceptable since there is a risk of material losses or even human health and life risk. Accordingly, there is the question of searching more universal control means over the fire occurrence or those ones that can improve fire safety index.

For this purpose, we have proposed technology of video surveillance, which, unlike the traditional automatic fire detectors, has several advantages:

- the ability to detect flames in the open area;
- the possibility of controlling large space with a minimum number of video cameras;
- camera location, unlike some types of traditional detectors, should not be near the fire source;
- the possibility of checking the correctness and identification of the immediate place of ignition;
- minimum time latency for fire detection;
- detecting the location of the fire and its source with much ease.

The described approach involves the use of traditional computer vision techniques to develop a basic generator of hypotheses. The specified generator allows detecting areas (segments) of an image or video stream, which can be taken as a «flame» by their visual features. In contrast to existing works, we aim to achieve the highest possible index of correct operations integrity rather than accuracy, providing maximum performance at the same time. The next stage of the development system for fire detection is the transmission of received ROI for further analysis by the convolutional neural network, which will provide the necessary accuracy of the verification.

II. RELATED WORKS

The majority of methods used to detect flames are based on the use of information about color, motion, or both of these characteristics at the same time.

Regarding the use of color as a way of flame features detection, color models such as RGB [1], [2] HSV [3], YCbCr [3], [4], YUV [5], [6] or their combinations [7] are widely used.

To detect moving objects in a video stream, a variety of methods based on spatial or temporal features are used to reveal possible changes. Among the existing methods, three most used and effective approaches can be pointed out: background subtraction [8], frame difference [9] and optical flow [10].

In addition to the above-mentioned methods for flame detecting in video sequences, there are some techniques which allow to develop classification based on wavelet

analysis [11], local binary patterns [12], hidden Markov models [13] etc.

III. MAIN RESTRICTIONS IN THE FLAME DETECTION TASK

Implementation of set tasks requires many efforts aimed at solving certain contradictions that may arise between theory and practice.

So, from the practical side, first of all, it is necessary to achieve high rates of correctly recognized target objects (True positive (TP)) and to minimize the number of calls on the objects that do not belong to requested category (True negative (TN)). This can only be achieved by minimizing false positive rate (FP) and objects that are related to the requested search category but were ignored by the detector (False negative (FN)).

At the same time, at the present stage of the machine learning algorithms development, the task of detecting flames in a video stream is not solved in full. Obtaining work efficiency sufficient for the system of life safety is possible only in the conditions of significant increase in the number of computing resources and that is quite difficult in modern realities. Without necessary resources, time delays might arise, which are unacceptable in the systems of such types, taking into account its specificity.

In particular, a number of external and internal factors can affect the quality of detection. Accordingly, the developed system should take into account:

- physical parameters: type of ignition source, type of combustible substance;
- environmental conditions: general weather conditions (fog, rain, etc.), visibility, time of day;
- camera placement: angle of inclination, height, observation area.

The main restrictions that we plan to cope with are the following:

1. Unknown conditions of observation. In order to ensure the versatility of flame detection system work, it is necessary to take into account the environment in which the observation is carried out (corridor, street, parking, etc.). At the same time, the conditions of observation at different facilities will also differ (digital stream quality, an angle of inclination and height of camera placement, etc.).
2. Peculiarities of combustion. In this case, it is considered expedient to take into account the presence of combustible substances and possible temperatures, which may lead to a change of flame color. Object overlays can also lead to detector's overlooking the ignition cases; therefore, it is necessary to pay attention to the unstable form of fire, especially the variability of dimensions.
3. Possibility of additional learning. Considering the above limitations, work of the proposed flame detector, especially at the initial stage, may be accompanied by incorrect triggers. That is why it is important to conduct continuous training, taking into account previous mistakes, thus avoiding their repetition in the future.

IV. CORE GENERATOR OF HYPOTHESES ABOUT FLAME PRESENCE IN VIDEO STREAM

Core generator of hypotheses was developed taking into account the main features of the flame as a research object, which allows generating so-called regions of interest (ROI). The generator includes the following operations: pre-processing, color segmentation and detection of moving objects. Compared to existing flame detection algorithms, developed generator of hypotheses has got an improved method of color segmentation and instead of classical methods of detecting moving objects a three-frame difference method is proposed.

A. Color segmentation

One of the basic ways that allows efficient visual separation of the flames from other objects is its color. Segmentation of a color-based image implies the use of certain boundaries for some color model or a set of models. Despite a number of works devoted to the color models for flames detecting in images, and rather high-performance indicators, still there is a question about the type of color model which allows classifying pixel as a flame. In this context, particular attention should be paid to the fact that in comparison with other studies, we are interested in the classification recall, not accuracy.

To solve this problem, it was decided to independently analyze the efficiency of segmentation recall by using four color models: RGB, HSV, CIE L*a*b, ra YCbCr.

For this purpose, 150 images with fire were taken. These images are completely variational. They are characterized by different resolution, noise level, illumination, shooting conditions, etc.

To select the boundaries in which the flame could be located the method of k-means clustering was used. In the general case, clustering allows you to divide the set of input vectors to groups according to their degree of similarity. The main purpose for using k-means method is minimizing of mean-square deviation at the points of each cluster.

The disadvantage of this approach is the need to independently select the number of clusters to achieve the most effective result. An example of the original images on which experimental studies were conducted is shown in Fig. 1. In most cases, the use of 4 clusters was sufficient (Fig. 2), but there were cases where the number of clusters had to be increased up to 8 (Fig. 3).



Fig. 1. Example of an analyzed images

Using the method of k-means, areas of the image in which the flame is located were obtained (last images on Fig. 2 and 3). The obtained results allowed to evaluate the correlation between different color channels. It allows determining color limits for flame locating. This, in turn, allowed deriving the rules under which the pixels of the image should be verified on belonging to «flame» category,

since they may have a certain similarity to other objects (red t-shirt, sun rays, etc.).



Fig. 2. Result of using k-means in an image with four clusters

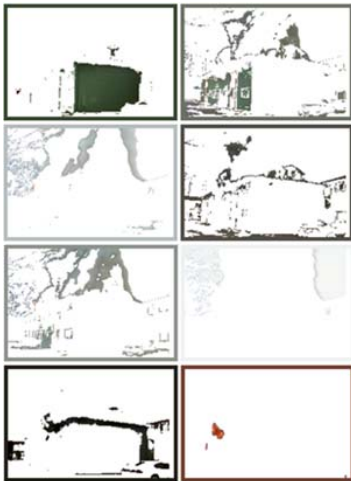


Fig. 3. The result of applying the k-means on an image with eighth cluster

Summarizing the experimental results, we can assert that the color models HSV, YCbCr and L*a*b showed good results during image segmentation. However, it was decided to use the model L*a*b, since it allows to provide not only effective split regions that are not similar to fire, but also those that can be referred to as "flames" in other models. Examples of segmented images are shown in Fig. 4.



Fig. 4. Examples of image segmentation using the color model L*a*b

B. Detecting only moving objects

Fire is characterized by a constant change in shape and boundaries. So, there is a possibility to minimize the number of false alarm by separating moving objects or comparing the flame behavior in time with the detected objects. Despite existing methods for detecting motion on video, it is important to consider the behavior of fire, which will reduce the number of false calls.

The proposed approach for detecting moving objects in a video sequence is based on the frame difference method. This method involves detecting changes between two sequential frames of the video stream. It is done by the elementary pixel subtraction of the current frame from the next frame. Mathematically this process is described:

$$D(x, y, t+1) = \begin{cases} 1, & |f(x, y, t) - f(x, y, t+1)| > T; \\ 0, & |f(x, y, t) - f(x, y, t+1)| \leq T. \end{cases} \quad (1)$$

where $f(x, y, t)$ is the frame of the video sequences in time t ; $f(x, y, t+1)$ is the next frame in time $t+1$; T is threshold value for decision.

However, analyzing the existing video dataset, it has been experimentally found that information obtained by comparing only two frames (current and next) is not sufficient for accurate verification of fire presence or absence in a video stream. In order to solve this problem, it was proposed to use a three-frame difference method, which allowed taking into account additional motion features of the object, thus enabling to provide a more accurate assessment of its belonging to a certain class. In the basic implementation of frame difference method, the difference between the previous and current frame is calculated as:

$$D_k(x, y) = |f_k(x, y) - f_{k-i}(x, y)|, i = 1, 2, \dots, \quad (2)$$

In turn, image binarization is calculated by the following formula:

$$R_k(x, y) = \begin{cases} 0, & D_k(x, y) \leq T; \\ 1, & D_k(x, y) > T. \end{cases} \quad (3)$$

The method of the three-frame difference analyzes not only the current and the previous frame, but also the following ($t+1$):

$$D_1(x, y) = \begin{cases} 1, & |f_k(x, y) - f_{k-1}(x, y)| \geq T; \\ 0, & |f_k(x, y) - f_{k-1}(x, y)| < T; \end{cases} \quad (4)$$

$$D_2(x, y) = \begin{cases} 1, & |f_{k+1}(x, y) - f_k(x, y)| \geq T; \\ 0, & |f_{k+1}(x, y) - f_k(x, y)| < T. \end{cases} \quad (5)$$

C. Morphological processing

In order to eliminate the artifacts that can be obtained as a result of using two above-mentioned algorithms, it is assumed that the use of morphological operations is necessary. The basis of morphological operations is a principle of transforming an image into another image by using a function or a set of functions that allow taking into consideration geometric information of an object. Such information allows getting a more accurate description of the form and size of objects that are obtained as a result of segmentation. In particular, morphology can be used to

reduce noise, identify borders, analyze textures and shapes, etc. There are four most common morphological operations: erosion, dilation, opening and closing, the symbol of which looks like the following \square , \ominus , \circ and \diamond for each operation, respectively. Mathematically, they are described as follows:

$$X \oplus H = \{(x, y) : H_{(x,y)} \cap X \neq \emptyset\}; \quad (6)$$

$$X \ominus H = \{(x, y) : H_{(x,y)} \subseteq X\}; \quad (7)$$

$$X \circ H = (X \ominus H) \oplus H; \quad (8)$$

$$X \diamond H = (X \oplus H) \ominus H, \quad (9)$$

where X is original image; $H \subseteq \mathbf{R}^2$ is structure element; $H_{(x,y)}$ is plural H to vector $(x, y) \in \mathbf{R}^2$.

Based on the foregoing, we can generalize application effect for each of the morphological operations. So, erosion can reduce the area of the image, dilation, on the contrary, expand it. Operation of opening removes the boundary protrusions at the object boundaries, and the closure fills the possible openings inside the segmented object, which is usually accompanied by an increase in its contour. During the development of ROI generation method, it was suggested to use binarization and opening to eliminate very small areas (visually cannot provide any information) and possible artifacts of the image (especially low-resolution video cameras). An example of noise reduction is shown in Fig. 5.



Fig. 5. Examples of image segmentation using the color model L^*a^*b

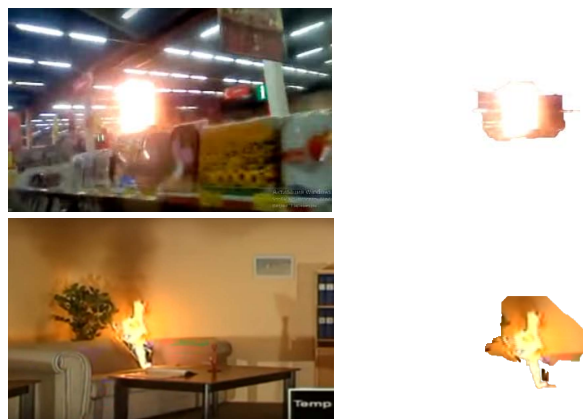


Fig. 6 The result of hypotheses generator works on video sequences in real time

V. EXPERIMENTS

To obtain experimental results a self-developed video dataset was used. It consists of 10 video sequences with flame or objects that can visually resemble a flame. Total video duration is 16 min 46 sec. Accuracy rate is 64.3%,

completeness rate is 99.9%. An example of the ROI generation is shown in Fig. 6. In the image on the left there is the frame received from a video sequence, on the right there is the result of its processing by generator of hypotheses.

VI. CONCLUSION

The generator of hypotheses is developed to select an area of the image that according to its visual appearance or time characteristics may resemble a flame. The first priority we set was to get the biggest index of completeness. Accordingly, although there are objects that can be mistakenly taken as «flame», we were able to greatly minimize the size of the image areas, which will need to be further verified by the classifier with the recall of 99.9%. In order to improve the accuracy of the generator of hypotheses it is considered expedient to use convolutional neural networks further on.

REFERENCES

- [1] P. Patel, and S. Tiwari, "Flame Detection using Image Processing Techniques", *Int. J. Comput. Applic.*, vol. 58, no. 18, pp. 13-16, 2012.
- [2] Thou-Ho Chen, Cheng-Liang Kao, and Sju-Mo Chang, "An intelligent real-time fire-detection method based on video processing," *IEEE 37th Annual 2003 International Carnahan Conference on Security Technology*, 2003. Proceedings., pp. 104-111, 2003.
- [3] Norsyahirah Izzati binti Zaidi, Nor Anis Aneza binti Lokman, Mohd Razali bin Daud, Hendriyawan Achmad and Khor Ai Chia, "Fire recognition using RGB and YCbCr color space," *ARNP Journal of Engineering and Applied Sciences*, vol. 10, no. 21, pp. 9786-9790, November 2015.
- [4] Turgay Çelik, Hüseyin Özkaramanlı, and Hasan Demirel, "Fire and smoke detection without sensors: image processing based approach," *15th European Signal Processing Conference*, pp. 1794-1798, 2007.
- [5] L. Rossi, M. Akhloufi, and Y. Tison, "On the use of stereovision to develop a novel instrumentation system to extract geometric fire fronts characteristics," *Fire Safety Journal*, vol. 46, no. 1-2, pp. 9-20, 2011.
- [6] C. Emmy Prema, S. S. Vinsley, and S. Suresh, "Multi Feature Analysis of Smoke in YUV Color Space for Early Forest Fire Detection," *Fire Technology*, vol. 52, no. 5, p. 1319-1342, 2016.
- [7] Thou-Ho (Chao-Ho) Chen, Ping-Hsueh Wu, and Yung-Chuen Chiou, "An Early Fire-Detection Method Based on Image Processing," in *Proc. Int. Conf. Image Process. (ICIP)*, vol. 3., pp. 1707-1710, Oct. 2004.
- [8] Guruh Fajar Shidik, Fajrian Nur Adnan, Catur Supriyanto, Ricardus Anggi Pramunendar, and Pulung Nurtantio Andono, "Multi color feature, background subtraction and time frame selection for fire detection," *2013 International Conference on Robotics, Biomimetics, Intelligent Computational Systems*, pp. 115-120, 2013.
- [9] T. Celik, "Fast and efficient method for fire detection using image processing," *ETRI journal*, vol. 32, no. 6, pp. 881-890, 2010.
- [10] Suchet Rinsurongkawong, Mongkol Ekpanyapong, and Matthew N. Dailey., "Fire detection for early fire alarm based on optical flow video processing," *9th International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology*, pp. 1-4, 2012 .
- [11] B. U. Toreyin, Y. Dedeoglu, and A. E. Cetin, "Wavelet based real-time smoke detection in video," *Signal Processing Conference*, 13th European, pp. 1-4, 2005.
- [12] Zhao-Guang Liu, Yang Yang, and Xiu-Hua Ji, "Flame detection algorithm based on a saliency detection technique and the uniform local binary pattern in the YCbCr color space," *Signal, Image and Video Processing*, vol. 10, no. 2, pp. 277-284, 2016.
- [13] B. Toreyin, Y. Dedeoglu, A. Cetin, "Flame detection in video using hidden markov models," *IEEE International Conference on Image Processing*, vol. 2, pp. 213-216, 2005.

Shallow Convolutional Neural Networks for Pattern Recognition Problems

Oleksii Gorokhovatskyi
Informatics and Computer Technologies
Simon Kuznets Kharkiv National University of Economics
Kharkiv, Ukraine
oleksii.gorokhovatskyi@gmail.com

Olena Peredrii
Informatics and Computer Technologies
Simon Kuznets Kharkiv National University of Economics
Kharkiv, Ukraine

Abstract—Paper describes an investigation of possible usage of shallow (limited by few layers only) convolutional neural networks to solve famous pattern classification problems. Brazilian coffee scenes, SAT-4/SAT-6, MNIST, UC Merced Land Use and CIFAR datasets were tested. It is shown that shallow convolution neural networks with partial training may be effective enough to produce the result close to state-of-the-art deep networks but also limitations are found.

Keywords— *image; recognition; classification; convolution; shallow neural network; layer; partial training; dataset*

I. INTRODUCTION

Artificial neural networks (ANN), deep neural networks (DNN) and convolutional neural networks (CNN) last decades [1] became one of the most effective ways to resolve complex pattern recognition, classification and machine learning problems because of their power and huge flexibility. Besides reaching high accuracy ANN have two basic problems, first one is related to the choice of network structure that is effective enough, the other one is related with the requirement to have powerful hardware to train network. Structures of state-of-the-art CNN like PatreNet [2], AlexNet [3], CaffeNet [4] (based on AlexNet architecture), VGG [5], GoogLeNet [6] became computationally complex last years, that leads to situation when solving of challenging pattern recognition problems according to a lot of publications seems to be possible only with usage of special hardware or optimization routines like GPU calculations and usage of small (shallow) CNN is underestimated.

The idea of paper is to investigate if some well-known image classification problems may be resolved with shallow CNNs only, which are possible to train and use just on a typical personal computer without special hardware and GPU calculations.

II. SHALLOW NETWORK STRUCTURE

One of the known problem with the implementation of a network is related to the huge amount of calculations, which may require a lot of time. This may be resolved with switching to GPU, that allows increasing speed significantly. Another problem is the complex net structure that may need a lot of memory. Finally, deep structure requires more parameters (like kernel size, quantity of neurons etc.) to be defined somehow. On the other hand, shallow NN may have more advantages to be used on mobile devices.

We are going to pay attention to such NN architectures and such training procedures that are possible to fit in memory at least partially and train in a reasonable amount of time.

A. Quantity of Layers

Looking at modern ANN like AlexNet [3] or GoogLeNet [6] or Microsoft ResNet [7] we may notice the huge amount of layers (from dozens up to hundreds) that require training time within days, weeks or even months using parallel GPU-calculations [8]. It is easy to find papers that describe the extremely effective solution of known pattern recognition problems using ANN mentioned above, but that doesn't mean that it is always reasonable to use NN with millions of parameters.

In order to make such CNN that is possible to train (within hours) and use (within seconds), we will consider shallow architectures that are limited by only ten layers including dropout ones.

B. Layer Types

The structure of traditional convolutional neural network is usually built as a sequence of layers of different neuron types that allows performing of specific operations on each stage of image processing. CNN usually consists of convolution, maxpooling, dropout and dense layers.

Convolution operator in computer vision and image processing problems is mostly used as a filtration layer that allows retrieving of specific features of an image. Let us denote image as I ($(w)width \times (h)height$ size) and convolution kernel as K with typically small square odd size (3×3 or 5×5 are the most popular) of a kernel k . Mathematically, convolution result is represented as:

$$I * K = \sum_{i=1}^k \sum_{j=1}^k K_{i,j} I_{i+x-1, y+j-1},$$

$$\text{where } i = \lfloor k/2 \rfloor, w - \lfloor k/2 \rfloor, j = \lfloor k/2 \rfloor, h - \lfloor k/2 \rfloor$$

Convolution means scanning of an image pixel by pixel overlaying with kernel window and computation of new values of convolved image with $(w-k+1) \times (h-k+1)$ size. Different kernel K values allow to apply the variety of specific filters like sharpening, blurring, edge detection etc. [9]. Additionally, convolution process may utilize other special parameters like stride or padding [10].

Maxpooling is a popular downsampling approach that applies some aggregation (maximization operator mostly) to image parts to leave only the most valuable values. A typical size of such parts is 2×2 , so each non-overlapping 2×2 part of image is replaced by a maximum of all values in this part.

Dense layers contain a combination of fully-connected neurons and usually used at the end of CNN structure to gather and generalize features after convolution and maxpooling layers. Training of dense layers is slower compared to other types, so only one or two such layers seems reasonable to use.

Dropout layers are used to prevent overfitting and speed up training process and the idea is to set zero values to some quantity of random input neurons. Exactly half of such inputs are dropped for this paper during training of a network.

C. Activation Function and Optimizer

A lot of different activation functions exist, but here only two types are used. Neurons in internal layers use rectifier activation according to

$$f(x) = \max(0, x),$$

where x is the weighted biased input from the previous neuron. Such activation is a good choice to make training faster because of simple gradient and somewhat more effective due to zero reaction to negative inputs.

Last dense layers use sigmoid activation

$$f(x) = \frac{1}{1+e^{-x}}$$

to produce output in the range between 0 and 1.

Default options $\alpha = 0.001, \beta_1 = 0.9, \beta_2 = 0.999, \varepsilon = 10^{-8}$ of Adam [11] optimizer were used for modeling. Stochastic gradient descent was tested too but Adam optimizer outperformed it easily both from speed and performance points of view.

III. RESULTS OF MODELLING

Usage of some shallow CNN architectures was tested on popular datasets defined below. Effective solutions for datasets A-C were found relatively easily, a lot of models were tested for each dataset, especially on those given in D section.


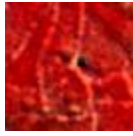







The modeling methodology was different dependently of original dataset structure, k-fold cross-validation was used for those which are already split to folds. Accordingly, if the dataset is split just on “Train” and “Test” part they were used for training and testing as a whole.

A. Brazilian Coffee Scenes

Brazilian Coffee Scenes dataset was proposed in [4, 12, 13] and it consists of 2876 RGB-NIR (Near-InfraRed) images of coffee and noncoffee plantations size of 64×64 pixels. Data was split by creators to 4 folds with 600 images each and the 5th fold with 476 images. All folds are balanced with coffee and noncoffee samples (50% each). We are focused on usage of only RGB channels and near-infrared channel was ignored.

This dataset is very challenging because of high interclass variance, different colorization of coffee regions and presence of distortions like shadows. Table I shows examples of coffee and noncoffee images (first and second row respectively) as well as samples of images that look pretty similar but belong to different classes (“coffee”-labeled images are above, similar noncoffee images are below).

TABLE I. EXAMPLES OF BRAZILIAN COFFEE SCENES DATASET IMAGES

Class label	Sample images		
Coffee			
Noncoffee			
Coffee Vs Noncoffee			

Some known results of this dataset recognition based on different technologies are gathered in Table II. Best results are related to CaffeNet, that includes 5 convolution layers, 3 maxpooling and 2 dense layers, or GoogLeNet, that contains 22 layers.

TABLE II. BRAZILIAN COFFEE SCENES DATASET KNOWN RESULTS

Paper & method	Results
Border-Interior Pixel Classification (BIC) [2, 14]	$87.03\% \pm 1.17\%$
Fine-tuned CaffeNet [4]	$94.45\% \pm 1.20\%$
Quaternion Orthogonal Matching Pursuit Q-OMP [15]	$90.75\% \pm 0.67\%$
Architecture II (LQPCANet – Linear Quaternion Principal Component Analysis) + GoogLeNet [1]	88.46%
GoogLeNet [6, 16]	91.83%
Multiple lAyeR feaTure mAtching(MARTA) generative adversarial networks (GANs) [17]	88.36%

Let’s look at recognition accuracy of shallow CNN that is shown in Fig. 1. It contains 2 convolution layers, 2 maxpooling and 2 dense layers. Additionally, 2 dropout layers were used in between to reduce overfitting possibility.

Results of recognition were gathered with 5-fold cross-validation strategy. 10 independent training and recognition experiments were performed for each fold, results were averaged. The common score was obtained by averaging of all folds results. We were able to get 86.64% of correct recognition (with a maximum value of 89.67% and minimum 78.83%). Training time for separate fold was about 2 minutes

(software/hardware description that was used is available in section IV), 21 epochs were performed during training, each image was resized down to 32x32 pixels. Training of the whole fold was done in 32 image batches.

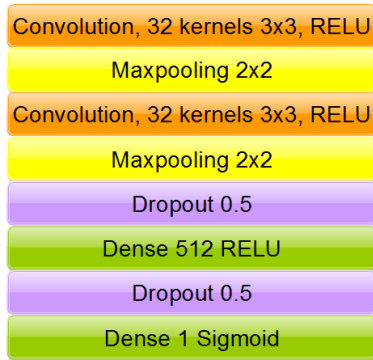


Fig. 1. CNN architecture to process Brazilian Coffee Scenes dataset.

B. SAT-4 and SAT-6

SAT-4 and SAT-6 airborne datasets were presented in [18, 19] and contain huge amount of different RGB-NIR aerial images with “barren land”, “trees”, “grassland” and “other” classes for SAT-4 and “barren land”, “trees”, “grassland”, “roads”, “buildings” and “water” classes for SAT-6. Each image has 28x28 pixels size and only one corresponding label, both datasets are challenging because of the huge amount of training and test images – 400000 and 100000 for SAT-4 and 324000 and 81000 respectively. Again, near-infrared channel was not used in this paper. Examples of SAT-6 images are shown in Table III.

TABLE III. EXAMPLES OF SAT-6 DATASETS IMAGES

Class label	Sample images
Building	
Barren land	
Trees	
Grassland	
Road	
Water	

Some results of this dataset processing based on different techniques are available in Table IV. As one can see very high results were achieved for both datasets.

TABLE IV. SAT-4 AND SAT-6 DATASETS KNOWN RESULTS

Paper & method	Results (SAT-4)	Results (SAT-6)
DeepSat [19]	97.95%	93.9%
SatCNN [20]	99.65%	99.54%
DropBand [21]	99.997%	99.994%
AlexNet, VGG [22]	99.98%	99.98%

Huge size of training and test set did not allow fitting them in memory at once even for shallow CNN that is presented in Fig. 2, so partial training process was

introduced. Let’s denote with S such amount of images that is possible to load to memory and train, so full image dataset (training or testing) size of N is split to N/S parts. Current weights of CNN are saved after training of each part, that allows to free memory, and restored before training of next part. Training of each part S was performed in 128 image batches and 30% of images were chosen randomly as validation set.

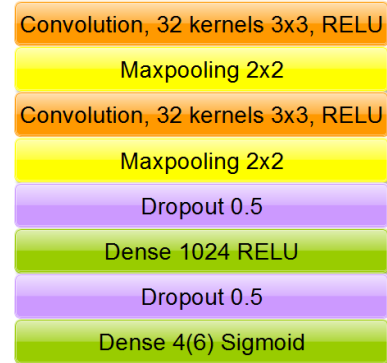


Fig. 2. CNN architecture to process SAT-4 and SAT-6 datasets.

Best modeling results we achieved are presented in Table V for SAT-4 and in Table VI for SAT-6 respectively. Each experiment was performed 5 times with averaging of scores and timings.

TABLE V. SAT-4 RECOGNITION RESULTS

Size of images	Epochs per part	Size of part S	Recognition rate	Training time for the whole dataset
20x20	1	50000	94.81%	10 min.
32x32	15	2000	97.87%	3 hrs. 20 min.
32x32	10	4000	97.99%	3 hrs. 20 min.

It is possible to see that S value should be chosen properly because neither too small nor too big values don’t allow to get best results. Also, it may be noticed that upscaling of images is preferable than downscaling. Correct recognition rate is high and comparable to DeepSat [19] investigation but is not high enough to be comparable directly to state-of-the-art approaches.

TABLE VI. SAT-6 RECOGNITION RESULTS

Size of images	Epochs per part	Size of part S	Recognition rate	Training time for the whole dataset
20x20	1	40500	96.15%	7 min.
32x32	15	2000	97.86%	2 hrs. 40 min.
32x32	10	4000	98.34%	1 hr. 40 min.

C. MNIST

MNIST [23, 24] is the other famous dataset of handwritten digit images that contains 60000 train images size of 28x28 and 10000 of test ones. The architecture of CNN is presented in Fig. 3, it allows to achieve average

98.52% of correct recognition rate with 10 epochs of training per part with size $S = 5000$ and resizing of input images to 32×32 . Order of train files was randomized before each of 5 experiments. Training of the whole dataset in this setup took about 1 hr. and 15 min. As earlier, 30% of samples in each part were used as validation data, changing of weights during learning was done in 128 batches.

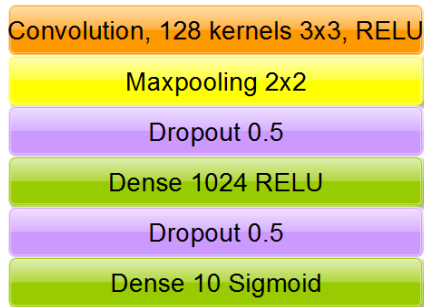


Fig. 3. CNN architecture to process MNIST dataset.

D. UC Merced Land Use and CIFAR

UC Merced Land Use dataset was introduced in [25] and contains 2100 aerial images of size 256×256 pixels that are split into 21 classes: agricultural, airplane, baseball diamond, beach, buildings, chaparral, dense residential, forest, freeway, golf course, harbor, intersection, medium residential, mobile home park, overpass, parking lot, river, runway, sparse residential, storage tanks and tennis courts. UC Merced Land Use dataset is very popular [4, 26 - 28] with high result reported above 99%.

CNN, which is presented in Fig. 4, was used to perform 5-fold cross validation recognition. Dataset was split to 5 folds with 420 images in each with balanced amount of every class representatives. 3 folds were used for training, another one for validation and the last one for testing. Training was performed 50 times with 30 epochs each time, the average correct recognition rate is 85.96% (minimum value is 81.86%, maximum one is 88.54%). Full training time was about the hour, weights for 64 images were updated simultaneously.

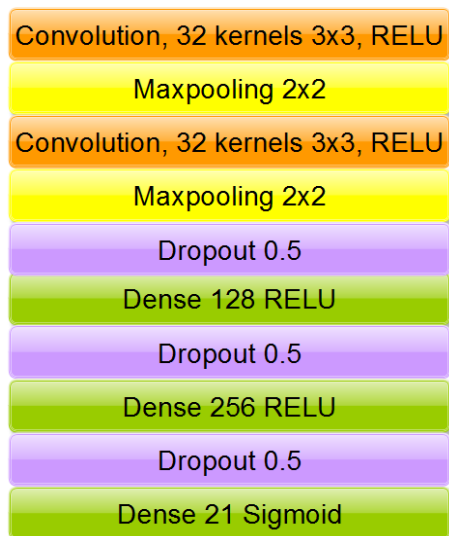


Fig. 4. CNN architecture to process UC Merced Land Use dataset.

Similar recognition rate (85.67%, minimum is 81.9%, the maximum is 89.52%) may be reached with lower CNN, e.g. network shown in Fig.1 with corresponding amount of output neurons, or even with less training iterations or epochs. But it is much harder to achieve better accuracy with such shallow architectures for this dataset.

CIFAR-10 and CIFAR-100 [29, 30] are datasets of tiny (32×32) color images representing 10 and 100 non-overlapping classes respectively, each dataset has 50000 of train and 10000 of test images. Best recognition results reported are over 96% for CIFAR-10 and over 75% for CIFAR-100 [31].

Applying CNN, that is shown in Fig. 4, allows getting 71.7% of correct recognition rate with training time about 4 hrs. We set up the size of the part to be $S = 2000$ and performed 10 epochs per part during training and 20 iterations. Increasing this quantity up to 30 with part size modification $S = 4000$ allows to improve recognition rate up to 74% with training time about of 6 hrs.

IV. TECHICAL NOTES

Results and all reported timings were achieved with Keras [32] deep learning library using default Theano backend, Python 2.7 programming language and without GPU-optimization. Hardware included a personal computer with Intel Core i7 4x 3.60GHz processor and 16GB RAM.

V. CONCLUSIONS

Resolving of practical pattern recognition (classification) problem using CNN seems to be related to the complex structure of network usually but it is possible to get similar results using only shallow networks like we presented on Fig.1 – Fig. 4. Not every shallow model is successful and not every problem may be resolved using this approach though, this was confirmed by testing of shallow models with different options.

Looking at datasets we successfully applied presented approach on (Brazilian coffee, SAT, MNIST) we may guess about the requirement for a problem to be resolved with shallow CNN. Samples of all these datasets have more common image information like color or shape on the same background in case of MNIST images. This information is mostly retained after downscaling to small size, besides that, all images are small initially.

Ways to recognize such datasets as CIFAR and UC Merced Land Use effectively enough with shallow networks were not found. Looks like CIFAR samples have important features which shallow CNN are unable to catch, whilst UC Merced Land Use images have 256×256 size and most details seem to be lost after downscaling.

Partial training approach we used in paper should be investigated deeply as that's unclear for now how training of each separate part influences other parts and the whole model.

REFERENCES

- [1] J. Wang, C. Luo, H. Huang, H. Zhao and S. Wang, "Transferring Pre-Trained Deep CNNs for Remote Scene Classification with General Features Learned from Linear PCA Network," Remote Sens. 2017, 9(3), 225; doi:10.3390/rs9030225.

- [2] K. Nogueira, W. O. Miranda and J. A. Dos Santos, "Improving spatial feature representation from aerial scenes by using convolutional networks," in: 28th IEEE SIBGRAPI Conference on Graphics, Patterns and Images, pp. 289–296, 2015.
- [3] A. Krizhevsky, I. Sutskever and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in Neural Information Processing Systems, pp. 1106–1114, 2012.
- [4] K. Nogueira, O. A. B. Penatti and J. A. Dos Santos, "Towards Better Exploiting Convolutional Neural Networks for Remote Sensing Scene Classification," [Online]. Available: <https://arxiv.org/pdf/1602.01517.pdf> [June 02, 2018].
- [5] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," [Online]. Available: <https://arxiv.org/pdf/1409.1556.pdf>. [March 15, 2017].
- [6] C. Szegedy, W. Liu, Y. Jia et al., "Going deeper with convolutions," [Online]. Available: <https://arxiv.org/pdf/1409.4842.pdf>. [March 10, 2017].
- [7] K. He, X. Zhang, S. Ren and J. Sun, "Deep Residual Learning for Image Recognition," [Online]. Available: <https://arxiv.org/pdf/1512.03385v1.pdf>. [May 12, 2017].
- [8] A. Deshpande, "The 9 Deep Learning Papers You Need To Know About (Understanding CNNs Part 3)," [Online]. Available: <https://adeshpande3.github.io/adeshpande3.github.io/The-9-Deep-Learning-Papers-You-Need-To-Know-About.html>. [May 15, 2017].
- [9] V. Powell, "Image Kernels," [Online]. Available: <http://setosa.io/ev/image-kernels/>. [June 02, 2017].
- [10] P. Veličković, "Deep learning for complete beginners: convolutional neural networks with keras," [Online]. Available: <https://cambridgespark.com/content/tutorials/convolutional-neural-networks-with-keras/index.html>. [June 02, 2017].
- [11] D. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization," [Online]. Available: <https://arxiv.org/abs/1412.6980v8.pdf>. [July 20, 2017].
- [12] Brazilian Coffee Scenes Dataset [Online]. Available: <http://www.patreeo.dcc.ufmg.br/downloads/brazilian-coffee-dataset/>. [May 17, 2017].
- [13] O. A. B. Penatti, K. Nogueira and J. A. Dos Santos, "Do deep features generalize from everyday objects to remote sensing and aerial scenes domains?," in IEEE Computer Vision and Pattern Recognition Workshops, pp. 44–51, 2015.
- [14] R. de O. Stehling, M. A. Nascimento and A. X. Falcao, "A compact and efficient image retrieval approach based on border/interior pixel classification," In Eleventh International Conference on Information and Knowledge Management (CIKM'02), pp.102–109, 2002.
- [15] V. Risojevic and Z. Babic, "Unsupervised Quaternion Feature Learning for Remote Sensing Image Classification," [Online]. Available: http://dsp.etfbl.net/aerial/unsupervised_final.pdf. [June 15, 2017].
- [16] M. Castelluccio, G. Poggi, C. Sansone and L. Verdoliva, "Land Use Classification in Remote Sensing Images by Convolutional Neural Networks," [Online]. Available: <https://arxiv.org/pdf/1508.00092.pdf>. [June 17, 2017].
- [17] DaoYu Lin, "Deep Unsupervised Representation Learning for Remote Sensing Images," [Online]. Available: <https://arxiv.org/pdf/1612.08879.pdf>. [June 20, 2017].
- [18] SAT-4 and SAT-6 airborne datasets [Online]. Available: <http://csc.lsu.edu/~saikat/deepsat/>. [June 20, 2017].
- [19] S. Basu, S. Ganguly, S. Mukhopadhyay, R. DiBiano, M. Karki and R. Nemani, "DeepSat – A Learning framework for Satellite Imagery," [Online]. Available: <http://bit.csc.lsu.edu/~saikat/publications/sigproc-sp.pdf>. [June 20, 2017].
- [20] Y. Zhong, F. Fei, Y. Liu, B. Zhao, H. Jiao and L. Zhang, "SatCNN: satellite image dataset classification using agile convolutional neural networks," Remote Sensing Letters, 8:2, 136-145, DOI: 10.1080/2150704X.2016.1235299. [Online]. Available: <http://dx.doi.org/10.1080/2150704X.2016.1235299>. [June 23, 2017].
- [21] N. Yang, H. Tang, H. Sun and X. Yang, "DropBand: a convolutional neural network with data augmentation for scene classification of VHR satellite images," [Online]. Available: <http://proceedings.utwente.nl/403/1/Yang-DropBand-91.pdf>. [June 25, 2017].
- [22] M. Papadomanolaki, M. Vakalopoulou, S. Zagoruyko and K. Karantzalos, "Benchmarking deep learning frameworks for the classification of very high resolution satellite multispectral data," in ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences, Volume III-7, XXIII SPRS Congress, Prague, Czech Republic, 12–19 July 2016.
- [23] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," Proceedings of the IEEE, 86(11):2278-2324, November 1998.
- [24] Y. LeCun, C. Cortes and C.J.C. Burges, "THE MNIST DATABASE of handwritten digits," [Online]. Available: <http://yann.lecun.com/exdb/mnist/>. [July 10, 2017].
- [25] Y. Yang and S. Newsam, "Bag-of-visual-words and spatial extensions for land-use classification," in: 18th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems, pp. 270–279, November 02 – 05, 2010.
- [26] M. Castelluccio, G. Poggi, C. Sansone and L. Verdoliva, "Land Use Classification in Remote Sensing Images by Convolutional Neural Networks," [Online]. Available: <https://pdfs.semanticscholar.org/4191/fe93bfd883740a881e6a60e54b371c2f241d.pdf>. [July 26, 2017].
- [27] F. P. S. Luus, B. P. Salmon, F. van den Bergh and B. T. J. Maharaj, "Multiview Deep Learning for Land-Use Classification," in IEEE Geoscience and Remote Sensing Letters, Vol. 12, pp. 2448 – 2452, 2015.
- [28] Y. Zhong, F. Fei and L. Zhang, "Large patch convolutional neural networks for the scene classification of high spatial resolution imagery," J. Appl. Remote Sens. 10(2), 025006 (2016), doi: 10.1117/1.JRS.10.025006.
- [29] The CIFAR-10 dataset [Online]. Available: <https://www.cs.toronto.edu/~kriz/cifar.html>. [July 27, 2017].
- [30] A. Krizhevsky, "Learning Multiple Layers of Features from Tiny Images," [Online]. Available: <https://www.cs.toronto.edu/~kriz/learning-features-2009-TR.pdf>. [July 27, 2017].
- [31] Classification datasets results, [Online]. Available: http://rodrigob.github.io/are_we_there_yet/build/classification_datasets_results.html. [July 27, 2017].
- [32] F. Chollet, "Keras," [Online]. Available: <https://github.com/fchollet/keras>. [July 30, 2017].

Quantization of the Space of Structural Image Features as a Way to Increase Recognition Performance

Volodymyr Gorokhovatskyi
Informatics department
Kharkiv National University of Radio Electronics
Kharkiv, Ukraine
gorohovatsky.vl@gmail.com

Putyatin Yevgenyi
Informatics department
Kharkiv National University of Radio Electronics
Kharkiv, Ukraine

Oleksii Gorokhovatskyi
Informatics and Computer Technologies department
Simon Kuznets Kharkiv National University of Economics
Kharkiv, Ukraine
oleksii.gorokhovatskyi@gmail.com

Peredrii Olena
Informatics and Computer Technologies department
Simon Kuznets Kharkiv National University of Economics
Kharkiv, Ukraine

Abstract — A modification of the structural image recognition method in computer vision systems is proposed. In order to improve the performance of recognition, quantization (clustering) is applied in the space of characteristic features that form the pattern of the object. Due to the transformation of structural objects descriptions from a set representation to a vector form, the amount of computation is reduced tens of times. The results of experiments that confirmed the effectiveness and increase of decision-making process are shown.

Keywords — *computer vision; structural recognition methods; set of characteristic features; descriptor; quantization; competitive learning; recognition performance; noise immunity*

I. INTRODUCTION

Structural methods for image recognition, where visual objects are represented by sets of characteristic features (points of interest), are widely used due to high efficiency associated with the practical ability to perform the investigation and recognition in complex conditions connected with the presence of a partial representation of the patterns [1-3]. In this case, the user of computer vision system has the ability to determine independently which specific part of the available description is acceptable for decision-making about the class of an object. Successful implementations of structural methods in problems of face recognition, animals recognition, as well as a number of iconic types images are known: coats of arms, paintings, logos, brands, etc. [4, 5].

The main factors that determine the effectiveness of computer recognition are the features detection method, the space of images within decision is made of, as well as the level of influence by external obstacles (noise, distortions). Usage of information about etalons during recognition process determines the quality of learning (tuning) of the system for specific initial data. Embedding of the learning step allows not only to improve recognition performance by adaptation of parameters but also ensures the universality and the ability of its functioning for arbitrary image databases [4].

An effective way to increase the effectiveness of the structural recognition methods in terms of speed and, in fact, without reducing the probabilistic characteristics, is the use of the vector quantization (granulation) in the space of structural features [4]. Vector quantization allows to approximate the space of key point descriptors by splitting them into subsets of equivalent elements. Due to the quantized representation, with the creation of set-vector mapping, the features space is transformed, as a result, the calculation of relevance of objects descriptions can be interpreted as distance or vector similarity [3]. Vector quantization without a teacher (self-learning) automatically classifies input elements, which makes it possible to apply it for a number of applications not related to image analysis, for example, in the field of digital telecommunications.

Applying quantization in the space of characteristic features (CF) of images allows to obtain subsets of close elements represented by a set of centers, and out of them, like of bricks, we form the image of an arbitrary visual object. Such models can be classified as structural-statistical. Formation of object's CF, for example, by SURF, ORB detectors, allows reaching invariance to permissible geometric transformations [2, 7].

The procedures for quantization and learning as its part in the organization of recognition process are often interpreted by researchers as a whole. However, vector quantization itself is, in our opinion, more objective, and learning, having a specific nature of the object of research, is more aimed to meaningful management of the adaptation process.

Goal of the work is to investigate and evaluate the effectiveness of quantization in structural recognition of images based on image descriptions in the form of sets of CF.

The objectives of investigation are the analysis of quantization qualities and learning process for the list of vectors-descriptors of images, as well as the evaluation of the effectiveness of learning and recognition for real image databases.

II. QUANTIZING OF THE DESCRIPTOR SPACE

Accordingly to VQ (Vector Quantization), we perform a discrete approximation of continuous (in the general case) input data from the set of vectors $W = \{x \mid x \in R^n\}$, $W \subseteq R^n$ using predefined set of k encoding vectors $M = \{m_i\}_{i=1}^k$, $m_i \in R^n$, $i = 1, 2, \dots, k$ [6,8]. Concurrent approximation of vector $x \in W$ means a search for a number v of vector $m_v \in M$ closest to it (in terms of Euclid $\rho(x, m_i)$ metrics usually) in the space of encoding vectors:

$$v = \arg \min_{i=1, \dots, k} \rho(x, m_i). \quad (1)$$

Model (1) is known as concurrent Kohonen learning [6]. Idea of a quantization is the formation of M on the basis of training set W accordingly to optimum of some criteria.

Normalization of W is required to ensure stable stationary learning process and equivalent influence of input vectors to result: $W = \{x^* \mid x^* = x/\|x\|, \|x\| = \sqrt{\sum_i x_i^2}\}$, assuming: $\|x^*\| = 1$.

In the space of CF formed by SURF, ORB feature detectors, this condition is fulfilled, so no additional signal normalization is required.

Quantization in general form can be formulated as a global optimization problem for some functional, that reflects the quality of the cluster system, while the total distances between the elements within the clusters are minimized, and the distances between the cluster centers are maximized. Taking into account, that quantization plays an auxiliary role here, the key criterion is the probability of correct recognition.

Batch processing during quantization is implemented in the form of a computational scheme, which is used in a situation where the set W is available as a whole at the beginning of learning, and all $x \in W$ are considered to be equivalent. A common version of batch processing – k-means (C-means) [6] is applicable for arbitrary metrics when comparing elements. The k-means algorithm arranges cluster centers (centroids) so that the average values for the lists of elements within the clusters differ as much as possible.

Stages of calculations applied to sets of CF:

1) random k vectors are assigned as initial centers $M = \{m_i\}_{i=1}^k$, $m_i \in W$;

2) for each $i = \overline{1, k}$ with learning according to (1) list $W_i \subseteq W$ is formed. This list contains elements, which have m_i as nearest encoding vector, i.e. form subsets $W_i = \{x \in W \mid \arg \min_v \rho(x, m_v) = i\}$; list $\{W_i\}$ in this case creates partition $W : W = \cup W_i$, $W_i \cap W_j = \emptyset$;

3) average value according to W_i is calculated as next m_i value: $m_i = \sum_{v=1}^{s(i)} x_v / s(i)$, where $s(i) = \text{card} W_i$ – amount of elements in W_i ;

4) steps 2 and 3 are repeated while list is unstable in terms of some criterion.

This algorithm is especially effective if the initial values of the vectors m_i are previously somehow coordinated with the training set. The algorithm does not contain the learning speed parameter, which is not required to be controlled during processing, and there are no problems with convergence. Algorithm is stopped if changes of the centers $\{m_i\}$ become insignificant between iterations, which is evaluated by criteria:

$$\Delta(M[h+1], M[h]) \leq \varepsilon_M,$$

where Δ is a measure of difference between two lists of centroids, $M[h] = \{m_i(h)\}_{i=1}^k$ are the values of list on step h of an iteration; ε_M – threshold value a priori.

As an example of Δ criteria sum of distances between lists may be used $\Delta = \sum_{i=1}^k \rho(m_i(h+1), m_i(h))$.

In order to make calculations more effective from performance point of view Δ value during each iteration is calculated only for centroids, that changes their value.

Experimental modeling of k-means scheme described above for sets of CF shows, that it is enough to do only a few iterations [4, 9].

K-means algorithm approximates the distribution function of the set of input samples by the criterion of minimum error, that is usually the sum of squares of deviations from cluster centers [8]:

$$E = \sum_{i=1}^k \sum_{v=1}^{s(i)} \rho^2(x_v, m_i), \quad (2)$$

where $s(i)$ is a power of i cluster.

Distance $\rho(x, m_i) = \min_{i=1, \dots, k} \rho(x, m_i)$ between element and its cluster center is quantization error. Iterative k-means algorithm converges to local minimum of error E . Minimization of (2) promotes quantization process to fit training data in a best way.

Numerous modifications of k-means method are known, e.g., k-median method, where in order to remove anomaly values median value $m_i = \text{med} \{x_v\}_{v=1}^{s(i)}$ is selected instead of average in each cluster. The median is defined as an element of a set whose total distance to the remaining elements is minimal [6]. If it is necessary to analyze overlapping clusters, where the values of the membership function for each of the clusters are calculated, the methods of Fuzzy Classifier Means [8] are applicable.

In case of online-learning, when CF $x[t] \in W$ comes into processing one by one, center m_v of cluster, that is the winner in (1) is corrected in a way:

$$m_v[t+1] = m_v[t] + \alpha[t](x[t] - m_v[t]), \quad (3)$$

on learning step $t = 1, 2, \dots, s$, where $s = \text{card} W$ is a size of training set (total quantity of CF for the whole dataset), $\alpha[t]$ is set up by researcher and specifies learning speed assuming $\alpha[t] \rightarrow 0$ and $t \rightarrow s$.

A huge variety of learning strategies and methods has been developed (3), including modeling of dynamics of network topology [8].

III. PROPERTIES OF STRUCTURAL IMAGE DESCRIPTIONS LEARNING

Characteristic features is a vector size of n (usually 32 or 64), that is calculated with usage of some detector to image brightness function. ORB and SURF [2, 7] detectors are the most widespread. Structural description of image is a finite set $O \subset R_1^n, R_1^n = \{z | z \in R^n, \|z\|=1\}$, where $R_1^n \subset R^n$ – is a subset of n -dimensional real vectors that have norm $\|z\|=1$ [2]. Implementation of the normalization condition allows usage of CF descriptors in the learning procedures directly.

During preprocessing stage set $Z = \{Z^j\}_{j=1}^J$ of dataset image features, that includes all patterns (Z^j – is an etalon, J – is the amount of classes), is split on finite amount of k clusters $M = \{M_i\}_{i=1}^k$, in a way that $M_i \cap M_d = \emptyset, M = Z$, clusters are defined with centers $m = \{m_i\}_{i=1}^k$. Clustering maps set of CF of the whole dataset into itself $Z \rightarrow Z$, each CF belongs to just one cluster. As a result of clustering situation $m_i \notin W$ is possible. Sets Z and Z^j are multisets, where close CF we consider as equivalent. After clustering is completed, we perform the "screening" of each etalon pattern, as a result, the description Z^j of etalon takes the form:

$$H[Z^j] = (h_1, h_2, \dots, h_i, \dots, h_k)^j, \quad (4)$$

where $h_i = \text{card}\{z \in Z^j \& z \in M_i\}$, $h_i \in C$ – amount of elements of etalon Z^j , that belong to M_i cluster, C – is a set of integers.

An interesting case is a quantization with the number of centers equal to the number of classes to recognize ($k = J$). At the learning stage, for each etalon, the "centers" of the attribute descriptions of the class are formed, and we can directly perform the recognition of a visual object, without going to clusters, in real time, because condition $Z^j = M_i$ is already established. Here, the quantization apparatus works as a tool for extracting the most significant distinguishing features of etalons.

The method of recognition of visual object represented by the description $O = \{o_i\}$ for $k = J$ can be reduced to counting of the number of voices of elements in accordance with the competitive rule of optimal closeness in the cluster system

$$o_i \rightarrow M_i | \arg \min_{d=1, \dots, J} \rho(o_i, m_d) = i. \quad (5)$$

As a result of calculation (5) $\forall o_i \in O$ we obtain an integer vector of class voices $q = (q_1, q_2, \dots, q_J)$ whose maximum component determines the class d for description O :

$$d = \arg \max_j q_j \quad (6)$$

At the same time, the usage of two methods proposed earlier [3,4] is possible, the first one is based on construction of an integral vector representation for the object O , and the second one is based on the summation of the vectors of the

specific weights of elements classified according to rule (5) to the nearest of the clusters.

The essence of learning for the system of structural recognition is reduced to such problems as the construction of a cluster system for a set of attributes of the image database, rational from recognition efficiency point of view, and the estimation of optimal threshold parameters for the classification of objects [10, 13-15].

Possibility of applying the batch mode (the entire W set is defined) may be considered as characteristic of learning in the CF space, as well as the potential of using the learning with the teacher, since the belonging of individual CF etalons from the image database is known, and each of etalons forms its own class. Other characteristics of recognition include the introduction of filtering individual CF identified as noise, as well as the requirement to equalize the number of features in the descriptions of etalons. Random selection or special procedures to filter "significant" features may be used for such equalization.

IV. PERFORMANCE ANALYSIS

The preliminary calculation of the number of computational operations shows that the gain in recognition speed when applying a cluster representation of the form (4) in comparison with the traditional voting of the entire set of CF of object and etalon directly proportional to the number of etalons and inversely proportional to the number of clusters. Specifically, for SURF characteristics, this advantage can be estimated by the value $\alpha \approx 64s / (64k + k^3)$ that for practical values $k = 8$ (the number of clusters) and $s = 300$ (the number of CF in etalon) is estimated as $\alpha \approx 19$. As you can see, the performance boost is dozens of times. The experiments confirm these calculations.

V. EXPERIMENTS

The confirmation of effectiveness of the proposed method is its effective work in applications [4, 9]. Software package was developed in the form of a web-service that implements the formation of CF-descriptors and image processing with the OpenCV library under the control of the Python server. The client application allows to provide input, recognition and clustering using the k-means method. Software simulation of the batch learning of the Kohonen network was performed for the situation when the number of etalons and clusters is the same, which makes it possible to implement recognition by direct assignment of CF to the etalon on the basis of models (5), (6). Quantization at the stage of preliminary processing was done by cluster transformation of the general content of all structural descriptions of etalons in dataset.

Table 1 shows an example of the distribution of CF SURF for each of the 4 etalons (images of dolphins) in the cluster system. Fig. 1 shows one of the images and the coordinates of its CF, generated by the SURF method.

The key factors for recognition efficiency are the distribution of descriptors over the cluster system, as well as the distribution of cluster elements within each etalon (Table 1). This composition depends on the method of clustering and learning technology. On the one hand, uniform distribution across clusters provides an equivalent

representation of the characteristics during decision making. On the other hand, the diversity of the distribution structure due to the predominance of some clusters over others contributes to the improvement of the quality of distinguishing objects from their descriptions.

TABLE I. SPLIT BY CLUSTERS

Etalons	Clusters			
	M_1	M_2	M_3	M_4
Z^1	35	9	17	13
Z^2	69	31	24	53
Z^3	67	14	68	6
Z^4	26	5	15	8



Fig. 1. Dolphin image and set of features (SURF)

Experiments showed that the modified method based on the cluster description in terms of recognition efficiency is almost inferior to the traditional approach with the calculation of the distance between sets. An exception is the case of an insignificant number of clusters (2-3).

Noise resistance of the method of structural recognition based on the vector transformation of descriptions is not inferior, and in some situations, for example, with additive noise, even higher than the traditional approach with voting. Proposed approach has sufficiently high noise immunity: with distortion of up to 30% of the total number of CF from the analyzed descriptions, the method provides an almost error-free recognition with a probability higher than 0.98 within the studied dataset.

Looking at comparison between the effectiveness of the use of SURF and ORB detectors, we note that ORB releases about twice as many CF, however, their dimension is smaller (32 versus 64 for SURF). A lot of SURF detectors "cover" the image of the object in more detail, displaying the features of its shape, while the ORB signs are "grouped" and often focus on the boundaries of the object, which is ineffective.

Our experiments confirmed that the implementation of the ORB detector using OpenCV and C++ is approximately 10-20 times faster than SURF and takes up about 0.006 seconds for one image. The choice of the detector is entirely determined by the type of images analyzed and the requirements for the applications.

VI. CONCLUSIONS

Quantization of the features space reduces the dimension of the recognition problem and provides adaptation to the image dataset. Proposed approach to the construction of methods for structural image recognition based on the quantization of structural descriptions and the transition to the space of descriptor vectors has the prospect of being used due to higher speed and maintaining a sufficient level of correct recognition and noise immunity. The further development of the approach can be the construction of a cluster system and the implementation of learning procedures within each of etalons, which should ensure a more careful accounting of the properties and improve the discernibility of the processed visual objects.

REFERENCES

- [1] V. Gorokhovatsky, "Structural analysis and intellectual data processing in computer vision," SMIIT, Kharkov, 2014. (in Russian).
- [2] H. Bay, T. Tuytelaars and L. Van Gool, "Surf: Speeded up robust features," in: 9th European Conference on Computer Vision, Graz, Austria, pp. 404–417, May 7-13, 2006.
- [3] V. Gorokhovatsky, "Efficient Estimation of Visual Object Relevance during Recognition through their Vector Descriptions," in: Telecommunications and Radio Engineering, Vol. 75, No 14, pp. 1271–1283, 2016.
- [4] V. Gorokhovatsky, A. Gorokhovatsky and A. Berestovsky, "Intellectual Data Processing and Self-Organization of Structural Features at Recognition of Visual Objects," in: Telecommunications and Radio Engineering, Vol. 75, No 2, pp. 155–168, 2016.
- [5] L. Shapiro and G. Stockman, "Computer vision," Prentice Hall, 2001.
- [6] T. Kohonen, "Self-Organizing Maps," Springer Science & Business Media, 2001.
- [7] E. Karami, S. Prasad and M. Shehata, "Image Matching Using SIFT, SURF, BRIEF and ORB: Performance Comparison for Distorted Images," [Online]. Available: <https://www.researchgate.net/publication/292157133> [July 30, 2017].
- [8] S. Osovski, "Sieci neuronowane do przetwarzania informacji," in: Oficyna Wydawnicza Politechniki Warszawskiej, 2000.
- [9] V. Gorokhovatsky, Y. Putyatin and V. Stolyarov, "Research of Effectiveness of Structural Image Classification Methods using Cluster Data Model," in: Radio Electronics, Computer Science, Control, vol. 3 (42), pp. 78–85, 2017.
- [10] O. Gorokhovatskyi, "Neocognitron As a Tool for Optical Marks Recognition," in: First IEEE International Conference on DataStream Mining & Processing (DSMP), Lviv, Ukraine, pp. 169 – 172, 23-27 August 2016.
- [11] E. Rublee, V. Rabaud, K. Konolige and G. Bradski, "ORB: an efficient alternative to SIFT or SURF," in: IEEE International Conference on Computer Vision (ICCV), pp. 2564 – 2571, 2011.
- [12] OpenCV Library [Online]. Available: <http://opencv.org> [November 15, 2017].
- [13] M. Sonka, V. Hlavac and R. Boyle, "Image Processing, Analysis and Machine Vision," Thomson, 2008.
- [14] R. Szeliski, "Computer Vision: Algorithms and Applications," Springer, 2010.
- [15] Y. Amit, "2D Object Detection and Recognition: models, algorithms and networks," The MIT Press, 2002.

Logic-Mathematical Model for Recognition the Dangerous Flight Events

Ali Al-Ammouri

*Electronics and Computing Technics Department
National Transport University
Kiev, Ukraine
ammourilion@ukr.net*

Hasan Al-Ammori

*Department of International Road Transportation and Customs
Control
National Transport University
Kiev, Ukraine
hasan.ammori@gmail.com*

Arsen Klochan

*Electronics and Computing Technics Department
National Transport University
Kyiv, Ukraine
VArsenchuk@gmail.com*

Anastasia. Degtiarova

*Electronics and Computing Technics Department
National Transport University
Kiev, Ukraine
Degtjarova@gmail.com*

Abstract—The paper deal with increasing the data reliability by the use of parallel informational reservation. The mathematical models of systems for parallel informational reservation are obtained in the work. The conducted researches have shown that parallel reservation of information allows creating reliable information systems at low information possibilities of separate information sources.

Keywords—*data reliability; mathematical models; parallel informational reservation; information-control system.*

I. INTRODUCTION

Real sensors have the finite accuracy of representing the controlled information. At the same time, the accuracy and reliability of information is determined both by the design features and technical reliability of the sensors and, as a rule, does not satisfy or weakly satisfy the requirements for the accuracy and reliability of information, which fed to the inputs of computer systems for automatic control of technological processes.

It is known that statistical processing can significantly increase both the accuracy and reliability of the monitored information. It is possible by feeding information to computationally controlled systems simultaneously from several sensors connected in parallel or from the same sensor in series with a given rate. Such information input methods are called respectively parallel and sequential information reservation, which in principle allow to significantly increase the accuracy and reliability of the monitored information.

II. EXISTING METHODS OF ESTIMATING THE EFFICIENCY

The analysis of informational and control systems (ICS) structures shows that their effectiveness depends on the reliability of the information upon which the relevant decisions are made. Therefore, it is necessary to take adequate actions to increase the reliability of information in the ICS. Solutions to the problem of validating the validity of data during the transmission and processing of information in process control systems are proposed in [1, 2, 3], where are considered the principles and methods of using statistic data redundancy in solving problems, related to control of information validity based on the minimum mean-

square error criterion for different distribution laws of controlled parameters. This and other statistical methods stipulate the collection a large array of data on the operation parameters of the ICS and don't allow to evaluate their effectiveness during operation. The use of information reservation allows to reduce the shortcomings of statistical methods and to shorten the evaluation time.

III. PROBLEM STATEMENT

Informational reservation is a way to ensure the effectiveness of information and control systems by introducing reservation based on the system's information characteristics. Particularly high requirements for information are demanded to the process of control, recognition and localization the dangerous flight situations in aircraft's ICS. One of the commonest method to ensuring the reliability of information is the parallel informational reservation. The parallel informational reservation provides for use few measuring channel. On the one hand, this leads to an increase the accuracy and reliability of information, and on the other hand, this leads to a deterioration the technical and economic characteristics of the system: increasing the mass, dimensions and the cost of the system. Therefore, there is a need for evaluating parallel reservation systems with a different coefficient of majority in order to determine the optimal structure of a parallel reservation system. The main objective of paper is to ensure and evaluate the reliability of information with the use of parallel informational reservation. The objects of the research are the information and control systems of aircraft and processes of control, recognition and localization of dangerous flight situations.

IV. PROBLEM SOLUTION

Parallel informational reservation is a way to ensure the effectiveness of information and control systems, when data comes from several sources and a decision is made about the presence of a controlled parameter by the majority principle "m of n", that means that the m of n sources confirm the occurrence of a controlled parameter.

In accordance with the physical representation of the work the information sources (IS), a real IS can be in one of three incompatible random states: correct detection, false

alarm and non-detection, determined by probabilities a , b and d , respectively. Such system can be represented using a trinomial probability distribution [4, 5], according to which the probability $p(n-m, m-k, k)$ that the n of k IS will not detect controlled phenomena at all, the $m-k$ IS will work with false alarm, and $n-m$ IS will provide the correct information about the controlled phenomenon. The probability $p(n-m, m-k, k)$ is described by the following expression:

$$P_{(n-m, m-k, k)} = C_n^{n-m} a^{n-m} C_m^{m-k} b^{m-k} d^k, \quad (1)$$

wherein $a + b + d = 1$.

The application of the above-mentioned general theoretical premises can be shown on specific examples of the implementation the fire recognition systems inside of aircraft engines [6, 7].

Let the data collection system consist of two IS and be organized so that at the output of this system a signal $F_{1,2}$ will appear when at least one of the input IS is triggered. Such system is shown in Fig. 1.

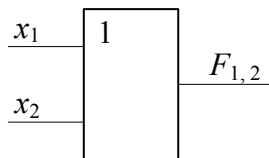


Fig. 1. The data collection system at $Q = 1$.

Through $F_{1,2}$ is designated the function of the system reaction, which consist from n IS, on incoming signals x_1, x_2 . Through Q is designated the majority index, which shows the number of IS from n , which are necessary to make a decision.

The reliability of information, obtained with such system can be estimated by three probabilistic characteristics, namely, $p_{cd1,2}$ – the probability of correct detection, $p_{fa1,2}$ – probability of false alarm and $p_{nd1,2}$ – probability of non-detection. Of course, condition $P_{cdQ,n} + P_{faQ,n} + P_{ndQ,n} = 1$ is always satisfied. It is easy to show that the probability $p_{cd1,2}$ is determined by the following probabilities: $p(a^2)$ – the probability that both IS will work correctly $p(ab)$ – the probability that one IS will work correctly and the second IS will give false information and $p(ad)$ – the probability that one IS will work correctly and the second IS will not detect necessary information. Similarly, the probability $p_{fa1,2}$ includes the probability that both IS will give false information, the probability that one IS will work false, and the second will not detect necessary information. Finally, the probability $p_{nd1,2}$ for this scheme is estimated by the probability that both IS will not detect necessary information.

If all IS are the same in their characteristics, then we can get the following dependencies:

$$\left. \begin{aligned} P_{cd1,2} &= a^2 + 2ab + 2ad; \\ P_{nd1,2} &= d^2; \\ P_{fa1,2} &= b^2 + 2bd. \end{aligned} \right\} \quad (2)$$

Let the data collection system, consisting of two IS, and be organized so that at the output of this system a signal $F_{2,2}$ will appear when both IS are triggered at the input. Such system is shown in Fig. 2.

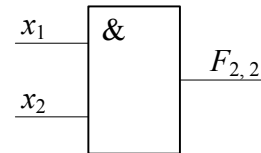


Fig. 2. The data collection system at $Q = 2$.

The probabilistic characteristics: $P_{cd2,2}$, $P_{fa2,2}$, $P_{nd2,2}$ of such system, for the identical in characteristics IS, can be described by the following expressions:

$$\left. \begin{aligned} P_{cd2,2} &= a^2 + 2ab; \\ P_{nd2,2} &= d^2 + 2ad + 2bd; \\ P_{fa2,2} &= b^2. \end{aligned} \right\} \quad (3)$$

On the basis of the mathematical models of formulas (2) and (3), we can construct graphical dependencies of the probabilistic characteristics p_{cd} , p_{fa} , p_{nd} from probabilities of correct detection by IS a , which are shown in Fig. 3

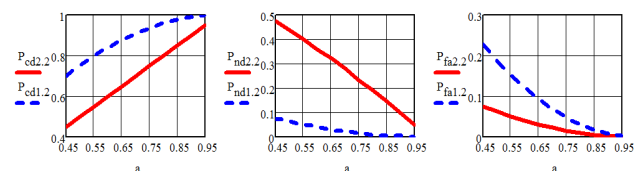


Fig. 3. The graphical dependencies for information system at $n = 2$.

Consider now a data collection system, which consist of three IS. Let this system be organized so that at the output of this system a signal $F_{1,3}$ will appear when at least one of the input IS is triggered. Such system is shown in Fig. 4.

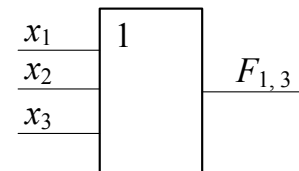


Fig. 4. The data collection system at $Q = 1$.

In view of the above the probabilistic characteristics: $P_{cd1,3}$, $P_{fa1,3}$, $P_{nd1,3}$ of such system, for the identical in characteristics IS, can be described by the following expressions:

$$\left. \begin{aligned} P_{cd1,3} &= a^3 + 3a^2b + 3a^2d + 3ab^2 + 3ad^2 + 6abd; \\ P_{nd1,3} &= d^3; \\ P_{fa1,3} &= b^3 + 3db^2 + 3bd^2. \end{aligned} \right\} \quad (4)$$

Let the data collection system, consisting of three IS, and be organized so that at the output of this system a signal $F_{2,3}$ will appear when at least two IS are triggered at the input. Such system is shown in Fig. 5.

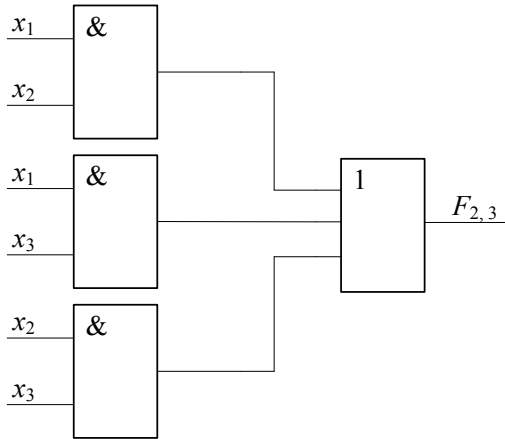


Fig. 5. The data collection system at $Q = 2$.

The probabilistic characteristics: $P_{cd2,3}$, $P_{fa2,3}$, $P_{nd2,3}$ of such system (Fig.5), for the identical in characteristics IS, can be described by the following expressions:

$$\left. \begin{aligned} P_{cd2,3} &= a^3 + 3a^2b + 3a^2d + 3ab^2 + 6abd; \\ P_{nd2,3} &= d^3 + 3ad^2 + 3bd^2; \\ P_{fa2,3} &= b^3 + 3db^2. \end{aligned} \right\} \quad (5)$$

Let the data collection system, consisting of three IS, and be organized so that at the output of this system a signal $F_{3,3}$ will appear when at least three IS are simultaneous triggered at the input. Such system is shown in Fig. 6.

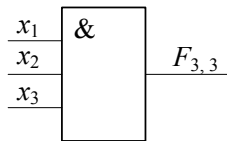


Fig. 6. The data collection system at $Q = 3$.

The probabilistic characteristics: $P_{cd3,3}$, $P_{fa3,3}$, $P_{nd3,3}$ of such system (Fig.6), for the identical in characteristics IS, can be described by the following expressions:

$$\left. \begin{aligned} P_{cd3,3} &= a^3 + 3a^2b + 3ab^2; \\ P_{nd3,3} &= d^3 + 3ad^2 + 3bd^2 + 3db^2 + 3a^2d + 6abd; \\ P_{fa3,3} &= b^3. \end{aligned} \right\} \quad (6)$$

On the basis of the mathematical models of formulas (4), (5) and (6) we can construct graphical dependencies of the

probabilistic characteristics P_{cd} , P_{fa} , P_{nd} from probabilities of correct detection by IS a , which are shown in Fig. 7.

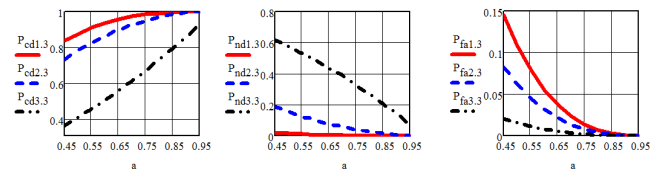


Fig. 7. The graphical dependencies for information system at $n = 3$.

Consider now a data collection system, which consist of four IS. Let this system be organized so that at the output of this system a signal $F_{1,4}$ will appear when at least one of the input IS is triggered. Such system is shown in Fig. 8.

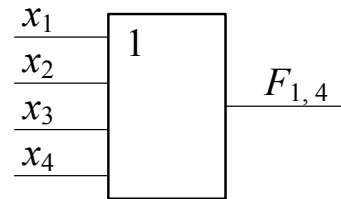


Fig. 8. The data collection system at $Q = 1$.

The probabilistic characteristics: $P_{cd1,4}$, $P_{fa1,4}$, $P_{nd1,4}$ of such system (Fig.8), for the identical in characteristics IS, can be described by the following expressions:

$$\left. \begin{aligned} P_{cd1,4} &= a^4 + 4a^3d + 4a^3b + 4a^3d + 4a^3b + 4ab^3 + 6a^2d^2 + \\ &+ 6a^2b^2 + 12a^2bd + 12ab^2d + 12abd^2; \\ P_{nd1,4} &= d^4; \\ P_{fa1,4} &= b^4 + 4db^3 + 4bd^3 + 6b^2d^2. \end{aligned} \right\} \quad (7)$$

Let the data collection system, consisting of four IS, and be organized so that at the output of this system a signal $F_{2,3}$ will appear when at least two IS are triggered at the input. Such system is shown in Fig. 9.

The probabilistic characteristics: $P_{cd2,4}$, $P_{fa2,4}$, $P_{nd2,4}$ of such system (Fig.9), for the identical in characteristics IS, can be described by the following expressions:

$$\left. \begin{aligned} P_{cd2,4} &= a^4 + 4a^3d + 4a^3b + 4ab^3 + 6a^2d^2 + 6a^2b^2 + \\ &+ 12a^2bd + 12ab^2d + 12abd^2; \\ P_{nd2,4} &= d^4 + 4bd^3 + 4ad^3; \\ P_{fa2,4} &= b^4 + 4db^3 + 6b^2d^2. \end{aligned} \right\} \quad (8)$$

Let the data collection system, consisting of four IS, and be organized so that at the output of this system a signal $F_{2,3}$ will appear when at least three IS are triggered at the input. Such system is shown in Fig. 10.

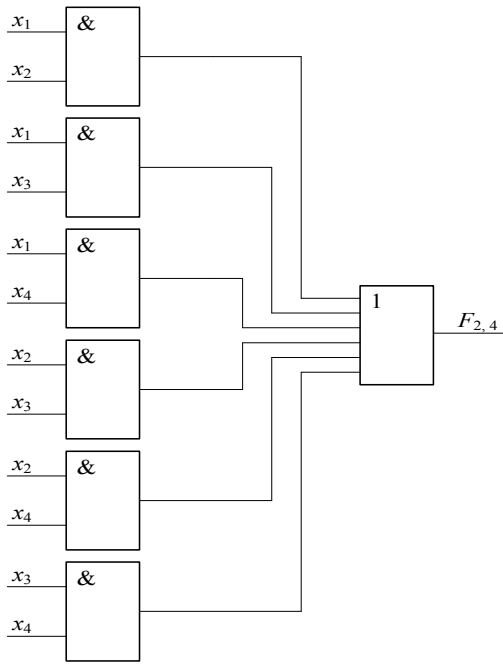


Fig. 9. The data collection system at $Q = 2$.

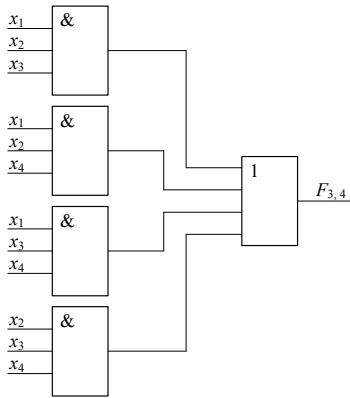


Fig. 10. The data collection system at $Q = 3$.

The probabilistic characteristics: $P_{cd3,4}$, $P_{fa3,4}$, $P_{nd3,4}$ of such system (Fig.10), for the identical in characteristics IS, can be described by the following expressions:

$$\left. \begin{aligned} P_{cd3,4} &= a^4 + 4a^3d + 4a^3b + 4ab^3 + 6a^2b^2 + 12a^2bd + 12ab^2d; \\ P_{nd3,4} &= d^4 + 4bd^3 + 6b^2d^2 + 4ad^3 + 6a^2d^2 + 12abd^2; \\ P_{fa3,4} &= b^4 + 4db^3. \end{aligned} \right\} (9)$$

The data collection system, which consist of four IS can be arranged and so that the output signal $F_{4,4}$ appears only when at least 4 sensors are triggered at the input. Such system is shown in Fig. 11.

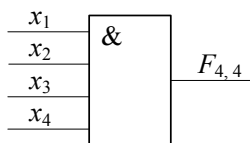


Fig. 11. The data collection system at $Q = 4$.

The probabilistic characteristics: $P_{cd4,4}$, $P_{fa4,4}$, $P_{nd4,4}$ of such system (Fig.10), for the identical in characteristics IS, can be described by the following expressions:

$$\left. \begin{aligned} P_{cd4,4} &= a^4 + 4a^3b + 4ab^3 + 6a^2b^2; \\ P_{nd4,4} &= d^4 + 4bd^3 + 4a^3d + 4db^3 + 6b^2d^2 + 4ad^3 + 6a^2d^2 + 12a^2bd + 12ab^2d + 12abd^2; \\ P_{fa4,4} &= b^4. \end{aligned} \right\} (10)$$

On the basis of the mathematical models of formulas (7), (8), (9) and (10) we can construct graphical dependencies of the probabilistic characteristics p_{cd} , p_{fa} , p_{nd} from probabilities of correct detection by IS a , which are shown in Fig. 12.

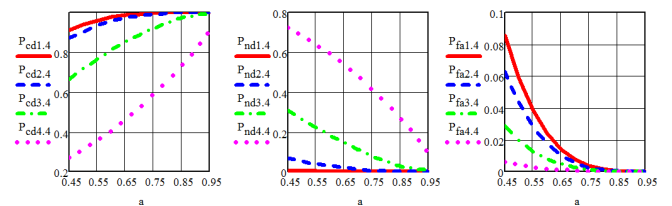


Fig. 12. The graphical dependencies for information system at $n = 4$.

The results of calculation the given probability characteristics p_{cd} , p_{fa} , p_{nd} are shown in Tables 1, 2 and 3.

The use of systems with majority logic and increasing the requirements for the IS (the operation threshold due to the use of comparators and the "restoration" the level of parametric redundancy in operation) can significantly increase the probability of correct detection p_{cd} and false alarm p_{fa} . So, with $a=0,9$, $b=0,05$, $d=0,05$, we obtain $p_{cd} = 0,994$, $p_{nd} = 0,003$, $p_{fa} = 0,003$, and when $a = 0,95$, $b = d = 0,0025$ the probability of a correct detection is $p_{cd} = 0,999$, with a further decrease p_{nd} and p_{fa} .

TABLE I. RESULTS OF CALCULATION THE PROBABILITY CHARACTERISTICS. PART I.

a	b	d	$F_{v,n}$	$P_{cdv,n}$	$P_{ndv,n}$	$P_{fav,n}$
1/3	1/3	1/3	$F_{1,2}$	0,556	0,111	0,333
			$F_{2,2}$	0,333	0,556	0,111
			$F_{1,3}$	0,705	0,035	0,26
			$F_{2,3}$	0,592	0,26	0,148
			$F_{3,3}$	0,26	0,705	0,035
			$F_{1,4}$	0,80247	0,01235	0,18519
			$F_{2,4}$	0,75309	0,11111	0,13580
			$F_{3,4}$	0,53086	0,40741	0,06173
			$F_{4,4}$	0,18518	0,80247	0,01235

TABLE II. RESULTS OF CALCULATION THE PROBABILITY CHARACTERISTICS. PART 2.

a	b	d	$F_{v,n}$	$P_{cdv,n}$	$P_{ndv,n}$	$P_{fav,n}$
1/2	1/4	1/4	$F_{1,2}$	0,75	0,062	0,108
			$F_{2,2}$	0,5	0,437	0,062
			$F_{1,3}$	0,876	0,016	0,108
			$F_{2,3}$	0,782	0,156	0,062
			$F_{3,3}$	0,406	0,58	0,014
			$F_{1,4}$	0,996093	0,000244	0,0036621
			$F_{2,4}$	0,9902344	0,0070801	0,0026855
			$F_{3,4}$	0,919921	0,078857	0,001221
			$F_{4,4}$	0,6914063	0,3083496	0,0002441

TABLE III. RESULTS OF CALCULATION THE PROBABILITY CHARACTERISTICS. PART 3.

l	b	d	$F_{v,n}$	$P_{cdv,n}$	$P_{ndv,n}$	$P_{fav,n}$
3/4	1/8	1/8	$F_{1,2}$	0,94	0,0156	0,0444
			$F_{2,2}$	0,75	0,234	0,016
			$F_{1,3}$	0,986	0,002	0,012
			$F_{2,3}$	0,948	0,043	0,009
			$F_{3,3}$	0,669	0,33	0,001
			$F_{1,4}$	0,93750	0,00391	0,05859
			$F_{2,4}$	0,90625	0,05078	0,04297
			$F_{3,4}$	0,71875	0,26172	0,01953
			$F_{4,4}$	0,31250	0,68359	0,00391

Analysis of the graphs shown in Fig. 3, Fig. 7, and Fig. 12 allows to make the following conclusions. With increasing the requirements for the reliability of information systems, namely, when Q is the maximum, then the probability of p_{nd} becomes higher than the probability of p_{fa} . That is, such system successfully suppresses the probability of p_{fa} and, at the same time, has no significant effect on p_{nd} . When Q is minimal, the information system successfully suppresses the probability of p_{nd} and has little effect on the probability p_{fa} . At $Q_{m,n}$, where m lies between 1 and n , the information system will successfully suppress both p_{fa} and p_{nd} simultaneously. Thus, depending on the technical and economic requirements, we can select the necessary structure for the system of recognizing dangerous flight events.

Analysis of the probability characteristics p_{cd} , p_{fa} , p_{nd} determined by formulas (2) - (10) allows us to make the following conclusions:

– To improve the quality of information systems consisting of n IS, in the sense of increasing the reliability of information, can be done at least by three ways: to increase in the number n of information sources; improving the IS characteristics a , b , d ; choosing the optimal information structure, by choosing the correct majority index Q .

– For information systems made of n information sources with identical probabilistic characteristics, the most acceptable is such structure, when the majority index $Q = n/2$.

V. CONCLUSION

The question of increasing the data reliability by the use of parallel informational reservation and mathematical models of system for parallel informational reservation has examined in the paper. The method of parallel informational reservation significantly reduces the probability of non-detection and has little effect on reducing the probability of false alarm. Application of the majority principle, allows reducing the probability of false alarm, but it is necessary to increase the number of parallel channels, which is due to economic constraints. The probability of false alarm can be reduced by the method of "hardening" (reducing sensitivity of sensors by raising their threshold) of individual information sources, while increasing the number of IS compensates the lacks of this method.

REFERENCES

- [1] I. Shubinsky, "Functional reliability of information systems. Methods of analysis," Moscow: Journal of Reliability, 2012 (in Russian)
- [2] T. Abezgaus, A. Tron, Yu. Kopenkin et al. "Handbook on probability calculations," Moscow: Military Publishing, 1989. (in Russian)
- [3] M. Rausand, and A. Hyland, "System reliability theory: models, statistical methods, and applications," New Jersey: A John Wiley & Sons, Inc., Publication, 2004.
- [4] A. Al-Ammouri, A. Kasyanenko, H. Al-Ammouri, and A. Degtiarova, "Optimization structure of onboard aircraft navigation systems," Proceedings of the IEEE-2016 4th International conference "Methods and Systems of Navigation and Motion Control (MSNMC)", Kyiv, Ukraine, pp. 288 – 290, October, 18-20, 2016.
- [5] A. Al-Ammouri, A. Degtiarova, A. Klochan, and H. Al-Ammouri, "Estimation the Efficiency of Information-Control System of UAV," Proceedings of the IEEE-2017 4th International conference "Actual problems of unmanned aerial vehicles developments(APUAVD)", Kyiv, Ukraine, pp. 200 – 203, October, 17-19, 2017.
- [6] Ali Al-Ammouri, "Probabilistic way to ensure the effectiveness of information systems," Management of projects, system analysis and logistic, Kyiv: NTU, vol. 3, pp. 178-180, 2006.
- [7] Ali Al-Ammouri, "Investigation of ways to improve the reliability of fire situation monitoring on board an aircraft," Problems of operation and reliability of aviation equipment, Kyiv: KMUGA, pp. 128-131, 1998.

The Multidimensional Extended Neo-Fuzzy System and its Fast Learning for Emotions Online Recognition

Yevgeniy Bodyanskiy
Control Systems Research Laboratory
Kharkiv National University of Radio
Electronics
Kharkiv, Ukraine
yevgeniy.bodyanskiy@nure.ua

Nonna Kulishova
Media Systems and Technologies
Department
Kharkiv National University of Radio
Electronics
Kharkiv, Ukraine
nokuliaux@gmail.com

Daria Malysheva
Control Systems Research Laboratory
Kharkiv National University of Radio
Electronics
Kharkiv, Ukraine
darly.malysheva@gmail.com

Abstract—Many tasks require human facial expressions automatic recognition in real time. Recent solutions to this problem using machine learning methods have been based on the applying of training data sets that include hundreds of thousands of samples. The formation of these data is too costly. In this paper, the architecture of a system using extended neo-fuzzy neurons for online emotions recognition is examined. We propose the algorithm which is based on the entropy criterion for learning the system and reducing the amount of training data thousands of times.

Keywords—extended neo-fuzzy system; online emotion recognition; entropy-information learning criterion

I. INTRODUCTION

Information technologies become an integral part of the lives of so many people; they are actively being introduced into education, business, healthcare, and entertainment. Many of these technologies are interactive, and they realize continuous two-way cooperation of a person with a computer or mobile device. One of the promising areas for the interfaces' development for such reciprocity is the approach that uses the recognition of people, their age, sex, state of health, emotional status on the real time video. This complex technical problem already finds its own solutions [1-10]. Usually, these decisions use the machine learning and neuro-fuzzy approach.

As a mathematical problem, the task of a user emotional status recognition by video is reduced to characteristic features detecting, and to the collected data clustering. The assumption of the modeled processes linearity is unreasonable. This leads to the need to select approaches that will be effective for nonlinear systems, especially in real-time conditions. Another problem is related to the fact that machine learning algorithms in this task require the training data sets in which the samples number can be tens or even hundreds of thousands. The creation of such sets is a serious, time-consuming task, significantly increasing the projects developing cost and implementation duration.

II. STANDARD AND EXTENDED NEO-FUZZY NEURON

The NFN could be very effective in the solving the problem of the person emotional state recognition from a

video. The neo-fuzzy neuron (NFN) was firstly proposed in the early 1990s by Uchino and Yamakawa [11-13] to simplify the complex nonlinear systems modeling. The NFN is computationally simple, has high approximation accuracy and the ability to minimize the chosen criterion of learning in real time.

Recently, there have been publications about the NFN applying results in different tasks. In [14-16] various architectures of the NFN and the corresponding learning algorithms are proposed. Practical tasks related to the study of induction motors vibration, bearings functioning, bacteria colonies number optimization and classification problems were successfully solved using NFN [17-20].

The standard NFN is constructed on the so-called nonlinear synapses - the elements that realize the fuzzy zero-order Takagi-Sugeno inference [21, 22]:

IF x_i IS x_{li} THEN THE $f(x_i)$ IS $w_{li}, l = 1, 2, \dots, h$.

This form corresponds to the transformation that the synapse performs:

$$f_i(x_i) = \sum_{l=1}^h w_{li} \mu_{li}(x_i) \quad (1)$$

where w_{li} is the synapse's weight, $\mu_{li}(x_i)$ - the membership function in the synapse, that fuzzify input component x_i l - weight number, $l = 1, 2, \dots, h$, i - synapse number $i = 1, 2, \dots, n$.

We improved the synapses possibilities - in one of the architectures [14] the so-called extended nonlinear synapse (ENS) was proposed, it is shown in Fig. 1.

The NFN extended nonlinear synapse realizes a fuzzy inference of an arbitrary order. For this, additional variables are used

$$y_{li}(x_i) = \mu_{li}(x_i) (w_{li}^0 + w_{li}^1 x_i + w_{li}^2 x_i^2 + \dots + w_{li}^p x_i^p), \quad (2)$$

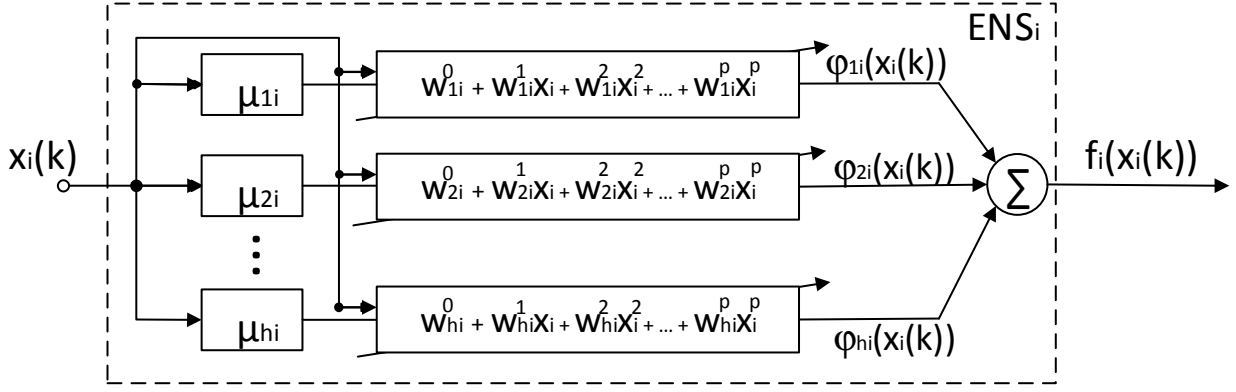


Fig. 1. Extended nonlinear synapse

$$\begin{aligned}
 f_i(x_i) &= \sum_{l=1}^h \mu_{li}(x_i) (w_{li}^0 + w_{li}^1 x_i + \\
 &+ w_{li}^2 x_i^2 + \dots + w_{li}^p x_i^p) = w_{li}^0 \mu_{li}(x_i) + \\
 &+ w_{li}^1 x_i \mu_{li}(x_i) + \dots + w_{li}^p x_i^p \mu_{li}(x_i) + \\
 &+ \dots + w_{hi}^p x_i^p \mu_{hi}(x_i),
 \end{aligned} \quad (3)$$

$$w_i = (w_{1i}^0, w_{1i}^1, \dots, w_{1i}^p, w_{2i}^0, \dots, w_{2i}^p, \dots, w_{hi}^p)^T, \quad (4)$$

therefore, we can write:

$$f_i(x_i) = w_i^T \tilde{\mu}_i(x_i), \quad (5)$$

$$y = \sum_{i=1}^n f_i(x_i) = \sum_{i=1}^n w_i^T \tilde{\mu}_i(x_i) = \tilde{w}^T \tilde{\mu}(x), \quad (6)$$

where $\tilde{w}^T = (w_1^T, \dots, w_i^T, \dots, w_n^T)^T$,

$$\tilde{\mu}(x) = (\tilde{\mu}_1^T(x_1), \dots, \tilde{\mu}_i^T(x_i), \dots, \tilde{\mu}_n^T(x_n))^T. \quad (7)$$

Thus, ENS implements the output of the form:

IF x_i IS x_{li} THEN THE $f(x_i)$ IS $w_{li}^0 + w_{li}^1 x_i + \dots + w_{li}^p x_i^p, l = 1, 2, \dots, h$,

that repeats with the formulation of Takagi-Sugeno r-th order inference.

Synapses are NFN structural blocks, which implements the mapping:

$$y = \sum_{i=1}^n f_i(x_i), \quad (8)$$

where x_i - the element of the vector of input data $x = (x_1, \dots, x_i, \dots, x_n)^T \in R^n$, i - component number, n - vector dimensionality, y - the scalar output of the NFN.

In the extended nonlinear synapse in Fig. 1 B-splines are used as membership function.

Thus, an extended neo-fuzzy neuron (ENFN), receiving an input vector $x(k) = (x_1(k), \dots, x_i(k), \dots, x_n(k))^T$ ($k = 1, 2, \dots$ - the current count of discrete time), generates a resulting scalar value

$$y(k) = \sum_{i=1}^n \sum_{l=1}^h w_{li}(k-1) \mu_{li}(x_i(k)) \quad (9)$$

where $w_{li}(k-1)$ are the values of the synaptic weights, that obtained as a result of training based on the previous $k-1$ observations. Fig. 2 shows how the elements of the ENFN [14] combine.

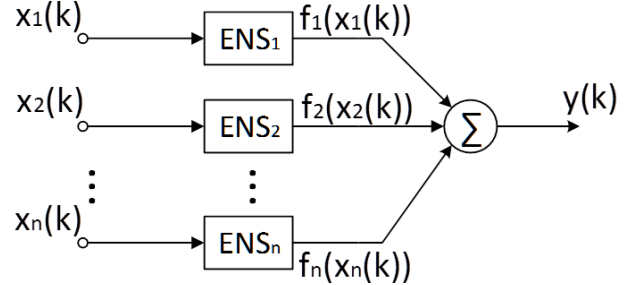


Fig. 2. Extended neo-fuzzy neuron

III. MULTIDIMENSIONAL EXTENDED NEO-FUZZY NEURON

The considered architecture of the ENFN was further developed in [23], where a multidimensional extended neo-fuzzy neuron (MENFN) was considered. In MENFN the input vector signal $x = (x_1, \dots, x_i, \dots, x_n)^T \in R^n$ generates the output vector response. This structure contains several layers. The input layer consists of ENFN; an intermediate layer of elements rejecting negative values, the output layer normalizes the output values and combines them into the resulting vector. For learning of the developed architecture, an

algorithm based on a gradient procedure was used. Writing the learning criterion in the form:

$$E(k) = \frac{1}{2}(d(k) - y(k))^2 = \frac{1}{2}e^2(k) = \frac{1}{2} \left(d(k) - \sum_{i=1}^n \sum_{l=1}^h w_{li} \mu_{li}(x_i(k)) \right)^2, \quad (10)$$

we obtain the learning algorithm:

$$\begin{cases} w(k) = w(k-1) + r^{-1}(k)e(k)\mu(x(k)), \\ r(k) = \alpha r(k-1) + \|\mu(x(k))\|^2, 0 \leq \alpha \leq 1. \end{cases} \quad (11)$$

where $d(k)$ - external data for training, $e(k)$ - error of learning, η - parameter of learning rate. Depending on the value of α , the (12) is converted to a Goodwin-Ramage-Caines algorithm [24] or an one-step Kaczmarz-Widrow-Hoff algorithm [25].

Despite its universal properties, this learning algorithm does not provide the fulfillment of strict requirements for the system learning in real time on a small training data samples number, which is present in the formulation of the user emotional status recognition task in a video sequence.

To accelerate the MENFN's learning, it was suggested to use the entropy-information learning criterion [26]:

$$E_j(t) = \frac{1}{2}(1+d_j(t)) \ln \frac{1+d_j(t)}{1+y_j(t)} + \frac{1}{2}(1-d_j(t)) \ln \frac{1-d_j(t)}{1-y_j(t)}. \quad (12)$$

It was noted in [26] that this criterion becomes essentially effective if the hyperbolic tangent is chosen as the activation functions for $y_j(t)$:

$$y_j(x) = \tanh(\tilde{w}^T x) \quad (13)$$

Then differentiating (13) respecting to w_{ij} , accounting (14), gives a simple learning algorithm of the form:

$$\frac{dw_{ij}}{dt} = \eta e_j(t) x_i, \quad (14)$$

where $e_j(t) = d_j(t) - y_j(t)$ is the local learning error. In the discrete case this simple expression takes the form:

$$w_j(k+1) = w_j(k) + \eta(k) e_j(k) x(k) \quad (15)$$

This recording simplicity ensures both the computational simplicity of the algorithm, as well as the high learning rate required for on-line applications.

As a result, the architecture of the MENFN acquired the following form (Fig. 3).

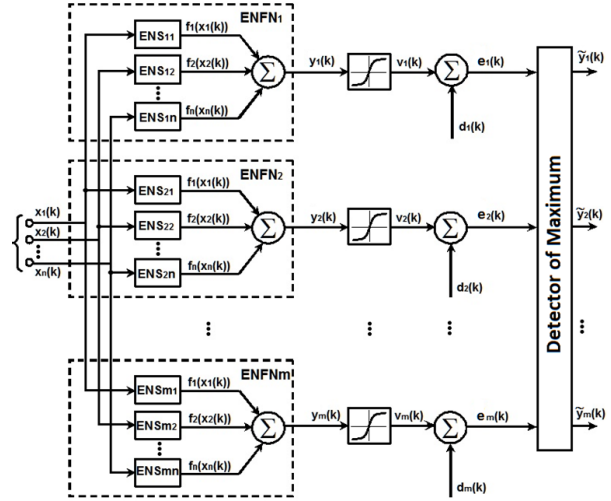


Fig. 3. Multidimensional extended neo-fuzzy neuron with activation functions $y_j(t) = \tanh(\tilde{w}^T x)$

The first layer consists of the ENFN, the number of which corresponds to the output vector $y_m(k)$ dimension. Nonlinear synapses amount, that forms each of the ENFNs, corresponds to the input feature vector $x_n(k)$ dimension. The next layer implements the function

$$v_j(k) = \psi(y_j(k)) = \tanh(y_j(k)) \quad (16)$$

The last layer of MENFN detects the maximums in the calculated learning algorithms values $v_j(k)$:

$$\tilde{y}(k) = \sup_{j=1}^m \{v_j(k)\}, \quad (17)$$

is necessary, if the learning vector set in the range $[0,1]$.

IV. EXPERIMENT

The proposed architecture ability to recognize individual emotions was investigated using photographs from two open bases - Psychological Image Collection at Stirling (PICS) [27], partly from the Extended Cohn-Kanade (CK+) database [28]. Some images are in public use as objects for recognition.

In the proposed MENFN architecture, 11 membership functions were used in each non-linear synaptic ENS. The amount of terms in the fuzzy inference rules is assumed to be 3, so that the network realizes the Takagi-Sugeno output of the second order in such form:

I IF x_i IS x_{li} THEN THE $f(x_i)$ IS $w_{li}^0 + w_{li}^1 x_i + \dots + w_{li}^p x_i^p, l = 1, 2, \dots, h$.

The NFNs amount m corresponds to the dimensionality of the output data vector. Seven basic emotions are selected for recognition: anger, disgust, fear, surprise, happiness, sadness, neutral expression. Therefore, $m = 7$. The character features vector contains the two-dimensional coordinates of 35 feature points position (Fig. 4).



Fig. 4. Examples of training images and position of characteristic points

To increase the network learning rate we decided to use variable value, that decreases as the number of the learning epoch and the photo position in the training sample increases:

$$\eta = \frac{1}{N_{EP} \cdot n_{InSet}} \quad (18)$$

where N_{EP} is the number of the current network learning epoch; n_{InSet} - current photo number in the set.

In this task, special attention was paid to a learning data set small size. To examine how the proposed architecture and learning algorithm will recognize facial expressions, small photo sets are used. Their dimensions are given in Table I.

TABLE I. DIMENSIONS OF TRAINING SETS OF PHOTOS FOR INDIVIDUAL EMOTIONS

Emotion	Anger	Disgust	Fear	Happiness	Sorrow	Surprise	Neutral
Data set size	49	66	35	45	19	50	80

The network learns to recognize each emotion separately in several algorithm steps; the resulting learning error is very small. The plots of the error depending on the learning epoch are shown in Fig. 5.

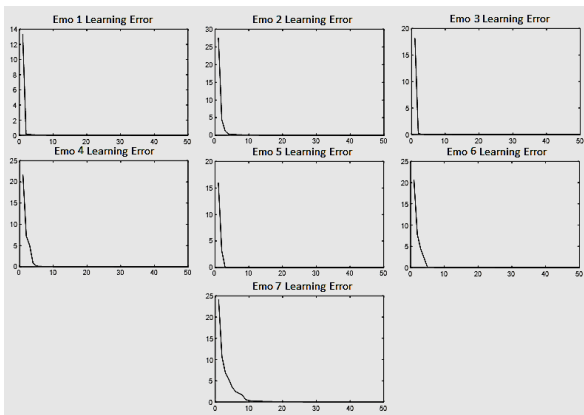


Fig. 5. The plots of the MENFN learning error on small sets for each of the seven emotions separately

Obviously, the learning rate varies from 2 to 15 epochs for different emotions. The sizes of learning sets do not exceed 100 samples each.

Then the architecture ability to learn from a mixed set was examined, and sets total size was 344 photos. The number of unrecognized emotions is given in Table II.

TABLE II. THE NUMBER OF UNRECOGNIZED EMOTIONS AS A RESULT OF MENFN LEARNING FROM A MIXED SET

	Primary emotions						
	Anger	Disgust	Fear	Happiness	Sorrow	Surprise	Neutral
The number of images in the learning set	1	0	2	2	0	0	2
The percentage of unrecognized images, %	2.04	0	5.71	4.44	0	0	2.5

A MENFN learning error change is shown in Fig. 6.

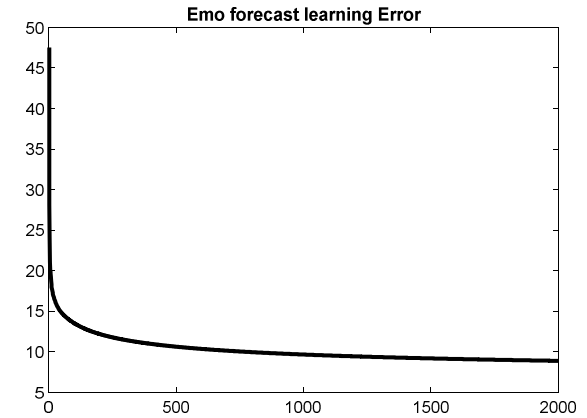


Fig. 6. MENFN learning error for the mixed data set

V. CONCLUSIONS

The paper proposes architecture of the MENFN. Its structure generalizes the standard NFN for the case of arbitrary order fuzzy inference procedure and multidimensional input and output data. The proposed learning algorithm allows effectively distribute the data between the different clusters. Considered MENFN has high learning rate, provided by the learning algorithm based on the entropy criterion; improved clustering properties, easy to numerical implementation.

REFERENCES

- [1] A. Kołakowska, A. Landowska, M. Szwoch, W. Szwoch, M. R. Wrobel, "Human-Computer Systems Interaction: Backgrounds and Applications," ch. 3, Emotion Recognition and Its Applications. Cham: Springer International Publishing, 2014, pp. 51 – 62.
- [2] Kaggle. Challenges in representation learning: Facial recognition challenge, 2013.
- [3] G.U. Kharat, S.V. Dudul, "Emotion Recognition from Facial Expression Using Neural Networks," in Human-Computer Systems Interaction. Advances in Intelligent and Soft Computing, vol 60, Z.S. Hippe, J.L. Kulikowski, Eds. Berlin, Heidelberg: Springer, 2009.
- [4] C. Shan, S. Gong, and P. W. McOwan, "Facial expression recognition based on local binary patterns: A comprehensive study," Image and Vision Computing, vol. 27, no. 6, 2009, pp. 803 – 816.
- [5] B. Fazel, J. Luetttin, "Automatic facial expression analysis: a survey", Pattern Recognition, 36(1), 2003, pp. 259 – 275.
- [6] Ch.-Yi Lee, Li-Ch. Liao, "Recognition of Facial Expression by Using Neural-Network System with Fuzzified Characteristic Distances Weights," IEEE Int. Conf. Fuzzy Systems FUZZ-IEEE 2008. [IEEE World Congress on Computational Intelligence, pp. 1694 – 1699, 2008].
- [7] N. Kulishova, "Emotion Recognition Using Sigma-Pi Neural Network," Proc. of 2016 IEEE First International Conference on Data Stream Mining & Processing (DSMP), Lviv, 2016, pp. 327 – 331.

- [8] A. Graves, J. Schmidhuber, C. Mayer, M. Wimmer, B. Radig, "Facial Expression Recognition with Recurrent Neural Networks," International Workshop on Cognition for Technical Systems, Munich, Germany, October 2008.
- [9] S. Ouelett, "Real-time emotion recognition for gaming using deep convolutional network features," CoRR, vol. abs./1408.3750, 2014.
- [10] B. Kim, J. Roh, S. Dong, and S. Lee, "Hierarchical committee of deep convolutional neural networks for robust facial expression recognition," Journal on Multimodal User Interfaces, 2016, pp. 1–17.
- [11] J. Miki, J. Yamakawa, "Analog implementation of neo-fuzzy neuron and its on-board learning," in Computational Intelligence and Applications, Ed. N.E. Mastorakis, Piraeus: WSES Press, 1999, pp. 144 – 149.
- [12] J. Yamakawa, E. Uchino, J. Miki, H. Kusanagi, "A neo-fuzzy neuron and its application to system identification and prediction of the system behavior," Proc. 2-nd Int. Conf. on Fuzzy Logic and Neural Networks "IIZUKA-92", Iizuka, Japan, 1992, pp. 477 – 483.
- [13] E. Uchino, J. Yamakawa, "Soft computing based signal prediction, restoration and filtering," in Intelligent Hybrid Systems: Fuzzy Logic, Neural Networks and Genetic Algorithms, Ed. Da Ruan, Boston: Kluwer Academic Publishers, 1997, pp. 331 – 349.
- [14] Ye.V. Bodyanskiy, N.Ye. Kulishova, "Extended neo-fuzzy neuron in the task of images filtering," Radioelectronics. Computer Science. Control, № 1(32), 2014, pp. 112 – 119.
- [15] Ye. Bodyanskiy, Y. Victorov, "The cascade of neo-fuzzy architecture and its online learning algorithm," Int. Book Series Inf. Sci. Comput., 17(1), 2010, pp. 110 – 116.
- [16] Ye. Bodyanskiy, I. Kokshenev, V. Kolodyazhniy, "An adaptive learning algorithm for a neo-fuzzy neuron," Proc. of the 3rd Conference of the European Society for Fuzzy Logic and Technology, pp. 375 – 379, 2005.
- [17] D. Zurita, M. Delgado, J.A. Carino, J.A. Ortega, G. Clerc, "Industrial Time Series Modelling by Means of the Neo-Fuzzy Neuron," IEEE Access, vol. 4, 2016, pp. 6151 – 6160.
- [18] M. Pandit, L. Srivastava, V. Singh, "On-line voltage security assessment using modified neo-fuzzy neuron based classifier," IEEE Int. Conf. Ind. Technol., 2006, pp. 899 – 904.
- [19] H.D. Kim, "Optimal learning of neo-fuzzy structure using bacteria foraging optimisation," Proceedings of the ICCA, 2005.
- [20] A.M. Silva, W. Caminhas, A. Lemos, F. Gomide, "A fast learning algorithm for evolving neo-fuzzy neuron," Applied Soft Computing, vol. 14, Part B, January 2014, pp. 194 – 209.
- [21] T. Takagi, M. Sugeno, "Fuzzy identification of systems and its application to modeling and control," IEEE Trans. On System, Man and Cybernetics, 15, 1985, pp. 116 – 132.
- [22] J.-S. Jang, C.-T. Sun, E. Mizutani, Neuro-Fuzzy and Soft Computing: A Computational Approach to Learning and Machine Intelligence. Upper Saddle River: Prentice Hall, 1997.
- [23] Z. Hu, Ye.V. Bodyanskiy, N.Ye. Kulishova, O.K. Tyshchenko, "A Multidimensional Extended Neo-Fuzzy Neuron for Facial Expression Recognition," International Journal of Intelligent Systems and Applications (IJISA), vol.9, No.9, 2017, pp.29 – 36.
- [24] G.C. Goodwin, P.J. Ramage, P.E. Caines, "Discrete time stochastic adaptive control," SIAM J. Control and Optimisation, 19, 1981, pp. 829 – 853.
- [25] S. Haykin, Neural Networks. A Comprehensive Foundation. Upper Saddle River: Prentice Hall, 1999.
- [26] A. Cichocki, R. Unbehauen, Neural Networks for Optimization and Signal Processing. Stuttgart: Teubner, 1993.
- [27] http://pics.psych.stir.ac.uk/2D_face_sets.htm
- [28] P. Lucey, J.F. Cohn, T. Kanade, J. Saragih, Z. Ambadar and I. Matthews, "The Extended Cohn-Kanade Dataset (CK+): A complete dataset for action unit and emotion-specified expression," Proceedings of IEEE workshop on CVPR for Human Communicative Behavior Analysis, San Francisco, USA, 2010.

Performance Evaluation and Comparison of Software for Face Recognition, based on Dlib and Opencv Library

Nataliya Boyko
*Department of Artificial Intelligence
Lviv Polytechnic National University
Lviv, Ukraine
Nataliya.i.boyko@lpnu.ua*

Oleg Basystiuk
*department of information systems and
networks
Lviv Polytechnic National University
Lviv, Ukraine
obasystiuk@gmail.com*

Nataliya Shakhovska
*Department of Artificial Intelligence
Lviv Polytechnic National University
Lviv, Ukraine
Nataliya.b.shakhovska@lpnu.ua*

Abstract— Overview and investigate time complexity of computer vision algorithms for face recognition. Main article idea is to compare two popular computer vision libraries, they are OpenCV and dlib, explore features, analyze pros and cons each of them and understand in what situation each of them suit the best. Method. The technologies of computer vision, which are used for face recognition was worked out. Research of two popular computer vision libraries was conducted. Their features are analyzed and the advantages and disadvantages of each of them are estimated. Examples of building recognition application based on histogram-oriented gradients for face finding, face landmark estimation for face orientation, and deep convolutional neural network to compare with known faces. The article generalizes the concept of face recognition. The scientific basis for facial recognition and the construction of a complete recognition system was described. The basic principles of the programs for face recognition are formulated. A comparative analysis of the productivity of both libraries in relation to - the time of execution to the number of iterations of the applied algorithms was presented. Also built two simple applications for face recognition based on these libraries and comparing their performance.

Keywords— *computer vision, face recognition, algorithms, performance, dlib, openCV, machine learning, HOG, face landmark estimation, DCNN, SVM*

I. INTRODUCTION

Nowadays technologies of artificial intelligence are actively developing; they open up huge possibilities in front of us. Analysis, forecasting, recognition came to a new level with the use of artificial intelligence and machine learning technologies. In recent years, an extremely promising field of research is computer vision. Technology in this area is most in demand in our everyday life.

One of the most popular computer vision problems that is being actively explored is face recognition. Moreover, the latest developments of technological leaders in the world are proof of this. The Apple Special Event in September presented the technology Face ID. The active introduction of recognition technology by social networks for tagging people in photos and Facebook as a leader in this direction. The algorithm developed by Facebook shows the recognition efficiency of 93%.

The result of this development was the creation of various types of libraries and APIs for face recognition.

Developed solutions especially for a particular area, or capable of solving several at once, written in a particular programming language or with the support of all popular languages. Often, among all the variety of tools, it's hard to understand which one is best suited for solving your problem.

This article will discuss computer vision, namely face detection based on dlib and OpenCV libraries. All the advantages and disadvantages of each of them will be appreciated, and the feasibility of using them in projects for constructing recognition systems. However, before the development and comparison of libraries, we need to find out the basic things and terms associated with this doctrine, we will also understand what tasks we face, as the developers of a successful software product for the face to face.

The research objective is to compare the performance of two popular computer vision libraries and build on their basis two simple recognition systems. To achieve this, the following tasks must be solved:

- consider the methods and basic principles of face recognition;
- analyze existing recognition technologies;

Perform a comparative analysis of the performance of OpenCV and dlib libraries based on the HOG method for searching and subsequent recognition to appear in Python.

II. LITERATURE REVIEW

In the present, technologies of computer vision are actively developing, with their help, we can solve problems more effectively, one of which is recognition. As a result of active development, developers receive a large number of libraries to solve problems associated with computer vision. Actually, in this article we face the task of determining the performance of these libraries. The best way to get detailed information about a particular system, library, or API is to get acquainted with the documentation that we will actually use for the research in this article [1, 2]. Works [6, 7] focus on the theoretical aspects of building a stable system for face recognition.

The researchers [3, 5] describe the actual methods and technologies for all stages of the development of the recognition system, since in the field of recognition, a huge number of unique solutions have been developed. In any system, there is a promising issue of its performance, and

especially this applies to recognition systems, so the authors [4] investigate the speed of the operation of recognition methods. Also, scholars [9, 10] describe the method of recognition using the method of support vectors machine (SVM), which can significantly improve the speed of the recognition process. Researchers [8] describe face landmarks estimation algorithms that are used to position faces. It enables you to increase the quality of the system by aligning the face for better recognition.

Based on the conducted analysis it was established that for the creation of the recognition system there are no uniform methods and technologies that would combine all stages of construction. Therefore, the work with a detailed description of technologies and tools for their implementation, the assessment of the advantages and disadvantages of each of them and performance research to address such problems is relevant.

III. MATERIALS AND METHODS

Before the developers of the recognition systems face, there are the following problems:

- Find faces - no matter whether the task of recognizing people in photographs, or video recognition, or anything else.
- Face positioning - photos are not often found on which a person stands directly in front of the lens, most often the face is turned, we face the task of positioning it as if the photo was taken directly.
- Defining unique facial features - this step can be called a full face recognition step (previous ones were preparatory), it analyzes the image and obtains unique digital values of the face.
- Identification of a person - we compare the received data with the data already available to us, if the data are similar, we will display the name of the person, if not, accordingly we have not known yet to us person.

This article will examine in detail all the steps to build a face recognition system and compare their implementation with the help of various libraries, as well as the speed of the work of each stage in different libraries of computer vision. Keep in mind that the research results may vary, depending on how the computer is filled with recognition. Also, all libraries that are used for research in this article are configured in the mode of use only CPU, without a GPU (NVIDIA CUDA).

The first step in the development of the system is important, since depending on how good the search was made. If we skip any face in this step, or we regard another object as a face in this step, then the recognition results can be considered unsatisfactory. One of the popular algorithms for finding facial expressions in the image was invented in 2005 by Neveet Dalal and Bill Triggs, HOG [2]. The algorithm's work consists of the following steps:

Calculation of gradient value - The most common way to calculate is to use a one-dimensional centered mask with subsequent filtering of the color or saturation of an image using such filtering kernels:

There are four methods of normalization, the best way to recommend this method:

$$\text{L2-norm: } f = \frac{V}{\|V\|_2 + e^2};$$

In formula: V is not normalized block vector, $\|V\|_k$ - the norm of the vector, k - some small constant (the exact meaning is not important to us).

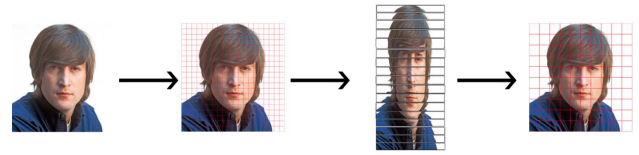


Fig. 1. Simplified visualization of the algorithm's work HOG.

After finding the face in the image, we face the task of positioning it, since in the vast majority of images, the faces are not centered, but are turned either from the angle and this will worsen any recognition later on. To solve this problem, use the face landmark estimation algorithm (face orientation estimation). The basic idea is to find 68 landmarks that are present on each face - the upper part of the chin, the inner edge of the eyebrows, the outer edge of the eye, the lower point of the nose, and the upper and lower points of the lip, and so on. After finding these landmarks with simple manipulations (rotation, increase or decrease), we get positioned in the center of the face.



Fig. 2. All 68 points of reference, which will recognize the algorithm on the face.

For positioning we will use the OpenFace additional library, it will perfectly "be friends" with the libraries we will use, it will look identical for both libraries, the performance will also be identical, so we will not compare it.

Here we come to the step that is directly related to the problem of face recognition. To distinguish between faces, we need to find unique features. And the first question that we face is how to organize this recognition. The easiest way to recognize faces is to compare the unknown face that was obtained in the second step with all the faces available in our database, hoping to find the same. Accordingly, during each recognition we refer to the database. The idea at first glance is good, but if we have huge data volumes, this will increase the sample time. Further consideration should be given to the option of passing these data to the cloud if we are talking about a large amount of data, which gives rise to the problem of delays between requests to and replies from the server. In addition, for most recognition projects, the speed of

recognition is in the first place, as you already understand, for such projects this method does not fit.

Because of the ineffectiveness and high computational complexity of the previous method, it was proposed to select several basic features of the face. We faced with the problem of choosing this rice. At first glance,, it seems that size or eye color, length of nose, size of the lips, eyebrows form - are the main characteristics by which a person recognizes faces. However, studies have shown that these features have no value to the computer face recognition. This problem is due to the fact, that the computer can not evaluate the face as a whole, but pixels the image.

To solve this non-trivial problem, it was suggested to use DCNN, which will be trained to identify 128 unique numerical facial features.

The process of training such a neural network works on the following principle:

- Upload the face image of a person we already know (classified).
- Upload another face image of the same person.
- Upload an image of another person's face.

The neural network will adjust the results of the obtained values so that 128 characteristics of the images loaded in steps 1 and 2 are as close as possible and the image loaded in step 3 differs from them as much as possible.

The idea of reducing the complex raw data to the list of computer generated numbers received the most development in the machine learning and was first used in the field of translation. The face-to-face approach, as suggested in the article, was firstly introduced by Google engineers in 2015.

The process of training the deep convolutional neural network is to generate unique numerical features of the face is a complex process that requires a huge database of faces and a significant computing ability of the computer.

Even using NVidia graphics card Telsa, since greater speed model train on NVidia graphics cards companies with presence support CUDA, the learning process took about 24 hours. However, as soon as the neural network is trained, it will be able to generate unique characteristics for any person, even for one that it has never seen before. So, this step is extremely important. The neural network is trained only once, but it will depend on the effectiveness of our face recognition system. If you can not train your own neural network to generate features, you can use an already trained network.[3].

The last step of the algorithm is to compare existing in our data, namely 128 features some faces that we received in the previous step, with all of our data about the people, if the data coincide, we can tell who is shown in the picture. This step can also be implemented in two ways:

- Database - as above described, with large volumes of data, this greatly reduces application performance. However, as before, this method of organization can be used for small volumes of data when it is a task of identifying a small group of people and there is not

enough photos for quality training of the method of reference vectors.

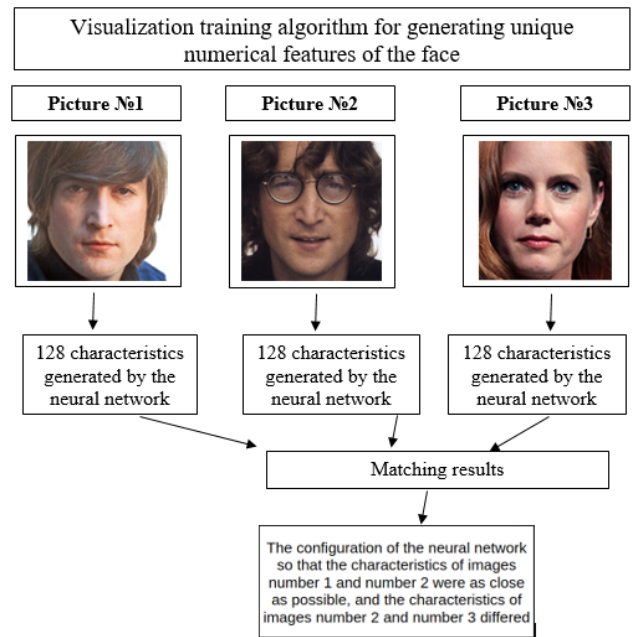


Fig. 3. Visualization of the learning algorithm to determine the unique numerical features of the face in the future.

- Support vector machine (SVM) - using existing features we train a classifier, the more homogeneous data set, the better for us because with this we can more accurately determine the face. Why the more homogeneous data, the better? For the classifier, we pass the value in the form of data set:

$$(\vec{x}_1, y_1), \dots, (\vec{x}_n, y_n),$$

where y is 1 or -1, and each one indicates the class to \vec{x}_i which the point \vec{x}_i belongs. During the training, support vector method, our challenge is to find the maximum separation hyperplane (see Fig.3), simply put, the dataset is divided into classes so that they are difficult to confuse during comparison. The classifier's working time is several milliseconds, ideal for face-to-face identification.

After weighing the pros and cons we will discuss identification method using support vector method. At this stage, there is also no reason to compare performance because the data set for learning is the same, accordingly, we create a classifier for two libraries, so at this point the speed will be the same for the two libraries.

Also, please note some features that can help you to improve the learning [7]:

- 1) Between the data (in our case, it's 128 facial features) there should be clear intervals so that they can be divided into groups, this depends on how well-trained the algorithm defines unique facial features.
- 2) Huge volumes of data will not provide the desired gain, but will only increase the learning time of the algorithm, to increase productivity, use data filtered

from the noise and any data from one class does not overlap the data of another.

IV. EXPERIMENTS

Let's make a study of all steps for building a recognition system based on the methods and technologies described above. Our first step is to search for the face based on the method proposed above, namely HOG. The result will be an array of coordinates for each found face in the photo.

An example of organizing this algorithm with dlib library tools:

```

Face_detector_dlib.py:
# Import the system and dlib libraries
import sys
import dlib
from skimage import io
# Call the basic function library to find faces
face_detector = dlib.get_frontal_face_detector()
# Go through all entered images
for images in sys.argv[1:]:
    print("Processing image: {}".format(images))
    image = io.imread(images)
    faces_detected_image= face_detector (image, 1)
    io.imsave('face_recognition.jpg', faces_detected_image)
    print("Number of faces detected:{}".format ( len(
faces_detected_image))
# Print the result
win = dlib.image_window()
win.clear_overlay()
win.set_image(image)
win.add_overlay(dets)
#Wait for "Enter" button to continue
dlib.hit_enter_to_continue()

```

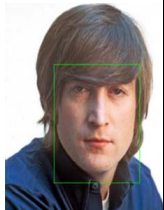
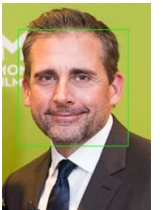
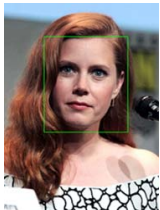
An example of organizing this algorithm with OpenCV library tools:

```

Face_detector_opencv.py:
# Import the system and OpenCV libraries
import sys
import cv2
from skimage import io
# Call the basic function library to find faces
lbp_face_cascade = cv2.CascadeClassifier('data/
lbpcascade_frontalface.xml')
# Go through all entered images
for images in sys.argv[1:]:
    print("Processing image: {}".format(images))
    image = io.imread(images)
    faces_detected_img = detect_faces(lbp_face_cascade, image,
scaleFactor=1.2)
    io.imsave('face_recognition.jpg', image)
    print("Number of faces detected:{}".format ( len(faces))
# Print the result
plt.imshow(convertToRGB(faces_detected_img))
#Wait for "Enter" button to continue
cv2.waitKey(0)

```

The results of the performance are as follows:

	Experiment №1	Experiment №2	Experiment №3
OpenCV			

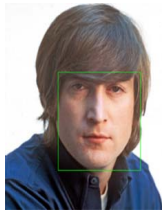

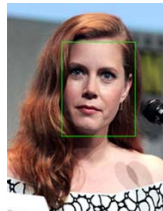
Recogniti on time (sec.)	0.5652	0.0934	0.6121
dlib			
Recogniti on time (sec.)	1.6584	0.2163	1.6889

Fig. 4. Comparison of the results and the time spent searching for faces using OpenCV libraries and dlib

As we can see at this step, the OpenCV library demonstrates better performance.

Going to the next step where we position the found face, I remind that for this we will use the face landmark estimation method, which we build on the basis of an additional OpenFace library.

```

Face_landmark_estimation.py:
# Import the system and OpenFace libraries
import sys
import dlib
import openface
from skimage import io
# Call the basic function library to find faces
face_detector = dlib.get_frontal_face_detector()
face_pose_predictor = dlib.shape_predictor(predictor_model)
face_aligner = openface.AlignDlib(predictor_model)
# Go through all entered images
for i, face_rectangle in enumerate(detected_faces):
    print("- Face #{} found at Left: {} Top: {} Right: {} Bottom: {}".format(i,
face_rectangle.left(), face_rectangle.top(), face_rectangle.right(),
face_rectangle.bottom()))
    pose_landmarks = face_pose_predictor(image, face_rect)
    alignedFace = face_aligner.align(534, image, face_rect, landmarkIndices
= openface.AlignDlib.OUTER_EYES_AND_NOSE)
# Save the result
io.imsave('aligned_face_{}.jpg', image, alignedFace)

```

The results are as follows:



Fig. 5. Visualization of the algorithm positioning faces

After receiving the characteristics, we only find such characteristics from the available data.

As described above, we will use SVM for this. Here's what an advertising implementation looks like:

```

SVM_example.py:
import sys
import numpy as np
from sklearn import svm
x = np.vectors();
y = np.array();
x.append(np.vector([1, 2, 3, -1, -2, -3]))
y.append(+1)
x.append(np.vector([-1, -2, -3, 1, 2, 3]))
y.append(-1)
svm = dlib.svm_c_trainer_linear()

```

```

svm.be_verbose()
svm.set_c(10)
classifier = svm.train(x, y)
print("prediction for first sample: {}".format(classifier(x[0])))
print("prediction for second sample: {}".format(classifier(x[1])))
with open('saved_model.pickle', 'wb') as handle:
pickle.dump(classifier, handle)

```

V. RESULTS

The final version of the experiment are two face detection systems, one is based on a dlib library, the second is based on OpenCV. The results of their work submitted to fig.6, which provides an interface of applications, in which we see the detected face.

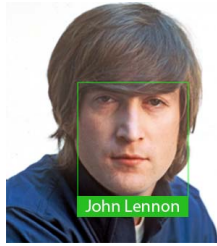


Fig. 6. Example of Face Detection

Based on the analysis of the submitted OpenCV and dlib libraries, it has been established that there is no single methods and technologies to create a distributed recognition information system that would combine all stages of the system construction. There is a huge number of methods for searching, positioning and in general for organizing the recognition process. By choosing technologies you should be very careful because depending on your needs you should use certain methods. Therefore, work with a detailed description of technologies for recognition is extremely relevant at present, as well as the development of new technologies and a way to solve this pressing problem.

VI. DISCUSSION

As mentioned above, currently the technology of computer vision and tools for face recognition are actively developing. For qualitative recognition, faces are developed not only software products but also unique hardware solutions. Global technology leaders invest in research and development a lot of money because this problem has wide opportunities for use. The direction for the development of face recognition is primarily the protection and increased protection systems. Currently the

most used in entertainment, I believe this trend will grow also. Proposed approach can be used for big data processing [11].

VII. CONCLUSIONS

As has been researched in this article, the OpenCV library is more productive, has better performance for face detection and detection. It also means that with OpenCV, it's better to build recognition applications for the IoT platform. Note that only HOG algorithm has been explored while searching for other algorithms, such as the Haar cascade, it works longer, but works out more in detail, if there are plenty of photos in the future for many, it is advisable to think about using this method. However, the design logic and the key points of creating an application recognition, was discussed in this article.

REFERENCES

- [1] OpenCV: OpenCV Tutorials [Electronic resource] – Access mode: https://docs.opencv.org/master/d9/df8/tutorial_root.html
- [2] Dlib Python API Tutorials [Electronic resource] – Access mode: <http://dlib.net/python/index.html>
- [3] Z. Rybchak, and O. Basystiuk, “Analysis of computer vision and image analysis technics,” ECONTechMOD: an international quarterly journal on economics of technology and modelling processes, Lublin: Polish Academy of Sciences, vol. 6, no. 2, pp. 79-84, 2017.
- [4] Face Detection Algorithms and Techniques [Electronic resource] – Access mode: <https://facedetection.com/algorithms/>
- [5] A toolkit for making real world machine learning and data analysis applications [Electronic resource] – Access mode: https://github.com/davisking/dlib/blob/master/python_examples/face_detector.py
- [6] R. Raja, Face Detection Using OpenCV and Python [Electronic resource] – Access mode: <https://www.superdatascience.com/opencv-face-detection/>
- [7] R. Raja, Face Recognition Using OpenCV and Python [Electronic resource] – Access mode: <https://www.superdatascience.com/opencv-face-recognition/>
- [8] A. Rosebrock, Facial landmarks with dlib, OpenCV, and Python [Electronic resource] – Access mode: <https://www.pyimagesearch.com/2017/04/03/facial-landmarks-dlib-opencv-python/>
- [9] Support Vector Machines [Electronic resource] – Access mode: <http://scikit-learn.org/stable/modules/svm.html>
- [10] F. Schroff, D. Kalenichenko, and J. Philbin, FaceNet: A Unified Embedding for Face Recognition and Clustering [Electronic resource] – Access mode: https://www.cv-foundation.org/openaccess/content_cvpr_2015/app/1A_089.pdf
- [11] N. Shakhovska, “The method of Big data processing,” XII International Conference on Computer sciences and information technologies (CSIT), Lviv, Ukraine, pp. 122-125, 2017

Partial Motion Blur Removal

Andriy Klyuvak
ABTO Software LLC
Lviv, Ukraine
klyuvak@ukr.net

Oksana Kliuva
Department of Business Economy and
Information Technology
Lviv University of Business and Law
Lviv, Ukraine
oksana_klyuvak@ukr.net

Ruslan Skrynkovskyy
Department of Business Economy and
Information Technology
Lviv University of Business and Law
Lviv, Ukraine
uan_lviv@ukr.net

Abstract—The work is focused on two types of images partial blur (a single object blur and background blur). Complex approach to the deblurring of such images is proposed. It includes such aspects as automated blur detection, blurred image PSF estimation, blurred patches extraction and processing etc.

Keywords—image deblurring, motion blur, single object, PSF, computer vision, image reconstruction, partial blur.

I. INTRODUCTION

One of the most difficult problems in the field of digital images processing is deblurring of images, distorted by partial blur. In contrast to stochastic blur caused by shaking of camera, the partial blur occurs in two following cases: Motion of one or more objects with fixed camera. Here we have blurred object image against unblurred back-ground (type 1); Camera strictly follows the moving object. Here we have the opposite situation: unblurred object image against blurred background (type 2). In the present paper we suggest a complex approach to removal of the first type of partial blur as well as some aspects of second type motion blur elimination. The complicity of this task is caused by the necessity of defining moving object shape, its motion's parameters (i.e. PSF – Point Spread Function also known as a convolution kernel), its deblurring without influencing image background, and, the most difficult, separation of areas which simultaneously belong to foreground (i.e. moving object) and background.

II. GENERAL APPROACH

In the initial time t_0 camera shutter opens and elements of sensitive matrix start acquiring some colours. At this infinitely short time period both moving foreground and background can be treated as not blurred. At time point t_n the shutter closes. During exposure time moving object covers some distance, which, taking into account discrete nature of digital photography, can be estimated as some number of points, say m . Hence, time interval $dt = t_n - t_0$ can be subdivided into m equal intervals. During each of these subintervals every point of moving object image puts its colour onto different picture point, adding its value to the value formed in this point before. Inner points of object apply its values to other inner points values, so here we have ordinary deblurring problem, which can be solved by applying of one of known deblurring algorithms, for example Richardson-Lucy's. But the picture on edges is quite different. These edges occupy the areas on the object perimeter on both sides of motion direction, and it as m points thick. As a result, the interfusion of background and foreground (moving object points) colours is observed. As whole exposure time is subdivided to m intervals, we can

state that every buffer point (point of mixed edges) of the $1/m$ time interval is influenced in $1/m$ by a moving object's corresponding point and in $1-1/m$ by other points, both background point colour and other foreground points, which also 'flew' over that position. If this point lies in the uttermost edge position, the point's colour is comprised of $1/m$ of foreground colour and $1-1/m$ of background colour. If this point is next, its colour is comprised of $1/m$ of this point's colour and $1-2/m$ of background colour and so on for whole buffer area depth, the last point of which would be comprised for $1/m$ of background colour [1;2]. This would be true in the case of uniform motion, in the case of nonuniform motion these proportions would shift (the shorter time some point stays in some position, the smaller impact it makes on forming the resulting colour in this position) but overall regularity would be the same. In order to deblur this blurred segment we need, first, to estimate blur parameters i.e. to calculate PSF (Point Spread Function) values. Discrete PSF in general case is a matrix which represents a blur kernel that influences an image. The sum of all elements of PSF matrix equals to 1. Signal distortion described by some blur kernel is that the distorted image's every point's value is calculated as follows:

$$d_i = \sum_j p_{i,j} u_j \quad (1)$$

where $p_{i,j}$ is a element of PSF matrix in position (i,j) which denotes what part of j point colour is observed in i point; d_i - i -th point colour value after convolution operation; u_j - j -th point ideal colour value (not distorted).

PSF dimensions and nonzero elements distribution depend on motion speed, uniformity and trajectory and also on exposure duration. Nonzero PSF elements are situated on trajectory lines, drawn by moving object, while camera is taking the picture. Their values correspond to motion speed between time intervals t_n and t_{n+1} , which equals to $1/m$ exposure time, where m – nonzero PSF elements number. Hence, all nonzero PSF elements are equal in case of uniform motion and placed in line in case of rectilinear motion. Horizontal motion would create PSF in form of row vector and vertical motion – column vector. Taking that, we can conclude that this buffer zone follows PSF matrix configuration: if PSF matrix is column vector 5 elements long, then buffer zone is 5 points (pixels) thick and is situated above and below moving object's image. That's why motion parameters estimation must precede all other operations with blurred domain.

III. MOVING OBJECT PSF ESTIMATION

A. Motion blur PSF peculiarities

There are two approaches to image reconstruction:

- Non blind reconstruction based on preliminary calculated PSF. PSF estimation can be based on information about distortion character. Hence, at first we can try to estimate a PSF, and then deconvolute an image using one of standard deconvolution methods with known PSF;

- Blind reconstruction by one of iterative algorithms. At first image undergoes reconstruction and then PSF is extracted, using the difference between distorted and reconstructed images. The sense of PSF estimation in this case is to obtain information about distortion character [2].

Let's take type 1 blur. During blur, every point of the moving object moves the trajectory of the whole object. That is why successively overlapping onto points of its trajectory moving point leaves its trace. If this point is situated in the zone with small gradients of colour, then its trace is overlapping with the traces of neighbouring trajectory points, which are similar by colour value. Quite a different situation would be with the points that differ in their background. Since almost every object is not absolutely homogeneous, there exist such points of interest. They could be corners, edge points, or other separated points. During their motion these points will be surrounded by other points of uniform background, and will leave their traces, which reproduces the trajectory of whole moving object. In fact, points, located on the motion trajectory in both sides, have colour of background, which by overlapping give approximately the same colour. Similarly, the neighbour points that are not on motion trajectory also form a colour of background near key points traces. This idea can be illustrated on Fig. 1. As shown on Fig. 1, a corner of a different colour A (within positions A1-A2-A3) and point with the other colour B (within positions B1-B2-B3) are performing motion together with whole moving object, to which they belong. Since they are related to the same object and have invariable positions, they move along the same trajectory. At the initial moment of frame's exposure corner's position is A1, whereas initial position of the point is B1. Final positions (at the final moment of frame's exposure) are A3 and B3 respectively. Positions A2 and B2 are represented in a moment of time between initialization and finalization of current frame exposure. As it is shown in Fig. 1, whole area *a* (from initial position of zone A1 to final position A3) will have colours clearly distinctive from those of neighbourhood, because area *a* is the result of spreading of the corner A with some admixture of background colours (at some places – «bufferzone» – these colours are mixed in definite proportions, and in other places only colours of the corner are mixed), whereas outside areas *a* and *b* only background colours influenced resulting points' colours.

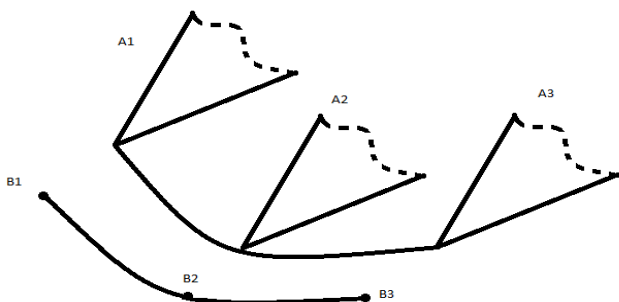


Fig. 1. Motion tracking inside moving object image.

The bottom edge of blurred area *a* will duplicate motion trajectory of the object. The same is observed for motion trace of point B: this trace will be standing out against a background and be coincident with motion trajectory of feature point. This can be illustrated using image and its corresponding PSF, which can be found in the paper [2]. Figure 2 illustrates how a small region with different colour has left its visual trace on the image and that this trace clearly corresponds to non-zero elements distribution of PSF matrix.

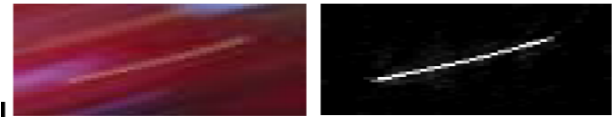


Fig. 2. Blur visual trace and corresponding PSF

So, in our case, the blurred image has the following: there are two regions in the image, which differ by colour from average colour of the background, and the shapes of bottom borders of those regions are equal. Using this, we can establish the main postulate of this method: the distribution of non-zero elements of PSF matrix of an image distorted by motion blur will correspond to the most frequent variant of edges of zones with colour distinguishable from neighbourhood.

B. Moving Object trajectory estimation.

As far as the method is based on colour segmentation, the algorithm involves gradients processing. Those gradients get checked for occurrence of frequent elements. The algorithm consists of following steps: conversion of an input coloured image to greyscale; gradients calculation; absolute values calculation; processing of partial derivatives in horizontal and vertical directions (two passes): a) calculation of maximum value; b) for every value of partial derivatives from 50 to 100% of maximum value the following is done:

- I. Threshold determination. Every run of «sifting» loop iteration differs only by threshold value. Using these values regions with different colour contrast can be outlined at every iteration.

- II. Estimation of regions where the values of partial derivatives are greater or equal to threshold value (estimation of blobs). The result of this step is a set of blob (mask).

- III. Excluding blobs with an area that is equal to 1. Isolated points with colours different from their neighbours are always present in images because of camera noise, but neither of these points can be PSF candidates, because such PSF represents no changes: after convolution with single element PSF any image transforms to itself.

- IV. Extraction of right and left or upper and bottom edges of every blob. Hence every blob gives us two new edge blobs; herewith the old one is to be discarded.

- V. Excluding blobs with an area that is equal to 1.

- VI. Searching for identical blobs.

Here we substituted the principle of absolute identity with the principle of almost identity [3]. We formulate this principle as: blobs are considered to be identical, either if they are absolutely equal to each other or if one of them completely duplicates a part of another, i.e. can completely be located in it. We can determine that one blob is equal to some part of another one using such method: two arrays – sums of all columns and sums of all rows of smaller blob –

must be subarrays of corresponding arrays of a larger blob, with the beginning in the same point. This case is shown in Fig. 3, where two subarrays of smaller blob start from point (x1, y1) in larger blob. Sometimes however starting point of a smaller blob lies on either strictly vertical or strictly horizontal segment of a larger blob (Fig. 4).

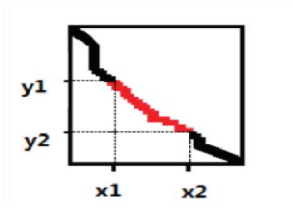


Fig. 3. Almost identical blobs

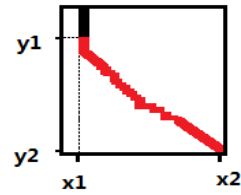


Fig. 4. Almost identical blobs with common point in vertical region

With such comparing on almost identity, the comparison function returns the larger blob. This larger blob will further be used for pairwise comparing process instead of previously used blob. Also, two blobs cannot be identical, if one of them significantly (more than 5 times) larger than the other blob, or if one of them contains gaps.

5. The variant, wherein the maximum of identities was found, is returned as final result.

Developed algorithm was implemented using MATLAB and was tested using a set of images with known motion blur PSF. In all cases the motion direction was estimated correctly. Received results completely reproduce trajectory of distorting motion, and its length deviates from its real value by 15-20%. Input images and results of processing in the form of binary PSF matrices are shown in Fig. 5, 6 and 7.



Fig. 5. Example 1



Fig. 6. Example 2



Fig. 7. Example 3

C. PSF values calculation.

The result of first step is PSF matrix structure, i.e. its dimensions and position of non-zero elements. On the second step the motion continuity of each point is estimated. Though in case of partial blur this step is performed after blur detection (4), it is integral part of PSF estimation algorithm. To determine this estimation we have to consider features of forming of blurred image. As far as digital image is obtained during quantization by photosensitive matrix, then every point during its motion appears in several positions. At that, the value of its colour saturation depends from time of this point being in current position. And time, in its turn, is defined by influence of point velocity – the less velocity, the greater time point stays in the position. Naturally the sum of influences of a point in every position of its trajectory equal to one.

$$\sum_{i=1}^m \sum_{j=1}^n a_{i,j} = 1 \quad (2)$$

where $a_{i,j}$ is dispersion coefficient of i -th point in j -th position of motion trajectory, n is the length of trajectory, equal to count of non-zero elements of PSF matrix.

If the i -th point is distributing its colour along of the entire trajectory, then its influences are present at every position of this trajectory. This makes a ground to assume, that between j -th positions of trajectories of all points of blurred image area there exists a relative correlation. Let's go into the matter of correlation in detail on example of two neighbour positions j and $j+1$ on a motion trajectory of two different points with dispersion coefficients $a_{1,j}$ and $a_{2,j}$ respectively. Let's denote their colours as c_1 and c_2 . In a blurred image there are four points with colours $S_{1,j}$, $S_{1,j+1}$, $S_{2,j}$, $S_{2,j+1}$.

$$S_{1,j} = a_{1,j}c_1 + \omega_{1,1} \quad (3)$$

$$S_{1,j+1} = a_{1,j+1}c_1 + \omega_{1,2} \quad (4)$$

$$S_{2,j} = a_{2,j}c_2 + \omega_{2,1} \quad (5)$$

$$S_{2,j+1} = a_{2,j+1}c_2 + \omega_{2,2} \quad (6)$$

where $\omega_{1,1}$, $\omega_{1,2}$, $\omega_{2,1}$, $\omega_{2,2}$ are the other colour components in blurred image, except a part, that is super induced in the result colour by colours c_1 and c_2 of two examined points of original image.

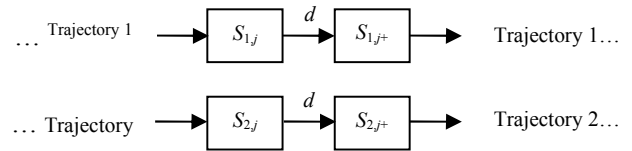


Fig. 8. Schematic relative positions of points in a blurred image

Two points move two parallel trajectories, both as translation result of each other (fig. 8). Let's calculate colour differences between two neighbour points in every trajectory:

$$d_1 = S_{1,j+1} - S_{1,j} \quad (7)$$

$$d_2 = S_{2,j+1} - S_{2,j} \quad (8)$$

Since displacements of two points are equal, their ratio can be found using (3)-(6):

$$\frac{d_1}{d_2} \sim \frac{(a_{1,j+1}c_1 + \omega_{1,2}) - (a_{1,j}c_1 + \omega_{1,1})}{(a_{2,j+1}c_2 + \omega_{2,2}) - (a_{2,j}c_2 + \omega_{2,1})} \quad (9)$$

$\omega_{i,j}$ are formed as the result of identical process of points' colours overlaying. That is why in the respect to trajectory neighbouring points' gradients they can be omitted. Formula (9) hence can be transformed as follows:

$$\frac{d_1}{d_2} \sim \frac{c_1(a_{1,j+1} - a_{1,j})}{c_2(a_{2,j+1} - a_{2,j})} \quad (10)$$

Unknown coefficients a_{ij} are equal in both cases, hence:

$$\frac{d_1}{d_2} \sim \frac{c_1}{c_2} \quad (11)$$

It may be concluded that, colour differences between pixels lying along trajectory lines have statistical significance and are correlating with those of other parallel trajectory lines.

The subsequent operations are carried out for every point of blurred image, if the aperture in this point does not out-step the blurred region of an image. Absolute values of chaining differences between neighbouring points are calculated starting from a current point of an image and along all non-zero elements of PSF matrix (since they are always arranged in a line):

$$d_{ij} = \left| C_{i+x,j+y} - C_{k+x,l+y} \right| \quad (12)$$

where C is processed image; i,j are coordinates of PSF matrix elements; x,y are coordinates of a point that is an input point for current iteration; k,l are coordinates of the next non-zero element of PSF matrix (that is trajectory neighbouring point to the current). Depending on trajectory character there are several variants available: 1) $k=i+1; l=j$; 2) $k=i; l=j+1$; 3) $k=i+1; l=j+1$.

After examining of all points, the mean values of differences for each element can be calculated.

$$D_{ij} = \frac{\sum_{k=1}^p d_{ijk}}{p} \quad (13)$$

where p is number of obtained matrices.

To meet condition, which was defined by restriction (2), the obtained result is to be normed:

$$PSF_{ij} = \frac{D_{ij}}{\sum_{j=1}^m \sum_{i=1}^n D_{ij}} \quad (14)$$

where m, n are dimensions of PSF matrix.

The overall result of this processing is PSF:

$$PSF = \left\langle PSF_{i,j} \right\rangle_{\substack{i=1..n \\ j=1..m}} \quad (15)$$

IV. BLUR DETECTION

It follows from the above that **every pixel of a blurred image must have smaller colour difference with its trajectory neighbours than with other neighbours**. This statement is the basis of motion blur detection method, proposed in this work.

Blur detection starts with calculation of colour standard deviations. It is carried imposing PSF-trace. PSF-trace reveals motion trajectory, hence we know trajectory neighbours of every pixel in an image. Every pixel has two major trajectory neighbours. Standard deviation of an every pixel is calculated in respect to these neighbours only. To this end we impose PSF-trace to an every pixel in its central element. These calculations are carried separately for every colour channel, and the final value for every pixel is the maximal value among all colour channels for that pixel. After that matrix values are normalized to be in range [0; 1].

As the result of previous action we get the matrix of trajectory-neighbour oriented standard deviations. But this matrix is to be binarized in order to become blur mask. If some point lies in blurred zone, its deviation value must be significantly lesser. So in order to find unblurred area these values are to be segmented by some minimal threshold, that is, if some value is greater than threshold value, then this

pixel belong to unblurred area, otherwise to the blurred one. Blurred area mask is inverted unblurred area mask. Usually calculated or selected threshold value lies in range [0.05; 0.1]. The result of standard deviation matrix binarization is the mask of not blurred zone. The following example of blur detection shows the result of the following MATLAB code execution. The code is based on not blind detection method.

Listing 1. Nonblind blur detection MATLAB code
`s = stdfilt(f, PSF > 0);
s = max(s,[],3);
bw = normal(s)>threshold;`



Fig. 9. Blurred image of type 2

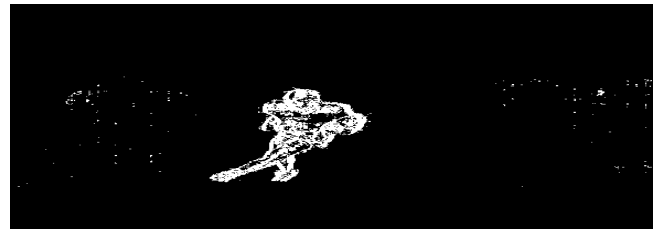


Fig. 10. Unblurred area mask before additional morphological processing

V. BLURRED/UNBLURRED PATCH EXTRACTION AND DEBLURRING

For type 1 motion blur blurred domain must be extracted from the native background and carried to homogeneously black background. Black background is chosen because black points are represented with zero values, hence are unable to influence results of subsequent operations and distort the resulting image. Simple copying of blurred object image with buffer zone has no sense, because part of admixed background will be copied as well. That's why foreground and background components of buffer zone must be properly separated taking into account rules of their mutual interference. The separation is carried with known PSF which is calculated while processing inner areas of moving object blurred image. Having PSF values we can estimate what part of buffer zone point colour is represented by foreground and background components. Removal of background component is equivalent to taking moving object image to black background. To do that we copy object's image to black background and calculate new values of buffer zone point's colours starting with edge points to the whole PSF depth using this formula:

$$u_j^* = \frac{u_j - b + a}{a} \quad (16)$$

where: u_j - point's colour old value; b - background colour; a - sum of PSF matrix values for current point and previously calculated points [4].

When we face opposite situation – the not blurred part must be preserved in its primitive state and must not affect the rest of the image at deblurring. We do not need to carry

anything to the black background, but unblurred patch is simply extracted according to the mask of unblurred patch.

Hence, the resulting background gets black hole in this place, which is to be filled by the method of image integration, otherwise black blot will distort the whole image. Reconstructing an image out of its derivatives using Frankot-Chellappa method [5] a new image with filled hole can be obtained [6]. This reconstructed image is suitable for successive deblurring.

In both cases deblurring is most often performed by one of non blind deblurring algorithm i.e. deblurring with known PSF (Viener filtration, Richardson-Lucy etc). In case of type 2 blur or total blur the technique of 4-sided mirroring can be applied in order to reduce the well-known ringing effect. After deblurring, processed image has to be placed back to its native canvas.

VI. PROCESSED ZONE IMPOSITION

The processed part of the image (foreground moving object), extracted and stuck to black background earlier, is to be imposed back to the native background. All pixels of the input unprocessed image that correspond to the blurred image are removed, i.e. substituted with black colour. To determine, which pixels are to be removed we impose the mask of blurred image, calculated previously (IV Blur detection). Hence the blurred zone is cut out of the canvas. Deblurred image is added to the background image. In this step two matrices are simply added: black pixels values cannot affect resulting values. Only nonzero values of both matrices remain in the resulting image, and deblurred patch takes its place.

Since deblurred image is always smaller than blurred, double-sized buffer zone on one of the edges of deblurred patch will remain black. It is filled with dominant colours on the background side of the edge. In case of second type motion blur this step is omitted, as the shapes of the patch and of the filled hole on the image background coincide. Foreground/background edge is smoothed, since after the imposition and colour refilling the edge may seem to be excessively sharp and hence look unnatural. For the first thing seam area is to be selected to perform smoothing over it. It is defined as the perimeter of the blurred zone mask, but expanded in both directions. Secondly, averaging filter is applied to the area, selected with this mask. This averaging creates more natural colour distribution and the image as whole looks now more natural, since the seam fades. Hence, whole image is assembled again (fig. 11) [7;8;9].



Fig. 11. Deblurred image.

VII. CONCLUSION

The paper is devoted to the methods of reconstruction of images distorted by partial motion blur – motion blur that affects not whole image but some part of it. The paper contains approaches to the deblurring of two types of partial blur: a) an image of a blurred moving object against unblurred background and b) an image of unblurred object surrounded with blurred background. One of the most important points of the paper is Point Spread Function (PSF) estimation method. It is based on the method of distorting motion trajectory estimation. Trajectory estimation method involves sharp edges motion tracing by means of colour gradients exploration. Hence the proposed method is able to calculate the trajectory and PSF from single image without any other preliminary data. Other developed methods are: the method of automated blur detection and the methods of blurred/unblurred areas separation. The first is used to detect the shape of blurred area and is based on pixels' colours standard deviation statistics and information about PSF character. This approach enables to detect a blur caused by some definite PSF with higher accuracy. The separation methods are utilised to separate blurred area for different partial blur types.

REFERENCES

- [1] S. Schuon, "The Nature of Motion Blur" http://ai.stanford.edu/~schuon/deblur/download/schuon_nature_of_motion_blur.pdf
- [2] Dong Gong, Jie Yang, Lingqiao Liu, Yanning Zhang, Ian Reid, Chunhua Shen, Anton van den Hengel, and Qinfeng Shi, "From Motion Blur to Motion Flow: a Deep Learning Solution for Removing Heterogeneous Motion Blur," Cornell University Library, Computer Vision and Pattern Recognition, Submitted on 8 Dec 2016 <https://arxiv.org/pdf/1612.02583v1.pdf>
- [3] R. Fergus, B. Singh, A. Hertzmann, S.T. Roweis, and W.T. Freeman, "Removing Camera Shake from a Single Photograph," ACM Trans. Graph. (SIGGRAPH), vol. 25(3) pp. 787–794, 2006
- [4] D. Peleshko, A. Klyuvak, M. Peleshko, and N. Kustra, "Motion Blurred Images PSF Estimation method," VIth International Scientific and Technical Conference "Computer Science And Information Technologies" (CSIT 2011), Lviv, Ukraine, pp.23-24, November 16 – 19, 2011.
- [5] D. Peleshko, A. Klyuvak, N. Kustra, and M. Navytka, "Deblurring of images distorted by separate objects' motion using blur kernel," V International Scientific and Technical Conference "Computer Science And Information Technologies" (CSIT 2010), Lviv, Ukraine, pp. 34-35, October 14-16, 2010.
- [6] R. T. Frankot and R. Chellappa, "A Method for Enforcing Integrability in Shape from Shading," IEEE PAMI vol 10, no 4, pp 439-451, July 1988.
- [7] D. Peleshko, O. Makoweychuk, and A. Klyuvak, "Reconstruction of Images Distorted by Partial Motion Blur with Not Blurred Zones Inclusions," Proceedings of the VII International Scientific and Technical Conference, "Computer Science And Information Technologies" (CSIT 2012), Lviv, Ukraine, pp.125-126, November 2012.
- [8] Q. Shan, J. Jia, and A. Agarwala, "High-Quality Motion Deblurring from a Single Image," ACM Transactions on Graphics, vol.27, no.3, 2008.
- [9] Q. Shan, W. Xiong, and J. Jia, "Rotational motion deblurring of a rigid object from a single image," ICCV 2007, IEEE 11th International Conference, pp. 1-8, 2007.

A Generalized Description for the Perceived Contrast of Image Elements

Sergei Yelmanov
Special Design Office of Television Systems
Lviv, Ukraine
sergei.yelmanov@gmail.com

Yuriy Romanyshyn^{1,2}
¹Department of Electronics and Computer Technologies
¹Lviv Polytechnic National University
²University of Warmia and Mazury
¹Lviv, Ukraine, ²Olsztyn, Poland
yuriy.romanyshyn1@gmail.com

Abstract—In this paper, the problem of assessing the perceived contrast of image elements for no-reference measurement of the global contrast of complex (multi-element) images is considered. A new method for assess the perceived contrast of elements of complex image is proposed on the basis of measuring of the contrast of these elements on a pre-normalized image with subsequent correction of the contrast value taking into account the dynamic range of the primary (original) image. A new generalized description of the perceived contrast of the image elements for different definitions of the contrast kernel is suggested. New definitions of the weighted and relative contrast of the image elements are proposed. A comparative analysis of the proposed and known definitions of the weighted and relative contrast of image elements was carried out.

Keywords—image, perceived contrast, image elements, global contrast, weighted contrast, relative contrast.

I. INTRODUCTION

Wide applying of modern technologies in imaging and image processing requires the solution of the task of no-reference assessing of the perceived quality of image [1].

Global contrast is the most important quantitative characteristic, which largely determines the overall perception of image [2, 3, 4]. At present, the development of new effective technologies for no-reference measuring the perceived contrast of complex images is relevant as never before [5].

The global (generalized) contrast of a complex (multi-element) image is determined by the contrast of its elements (objects and background). The contrast of the image elements (two objects or the object and the background) determines the difference in their objective characteristics. The contrast of the two image elements is usually defined on the basis of the difference in their brightness values [6]. The method of measuring the contrast of image elements appreciably defines the accuracy of measuring the perceived contrast for the image as a whole. Currently, there are various approaches to measuring the contrast of image elements [6, 7]. However, the known definitions of the contrast of image elements have a number of disadvantages that significantly reduce the effectiveness of their practical use [7, 8, 9]. To address these disadvantages, we propose a new method of measuring the perceived contrast of elements of complex image on the basis of measuring of the contrast of these elements for normalized image and of the dynamic range of the initial image.

The problem of increasing the accuracy of measuring the perceived contrast of elements of complex image is considered in this paper. The object of the study is the process of measuring the contrast of complex images to assess their quality. The purpose of the work is to increase the accuracy of measurement the perceived contrast of elements of complex image. The subject of the study is methods of measuring the perceived contrast of elements of complex image. The main known approaches to measuring the image contrast are considered (Section II). A new method of measuring the perceived contrast of two image elements (objects and background) is proposed by measuring their contrast on a normalized image with subsequent correction of the contrast value taking into account the dynamic range of the initial image (Section III). A generalized description of the perceived contrast of elements of complex images for various definitions of the contrast kernel is proposed. New definitions of perceived contrast of image elements for weighted and absolute contrast are proposed. The research of known and proposed definitions of a weighted and relative contrast to evaluate the efficiency of measuring of perceived contrast of image elements was carried out (Sections IV and V).

II. THE DEFINITION OF IMAGE CONTRAST

The global contrast of a complex multi-element image is determined on the basis of contrast values for all pairs of its elements (objects and background) [2]. The contrast of the two elements of the image (two objects or an object and a background) characterizes the distinction in their objective quantitative characteristics [6, 7, 10, 11].

A. The definition of global contrast for complex images

At present there are various approaches to assessing the global contrast of a complex multi-element image on the basis of measuring the values of contrast for all pairs of its elements.

The global (generalized) contrast of a complex multi-element image is usually defined as the average value of contrast for all pairs (i, j) of image elements (of objects and background) [2]:

$$C_{gen} = \int_{-1}^1 |C_{ij}| \cdot h(C_{ij}) dC_{ij}, \quad (1)$$

where C_{ij} – contrast of a pair (i, j) of image elements; $h(C_{ij})$ – probability density function for contrast C_{ij} .

However, the assessment of the distribution of $h(C_{ij})$ itself is quite a challenge. Therefore expression (1) is often represented in the form [2]:

$$C_{gen} = \int_0^1 \int_0^1 C_{ij} |p(L_i, L_j) dL_i dL_j, \quad (2)$$

where L_i, L_j - brightness of the image elements i and j ; $p(L_i, L_j)$ - two-dimensional distribution of brightness of image elements.

In [2] the definition of global contrast of multi-element image as the average value of the contrast of pairs of image elements relative to a preset adaptation level was proposed:

$$C'_{gen} = \int_0^1 \int_0^1 C_{ij0} |p(L_i, L_j) dL_i dL_j, \quad (3)$$

where C_{ij0} - contrast of two image elements at a preset adaptation level value L_0 .

In [6] the value L_0 of adaptation level is equal to the average value of brightness of the current image.

To the practical implementation of these approaches (2) and (3), it is necessary to solve the problems of estimating the two-dimensional distribution $p(L_i, L_j)$ of brightness and of choosing the definition for contrast of two image elements.

To simplify the calculations, in [2] the estimate of the two-dimensional distribution $p(L_i, L_j)$ of the brightness of the image elements (objects and background) has been suggested:

$$p(L_i, L_j) = p(L_i) p(L_j), \quad (4)$$

where $p(L_i)$ - probability density function of brightness.

In this case, expressions (1) and (2) on the basis of (4) take the form [2]:

$$C_{com} = \int_0^1 \int_0^1 C_{ij} |p(L_i)p(L_j) dL_i dL_j, \quad (5)$$

$$C'_{com} = \int_0^1 \int_0^1 C_{ij0} |p(L_i)p(L_j) dL_i dL_j, \quad (6)$$

where C_{com}, C'_{com} - definitions of complete integral contrast of image.

In [2] another approach to the estimation of the two-dimensional distribution of brightness of image elements has been suggested:

$$p(L_i, L_j) = p(L_i) \cdot \delta(L_i - L_j), \quad (7)$$

where $\delta(\cdot)$ - delta function.

In [2] for (3) on the basis (7) the definition of incomplete integral contrast of image was proposed:

$$C_{inc} = \int_0^1 C_{i0} |p(L_i) dL_i, \quad (8)$$

where C_{i0} - contrast of i -th element of image relative to adaptation level L_0 .

Expressions (5), (6) and (8) are no-reference histogram-based metrics of global contrast for multi-element images.

However, for the practical implementation of the examined approaches (5), (6) and (8) very important problem is also the choice of concrete definition of contrast for two image elements (two objects or an object and a background).

The choice of the definition of contrast for two image elements appreciably defines the accuracy of measurement of the global contrast for complex multi-element images.

It is assumed that the definition of contrast must satisfy the following basic requirements.

B. The basic requirements to the contrast definition

It is traditionally supposed that the contrast of two image elements is a dimensionless function and must satisfy the following basic requirements [7, 8]:

1) conditions for equality and asymmetry of the influence the arguments L_1 and L_2 [7]:

$$|C(L_1, L_2)| = |C(L_2, L_1)|, \quad (9)$$

$$C(L_1, L_2) = -C(L_2, L_1); \quad (10)$$

2) unambiguity and certainty of conditions under which the equality to zero is achieved [7]:

$$C(L_1, L_2) = 0 \text{ only when } L_1 = L_2; \quad (11)$$

3) condition of limitations of the range of contrast values [7]:

$$C(L_1, L_2) \leq C_{max}, \forall L_1, L_2 \in [0, 1], \quad (12)$$

where $|C_{max}|$ - maximum absolute value of contrast, it is usually assumed that:

$$C_{max} = 1, |C(L_1, L_2)| \in [0, 1]; \quad (13)$$

4) unambiguity and certainty of the conditions under which the maximum absolute value of the contrast is achieved [7]:

$$|C(L_1, L_2)| \rightarrow \begin{cases} = |C_{max}|, & \text{if } |L_1 - L_2| = L_{max} - L_{min} \\ < |C_{max}|, & \text{otherwise} \end{cases}, \quad (14)$$

where L_{min}, L_{max} - minimum and maximum values of brightness on image.

In [8] the requirements on the invariance of the definitions of contrast relative to linear transformations of the brightness scale were discussed:

$$C(k \cdot L_1 + b, k \cdot L_2 + b) = \text{sign}(k) \cdot C(L_1, L_2), \quad (15)$$

$$k \neq 0 \forall L_1, L_2, (k \cdot L_1 + b), (k \cdot L_2 + b) \in [0, 1].$$

The requirement (15) implies a number of important consequences:

a) condition of invariance to image inversion:

$$C(\bar{L}_1, \bar{L}_2) = -C(L_1, L_2), \quad (16)$$

$$\bar{L} = 1 - L; \quad (17)$$

b) condition of invariance to linear stretching of the dynamic range of image brightness:

$$C(\tilde{L}_1, \tilde{L}_2) = C(L_1, L_2), \quad (18)$$

$$\tilde{L}_i = \frac{L_i - L_{\min}}{L_{\max} - L_{\min}}. \quad (19)$$

Expressions (9)-(16), (18) define the basic requirements for contrast definition of image elements.

C. Known definitions of contrast of image elements

There are various approaches to definition the contrast of simple two-element images.

The contrast of two elements of a simple image is most often characterized by a difference in their brightness.

In [2] the definition of weighted contrast of two elements of complex image relative to a preset adaptation level has been suggested:

$$C^{wei_1}(L_1, L_2) = \frac{L_1 \cdot L_2 - L_0^2}{L_1 \cdot L_2 + L_0^2}. \quad (20)$$

In [6] the definition of weighted contrast of the image elements on the basis of the contrast law of light perception has been proposed:

$$C^{wei_2}(L_1, L_2) = \frac{L_1^2 - L_2^2}{L_1^2 + L_2^2}. \quad (21)$$

The most widely used at present definition of weighted contrast is defined as [3]:

$$C^{wei_3}(L_1, L_2) = \frac{L_1 - L_2}{L_1 + L_2}. \quad (22)$$

In [9], as an assessment of the perceived contrast, a weighted contrast (22) of the elements of the pre-inverted image (17) was proposed:

$$C^{wei_4}(L_1, L_2) = \frac{L_1 - L_2}{2 - L_1 - L_2}. \quad (23)$$

Another known definition of the contrast of image elements is the relative contrast, which is most often defined as [7, 11]:

$$C^{rel_1}(L_1, L_2) = \frac{L_1 - L_2}{\max(L_1, L_2)}, \quad (24)$$

$$C^{rel_2}(L_1, L_2) = \frac{L_1 - L_2}{1 - \min(L_1, L_2)}. \quad (25)$$

3D-graphs of surfaces for the weighted (20)-(23) and relative (24)-(25) contrast for primary image Lena (Fig.1) [12] are shown in Fig. 2 - Fig. 7.

However, known definitions (20)-(23) and (24)-(25) of weighted and relative contrast have significant disadvantages.

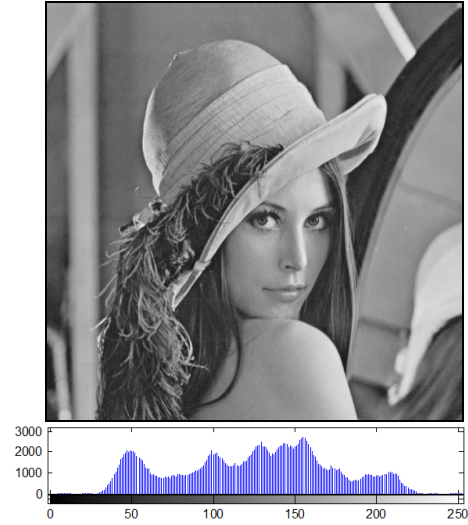


Fig. 1. Appearance of the primary image Lena [12] and its histogram ($L_{\min} = 0.0980$; $L_{\max} = 0.9608$; $L_0 = 0.4864$)

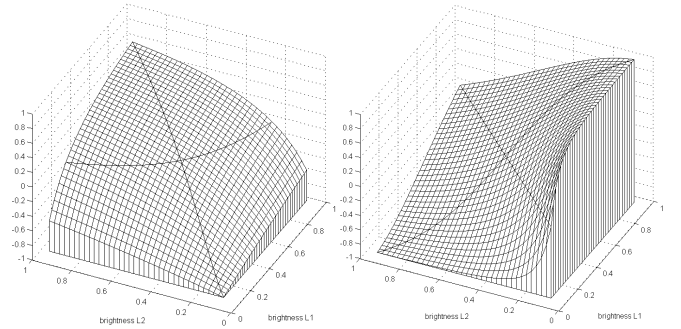


Fig. 2. $C^{wei_1}(L_1, L_2)$ (20)

Fig. 3. $C^{wei_2}(L_1, L_2)$ (21)

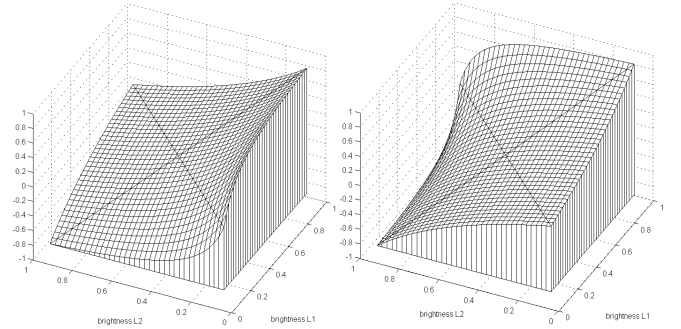


Fig. 4. $C^{wei_3}(L_1, L_2)$ (22)

Fig. 5. $C^{wei_4}(L_1, L_2)$ (23)

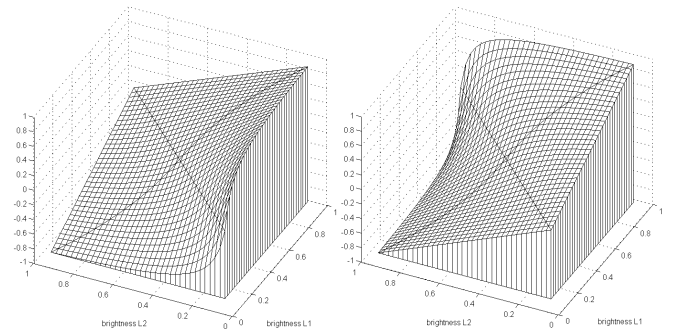


Fig. 6. $C^{rel_1}(L_1, L_2)$ (24)

Fig. 7. $C^{rel_2}(L_1, L_2)$ (25)

A significant disadvantage of the known definitions of weighted and relative contrast is their relatively low efficiency in measuring the perceived global contrast for complex multi-element images with limited dynamic range of brightness, since when measuring the contrast of image elements the characteristics (size and location) of the dynamic range of image brightness are not taken into account [9].

To address this disadvantage, we propose a new method for assessment the contrast of elements of complex image on the basis of measuring of the contrast of these elements on a pre-normalized image with subsequent correction of the contrast value taking into account the dynamic range of the initial (original) image.

III. THE PROPOSED METHOD

In this paper a new method for measuring the perceived contrast of image elements of complex multi-element images is proposed based on measuring the contrast of image elements on a pre-normalized image and estimating the dynamic range of the original image.

To define the contrast of two image elements on multi-element image for the chosen (specified) definition of contrast $C(L_1, L_2)$ (of contrast kernel), we propose a generalized description of the perceived contrast of image elements based on an analytical definition of the contrast of two image elements on a pre-normalized image and the value of maximum contrast for the original image:

$$\tilde{C}(L_1, L_2) = |C_{\max}| \cdot C(\tilde{L}_1, \tilde{L}_2), \quad (26)$$

where $C(\tilde{L}_1, \tilde{L}_2)$ - contrast of two image elements on a normalized image using (19); $|C_{\max}|$ - maximum absolute value of contrast for the original image, normalizing factor, $|C_{\max}| \leq 1$.

Taking into account that according to (14):

$$|C_{\max}| = C(L_{\max}, L_{\min}), \quad (27)$$

and considering that according to (19):

$$C(\tilde{L}_1, \tilde{L}_2) = C\left(\frac{L_1 - L_{\min}}{L_{\max} - L_{\min}}, \frac{L_2 - L_{\min}}{L_{\max} - L_{\min}}\right), \quad (28)$$

the expression (1) can be written in the form:

$$\tilde{C}(L_1, L_2) = C(L_{\max}, L_{\min}) \cdot C\left(\frac{L_1 - L_{\min}}{L_{\max} - L_{\min}}, \frac{L_2 - L_{\min}}{L_{\max} - L_{\min}}\right). \quad (29)$$

Expressions (26) and (29) describe the proposed method of assessing the perceived contrast of the elements of complex image for specified definition $C(L_1, L_2)$ of contrast kernel.

To demonstrate the possibilities of the proposed method (26), let us consider known definitions of weighted (22) and relative (23), (25) contrast, which are most often used to estimate the perceived contrast of image elements.

A. The proposed definition for weighted contrast

At present, the known definition (22) of weighted contrast is most often used to define the perceived contrast of image elements.

Expression (28) for the weighted contrast (22) takes the form:

$$C^{wei_3}(\tilde{L}_1, \tilde{L}_2) = \frac{L_1 - L_2}{L_1 + L_2 - 2L_{\min}}. \quad (30)$$

In this case, the maximum value of the weighted contrast (22) for the original image is equal to:

$$|C_{\max}^{wei_3}| = C^{wei_3}(L_{\max}, L_{\min}) = \frac{L_{\max} - L_{\min}}{L_{\max} + L_{\min}}. \quad (31)$$

Taking into account (26), (30) and (31), the proposed definition of the perceived contrast of the image elements on the basis of the weighted contrast kernel (22) takes the form:

$$\tilde{C}^{wei_3}(L_1, L_2) = \frac{L_{\max} - L_{\min}}{L_{\max} + L_{\min}} \cdot \frac{L_1 - L_2}{L_1 + L_2 - 2L_{\min}}. \quad (32)$$

By analogy with (30) - (32), the proposed definition of perceived contrast using the definition (23) of weighted contrast has the form:

$$C^{wei_4}(\tilde{L}_1, \tilde{L}_2) = \frac{L_1 - L_2}{2L_{\max} - L_1 - L_2}, \quad (33)$$

$$|C_{\max}^{wei_4}| = C^{wei_4}(L_{\max}, L_{\min}) = \frac{L_{\max} - L_{\min}}{2 - L_{\max} - L_{\min}}, \quad (34)$$

$$\tilde{C}^{wei_4}(L_1, L_2) = \frac{L_{\max} - L_{\min}}{2 - L_{\max} - L_{\min}} \cdot \frac{L_1 - L_2}{2L_{\max} - L_1 - L_2}. \quad (35)$$

Expressions (26) and (29) describe the proposed method for assessing the perceived contrast of elements of complex image for known definitions (22) and (23) of weighted contrast.

B. The proposed definition for relative contrast

Expressions (27)-(29) for the definition (24) of relative contrast take the form:

$$C^{rel_1}(\tilde{L}_1, \tilde{L}_2) = \frac{L_1 - L_2}{\max(L_1, L_2) - L_{\min}}, \quad (36)$$

$$C^{rel_1}(L_{\max}, L_{\min}) = \frac{L_{\max} - L_{\min}}{L_{\max}}, \quad (37)$$

$$\tilde{C}^{rel_1}(L_1, L_2) = \frac{L_{\max} - L_{\min}}{L_{\max}} \cdot \frac{L_1 - L_2}{\max(L_1, L_2) - L_{\min}}. \quad (38)$$

The proposed definition of perceived contrast using the definition (25) of relative contrast has the form:

$$C^{rel_2}(\tilde{L}_1, \tilde{L}_2) = \frac{L_1 - L_2}{L_{\max} - \min(L_1, L_2)}, \quad (39)$$

$$C^{rel_2}(L_{\max}, L_{\min}) = \frac{L_{\max} - L_{\min}}{1 - L_{\min}}, \quad (40)$$

$$\tilde{C}^{rel_2}(L_1, L_2) = \frac{L_{\max} - L_{\min}}{1 - L_{\min}} \cdot \frac{L_1 - L_2}{L_{\max} - \min(L_1, L_2)}. \quad (41)$$

Proposed assessments (32), (35), (38), (41) of perceived contrast on the basis of the known definitions (22) - (25) of contrast are the basis for the no-reference metrics of global contrast for multi-element images (Section II.A).

3D-graphs of surfaces for the proposed assessments (32), (35), (38), (41) of weighted and relative contrast for primary image Lena (Fig.1) [12] are shown in Fig. 8 - Fig. 11.

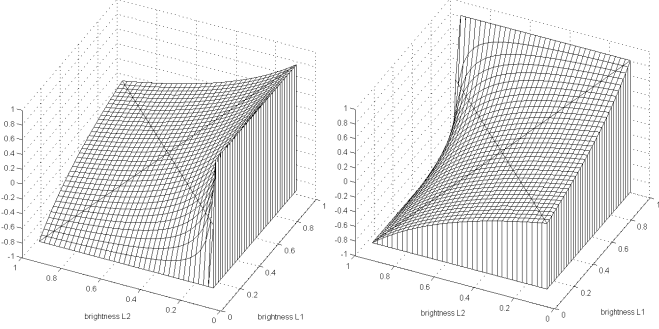


Fig. 8. $\tilde{C}^{wei_3}(L_1, L_2)$ (32)

Fig. 9. $\tilde{C}^{wei_4}(L_1, L_2)$ (35)

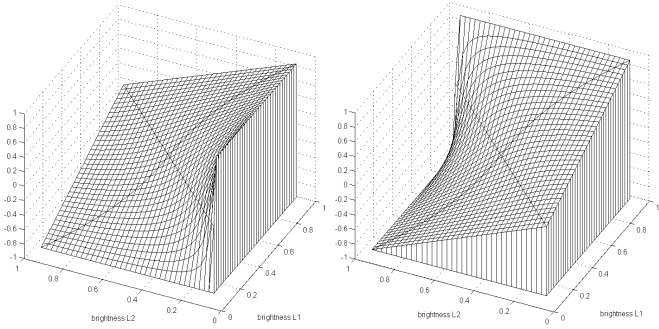


Fig. 10. $\tilde{C}^{rel_1}(L_1, L_2)$ (38)

Fig. 11. $\tilde{C}^{rel_2}(L_1, L_2)$ (41)

Comparative analysis of known and proposed definitions of contrast of image elements was carried out in Section IV and Section V.

IV. RESEARCH

Experimental research was carried out by measuring of global contrast using known and proposed definitions of contrast of image elements for a group of nine test images.

The group of test images consists of nine real images with a complex structure and a limited dynamic range of brightness, the appearance of which is shown in Fig. 12.

Research was carried out by measuring of complete and incomplete integral contrast using known and proposed definitions of contrast of image elements, namely:

- 1) complete contrast (6) using weighted contrast (20);
- 2) complete contrast (5) using weighted contrast (22);
- 3) incomplete contrast (8) using weighted contrast (21);
- 4) known definition of incomplete contrast (8) using linear contrast [7]:

$$C_{inc}^{lin} = \int_0^1 \left| \frac{L - L_0}{LMAX} + \frac{1}{2} - \left| \frac{L - L_0}{LMAX} - \frac{1}{2} \right| \right| p(L) dL, \quad (42)$$

where $LMAX$ - maximum possible value of brightness;

- 5) complete integral contrast (6) using proposed definition (32) of the perceived contrast for weighted contrast;
- 6) incomplete integral contrast (8) using proposed definition (32) of the perceived contrast for weighted contrast.

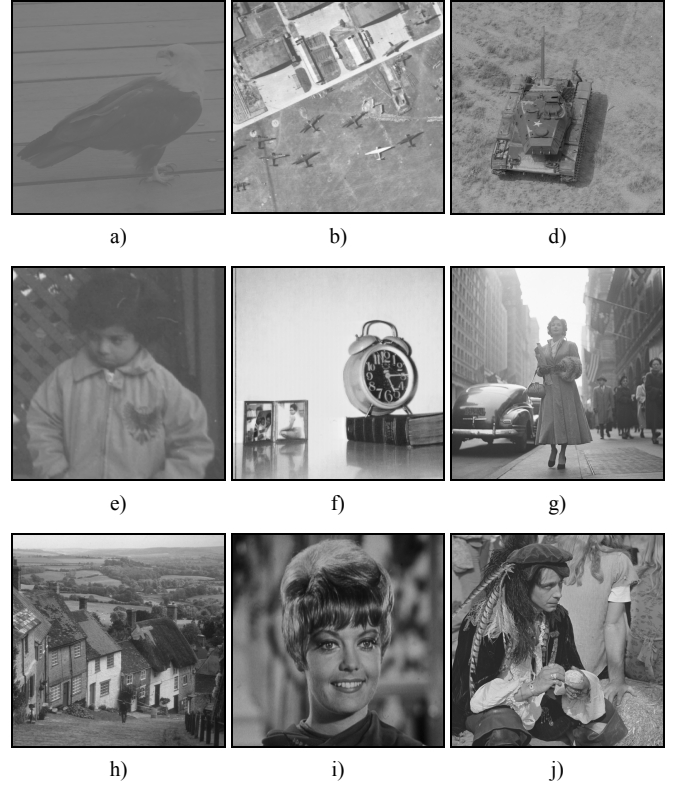


Fig. 12. The appearance of test images

The results of measuring of global contrast for test images (Fig. 12.a – Fig. 12.j) are shown in Table I.

TABLE I. RESULTS OF CONTRAST MEASUREMENT FOR TEST IMAGES

	test images								
	a	b	d	e	f	g	h	i	j
$C_{com}^{wei_1}$	0.050	0.104	0.131	0.115	0.223	0.224	0.253	0.271	0.278
$C_{com}^{wei_3}$	0.044	0.101	0.119	0.113	0.190	0.218	0.246	0.260	0.262
$C_{inc}^{wei_2}$	0.068	0.134	0.162	0.175	0.264	0.298	0.325	0.347	0.363
C_{inc}^{lin}	0.062	0.174	0.162	0.153	0.372	0.339	0.304	0.256	0.323
$\tilde{C}_{com}^{wei_3}$	0.048	0.109	0.120	0.190	0.189	0.236	0.251	0.260	0.321
$\tilde{C}_{inc}^{wei_3}$	0.036	0.074	0.085	0.153	0.143	0.172	0.182	0.193	0.241

Graphs of values of global contrast for test images (Fig. 12.a - Fig. 12.j) are shown in Fig. 13.

Analysis of results of the research is carried out in Section V and Section VI.

V. DISCUSSION

The results of measurements of contrast for test images show that the value of the integral contrast significantly depends on the choice of the definition of contrast kernel

(Fig. 13). The values of the known definitions (20) - (25) of the weighted and relative contrast heavily depend on the changes in the average value of the brightness under the additive transformations of the brightness scale of image.

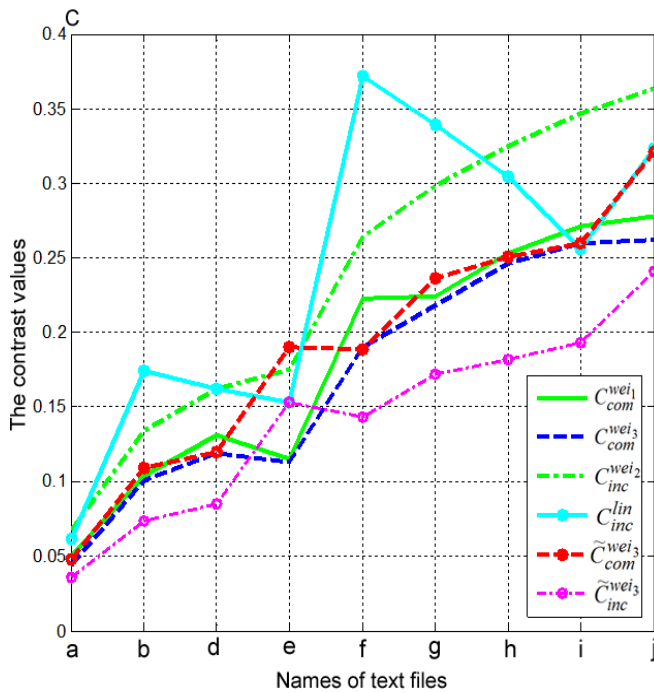


Fig. 13. Contrast values for test images (Fig. 12).

A proposed generalized description (26), (28) of the perceived contrast of elements of complex images is invariant to linear transformations of image brightness scale.

Values of known (22), (23), (24), (25) and proposed (32), (35), (38), (41) definitions for weighted and relative contrast coincide for pre-normalized images, for which $L_{min}=0$ and $L_{max}=1$.

The values $\tilde{C}_{com}^{wei_3}$ of the complete integral contrast (5) on the basis of the proposed definition (32) of weighted contrast are proportional to the values of the incomplete integral contrast $\tilde{C}_{inc}^{wei_3}$ (8) using weighted contrast (32).

The assessments (5), (8) on the basis of integral contrast of image using proposed definitions of weighted contrast (32) are the closest to the expert estimates of image contrast and are best suited to quantitative assessment of global contrast of images with complex structure and limited dynamic range of brightness.

VI. CONCLUSION

The problem of increasing the accuracy of measuring the perceived contrast of elements (objects and background) of a complex multi-element image was considered.

A new method of measuring the perceived contrast of two image elements (objects and background) by measuring

their contrast on a normalized image with subsequent correction of the contrast value taking into account the dynamic range of the initial image was proposed.

A new generalized description of the perceived contrast of elements of complex images for various definitions of the contrast kernel was proposed.

The proposed generalized description of the perceived contrast satisfies the basic requirements for the definition of contrast and provides a sufficiently accurate quantitative assessment of the contrast of image elements for complex images, also allow predict the perceived contrast of the image when subjective expert assessments.

New definitions of the weighted and relative contrast of the image elements were proposed.

The proposed definitions of weighted and relative contrast allow to increase the accuracy and reliability of measuring the global contrast for multi-element monochrome images with limited dynamic range of brightness.

The proposed definitions of weighted and relative contrast can be recommended to estimate the generalized contrast of images in imaging, image processing and analysis in automatic mode.

REFERENCES

- [1] Z. Wang, and A.C. Bovik, Modern Image Quality Assessment. Morgan and Claypool Publishers, New York, 2006.
- [2] V. F. Nesteruk, and V. A. Sokolova, "Questions of the theory of perception of subject images and a quantitative assessment of their contrast", Optiko-electronic industry, no. 5, pp. 11-131, 1980.
- [3] W. K. Pratt, Digital Image Processing: PIKS Inside, 3rd edn, John Wiley & Sons, New York, 2001.
- [4] R.C. Gonzalez, and R.E. Woods, Digital Image Processing. 2nd edn, Prentice Hall, New Jersey, 2002.
- [5] Z. Wang, H.R. Sheikh, and A.C. Bovik, "Objective video quality assessment", in Furht, B. and Marqure, O. (ed.), The Handbook of Video Databases: Design and Applications, CRC Press, Austin, USA, vol. 41, pp. 1041-1078, 2003.
- [6] V.F. Nesteruk, and N.N. Porfiryeva, "Contrast law of light perception", Optics and spectroscopy, vol. XXIX, no. 6, pp. 1138 – 1143, 1970.
- [7] R.A. Vorobel, Loharyfmichna obrobka zobrazhen' [Logarithmic Image Processing], Naukova Dumka, Kyiv, Ukraine, 2012.
- [8] E. Yelmanova, "Quantitative Assessment of Contrast of the Image Elements", Herald of Lviv Polytechnic National University, Series of Radio Electronics and Telecommunication, vol. 818, pp. 69-75, 2015.
- [9] E. Yelmanova, and Y. Romanyshyn, "Definitions of Weighted and Relative Contrasts of Elements of Monochrome Images", XII-th International Scientific and Technical Conference "Computer Science and Information Technologies" (CSIT'2017), Lviv, Ukraine, pp. 381-384, September 05 – 08, 2017.
- [10] E. Peli, "Contrast in Complex Images" in Journal of the Optical Society of America A (JOSA A), vol. 7, no. 10, pp. 2032-2040, 1990.
- [11] P. Whittle, "Increments and decrements: luminance discrimination", Vision Research, vol. 26, iss. 10, pp. 1677–1691, 1986
- [1] Public-Domain Test Images for Homework's and Projects, <http://homepages.cae.wisc.edu/~/images/Lena.tif>.

Method for Determining the Rational Time Intervals for Detecting Objects by Thermal Imager

Maksym Korobchynskyi

Military-Diplomatic Academy named after Eugene Bereznyak
Kiev, Ukraine
maks_kor@ukr.net

Alexander Mariliv

Military-Diplomatic Academy named after Eugene Bereznyak
Kiev, Ukraine
spirit148@i.ua

Mihail Slonov

Military-Diplomatic Academy named after Eugene Bereznyak
Kiev, Ukraine
slonovmu@gmail.com

Serhii Mieshkov

Military-Diplomatic Academy named after Eugene Bereznyak
Kiev, Ukraine
sergeymechkov081971@gmail.com

Abstract—The developed method allows to provide the maximum possible distance of the thermal imaging observation by determining the rational time intervals of a day when the thermal contrast between object and background is maximal. The method takes into account the influence of cyclic heat transfer in the environment on the change in the recognition properties of the thermal image during the day. This method allows determining the time intervals of the day when the quality of the thermal image will be maximal or minimal.

Keywords—object, environment, thermal image, time of day

I. INTRODUCTION

The level of thermal radiation of object Φ is important, when operator is using thermal imagers. Thermal radiation changes under the influence of heat exchange processes between the object and the environment. Also it depends on the thermo-physical properties of the materials from which the object was produced. Therefore, each object will be detected on a thermal image with varying degrees of probability P . During the day, the level of thermal radiation $\Phi(t)$ have cyclical changes. Determining the time of day, when the level of thermal radiation will be maximum allow to detect objects from a higher distance. But, at the present time, there are no methods to determining time intervals, when using of thermal imager is the most rational.

II. THE FORMULATION OF THE PROBLEM

Today there isn't information about availability of techniques for the use of thermal imagers, which take into account the cyclic changes of the recognition properties of the thermal image during the day. The scientific and methodical apparatus for calculating the surface temperature of the objects with cyclic effects of heat exchange processes of the environment is considered in [1-3]. But the problem of detecting objects using thermal imagers is not considered in [1-3]. In [4-6] the focus is on developing technologies (techniques) recognition of multispectral images (including thermal images), discussed issues related to direct detection (shape, contrast, shadow) tell-tale signs of objects. But the influence of the change of environment is not considered.

The cyclic effects of heat exchange processes in the environment on the nature of the distribution of the temperature of the object and background is described in [7-9]. However, in these sources, there are no work on ways to

determine rational intervals of time, when the use of thermal imagers is the most efficient. Information obtained by analysis of the thermal image, determined by its quality. The lower the quality gives, less probability to detect the object. Improving image quality is described in the works [4-5, 7]. Authors emphasized that the quality of the image is derived from the hardware perfection of the thermal imager and the conditions for thermal imaging observation.

One of the factors of the conditions is the change of the object and background temperature during the day. Features of thermal imaging observation of objects during the day were investigated in [9-10]. But they were limited just by carry out of experiments. These experiments showed a significant impact on the quality of the thermal image by changing of recognition properties of objects. So, we have the cyclic changes in the recognition properties of the thermal image $P = \{d, x\}$ but don't have prediction of intervals of the day Δt_r , when such properties are maximal $P \rightarrow \max$. Solving this problem will allow to obtain a better thermal image, and as a result, increase the maximum possible distance of the object detection $L \rightarrow \max$.

III. METHOD OF SOLVING THE PROBLEM

Equation (1) allows you to take into account the main factors that affect at the probability of object detection:

$$P(t) = \exp \left[- \frac{\ln P_g}{\lg \frac{1+K(t)}{1-K(t)}} \left(\frac{d}{x} \right)^2 \right], \quad (1)$$

where $P(t)$ is probability of object detection; t is time of day, hour; P_g is the given value of the probability of object detection; d is spatial disparity in the plane of the object, m; x is linear size of recognizable feature of the object, m; $K(t)$ is thermal contrast [6].

By equation (1), time-varying parameter is the value of thermal contrast $K(t)$. In accordance with the law of Stefan-Boltzmann, thermal contrast describes the ratio of the temperature difference of the object $T_{ob}(t)$ and background $T_f(t)$ to their sum. Equation of the thermal contrast has the form:

$$K(t) = \frac{\varepsilon_{ob} T_{ob}^4(t) - \varepsilon_f T_f^4(t)}{\varepsilon_{ob} T_{ob}^4(t) + \varepsilon_f T_f^4(t)}, \quad (2)$$

where ε_{ob} , ε_f is radiation ratio of the object and background; $T_{ob}(t)$, $T_f(t)$ is temperature of the object and background, K [11].

Changing of the object and the background temperature during the day and the effect of heat exchange processes can be represented by equation:

$$\begin{cases} \Phi(t) = \sum_{k=1}^n \Phi_{c,k}(t) + \sum_{r=1}^h \Phi_{rad,r}(t) + \sum_{i=1}^b \Phi_{cd,i}(t); \\ \frac{dT(t)}{dt} = a \nabla^2 T(t) + \frac{P_v(t)}{c\rho}, \end{cases} \quad (3)$$

where $\Phi_c(t)$, $\Phi_{rad}(t)$ and $\Phi_{cd}(t)$ is heat flux, which comes to the object by convective, radiative and conductive heat exchange, W ; n , h and b is discretization of the object by sections what determine convective, radiative and conductive heat exchange; a is coefficient of temperature conductivity, m^2/s ; ρ is density, kg/m^3 ; c is specific heat capacity, $J/(kg \cdot K)$; ∇^2 is the Laplace operator, $1/m$; $P_v(t)$ is heat stress, W/m^3 [8].

Solution of equation (3) allows determining the effect of heat exchange processes in the environment by changing the temperature of the object and background during the day. Versatility of equation (3) is possibility to simultaneously take into account all types of heat exchange processes between the environment, object and background.

Determine through equation (1) the distance of thermal imaging observation L . Imagine spatial disparity in the plane of the object d as equation:

$$d = \frac{LD}{f}, \quad (4)$$

where D is matrix pixel size of the thermal imager, mm ; L is distance of thermal imaging observation, m ; f is focal length of the thermal imager, m [12];

Substituted at equation (1) equation (4), then:

$$P(t) = \exp \left[- \frac{\ln P_g}{\lg \frac{1+K(t)}{1-K(t)}} \left(\frac{LD}{xf} \right)^2 \right]. \quad (5)$$

From equation (5) we can determine how will change distance of thermal imaging observation during the day:

$$L(t) = \frac{xf}{nD} \sqrt{\log_{P_g} P \cdot \lg \frac{1+K(t)}{1-K(t)}}, \quad (6)$$

where n is number of pixels what comes to the distinguishing feature of the object.

Based on equations (1) – (6) developed method for determining the rational time intervals for detecting objects by thermal imager.

The implementation of the method involves such assumptions: prognostication of the rational time intervals is carried out on the eve of thermal imaging observation, because of high accuracy of weather forecast; distance of thermal imaging observation is determined from the terms of the task. Method for determining the rational time intervals for detecting objects by thermal imager consists of six steps: first – forecast change of $T_{ob}(t)$ and $T_f(t)$ during the day (equation (3)); second – forecast change of $K(t)$ during the day (equation (2)); third – determination of the number of pixels n , which is necessary for detection of recognizable feature of the object; fourth – forecast change of $P(t)$ during the day (equation (1) or (5)); fifth – plotting $P(t)$; sixth – definition of conditions for which time intervals considered rational Δt_r , and finding time indicators of start and end each of the specified intervals (Fig. 1).

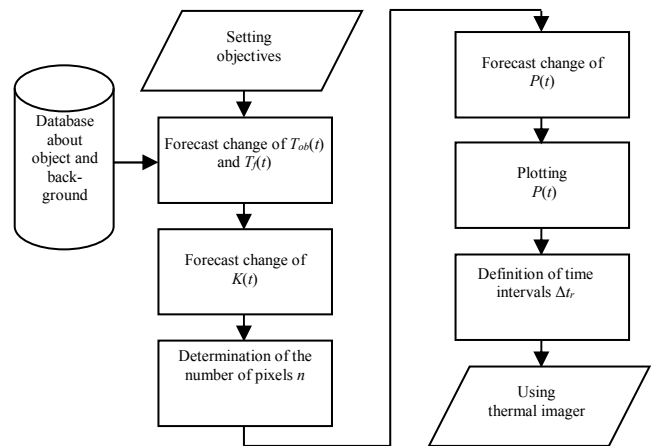


Fig. 1. Scheme of the method

In the first step prognostication $T_{ob}(t)$ i $T_f(t)$ is carried out by means of a solution equations (3), where the methods of the theory of heat and mass transfer are based. Data about thermophysical and mass-dimensional parameters of objects and backgrounds are determined from the reference literature. In the second step by using equations (2) finds changes in thermal contrast $K(t)$ during the day. The value of the radiation factor of the object and the background for different materials is individual and determined from the reference literature.

In the third step it is necessary to determine with what spatial disparity have to find the object. According to Johnson, to detect an object the number of pixels is $n = 2...3$, recognition – $n = 6...9$, classification – $n = 12...15$ and object identification – $n > 18$.

In the fourth step by using equations (1) or (5) are forecasting of changes $P(t)$ during the day. Equations (1) parameter d determines the scale of the thermal image and depends of distance of thermal imaging observation.

Parameter x determines from the description of the overall dimensions of the object. In equations (5) parameter D and f determines from the data on the tactical and technical characteristics of the thermal imager. Difference between equation (1) and (5) is that in equation (1) you need to have information about the magnitude of the thermal image but in equation (5) you need to have information about tactical and technical characteristics of the thermal imager.

In the fifth step by having received information about changing the probability of object detection $P(t)$ during the day we construct graphic addition.

In the sixth step by using the given value of the probability of object detection P_g we consider intervals what are rational, when $\Delta t_r \geq P_g$, or irrational $\Delta t_n < P_g$. Then for rational and irrational time intervals, are the numerical values of the beginning and end of each interval. To find distance of thermal imaging observation L in rational time intervals Δt_r if given value of the probability of object detection P_g is used an equation (6).

IV. THE IMPLAMENTATION OF THE METOD

To check the adequacy of the developed method experimental studies were conducted for a specific “object-background” combination. The results of experimental studies were determined the value of the temperature of the object and the background, thermal contrasts, probability of object detection and rational time intervals. Also, during the experimental studies were determined thermal images to specific “object-background” combination with a frequency of 1 hour throughout the day.

Examples of thermal imaging images obtained during experimental studies presented on Fig. 2 and confirming the impact cyclic heat exchange in the environment to change the recognizable properties thermal image during the day.

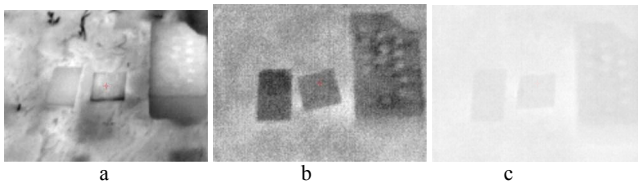


Fig. 2. Thermal images of objects at different times of day (a – 13:00, b – 20:30, c – 5:30)

During the experiment, a change in the values of the temperature of the brick and the soil was studied. The obtained results are presented as graphical dependence on Fig. 3. The research was conducted by using a thermal imager Fluke Ti30 from 13 to 14 august 2017, in the open air without clouds.

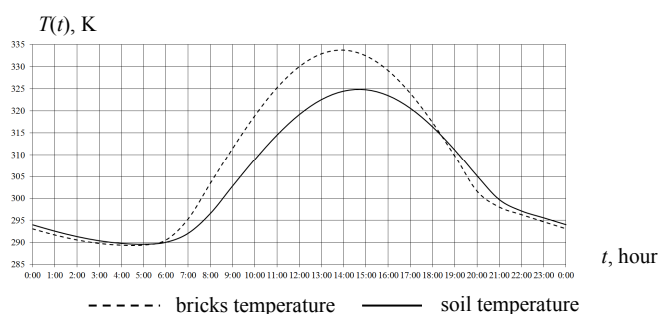


Fig. 3. Change in temperature of object and background during the day

Next, according to equation (1) algorithmically determined value of probability of object detection. Data from Table. 1 were used for calculations. Results of change values probability $P(t)$ are presented on Fig. 4.

TABLE I. DATA FOR THE CALCULATION OF TEMPERATURE CHANGE OF THE OBJECT AND THE BACKGROUND DURING THE DAY

Parameters	Object	Background
ε	0,7	0,7
P_g	0,8	0,8
n	8	8
d, m	0,015	0,015
x, m	0,12	0,12

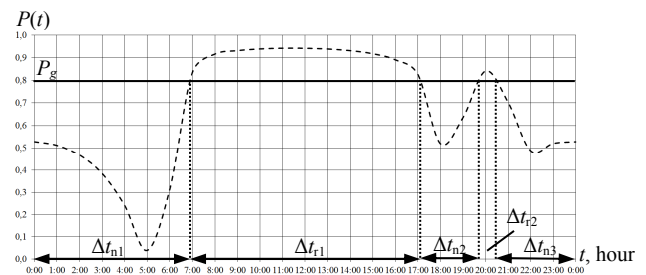


Fig. 4. Change the probability of detecting an object during the day

According to Fig. 4 rational time intervals Δt_r will be: $\Delta t_{r1} \in [6:55 - 17:10]$ and $\Delta t_{r2} \in [19:40 - 20:30]$. Irrational time intervals Δt_n are: $\Delta t_{n1} \in [0:00 - 6:55]$; $\Delta t_{n2} \in [17:10 - 19:40]$ and $\Delta t_{n3} \in [20:30 - 0:00]$. Thus, during the day is observed 2 rational and 3 irrational time intervals for using thermal imager.

To definition of numerical values of efficiency of use a thermal imager in rational time intervals by equation (6), we should calculated value of distances. This distance guarantees the detection of an object with given value of the probability P_g in rational Δt_r and Δt_n irrational. Next is determined by how many times distances guarantees the detection of an object in rational time intervals L_r bigger then in irrational time intervals L_n . Under distance, the guaranteed detection of an object is understood as a distance, which fulfills the condition $P(t) \geq P_g$. Data for calculations was used from Table. 2. Results of calculations distance of thermal imaging observation for bricks on the background of the soil during the day for thermal imager Fluke Ti30 presented on Fig. 5.

TABLE II. DATA FOR CALCULATION OF CHANGE DISTANCE OF DETECTION OF THE OBJECT DURING THE DAY

Parameters	Value
P	0,8
D, M	0,000025
f, M	0,025

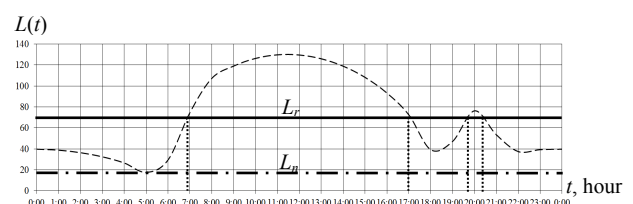


Fig. 5. Figure 5. Change the detection distance of the object during the day

According to Fig. 5 $L_r = 74,5$ m and $L_n = 17,7$ m. Thus, the use of the method allows guaranteed to detect an object in rational time intervals from distance, that in 4,2 times higher than the same value at irrational time intervals.

V. CONCLUSION

The results of the scientific research carried out are important for the field observation of objects in conditions of low contrast of the images. Immediately, the research results can be used when planning the time of day for using thermal imager.

New method for determining the rational time intervals for detecting objects by thermal imager based on the consideration of cyclic heat transfer in the environment. This will increase the detection range of objects by thermal imagers due to their use at rational intervals of the day for a given value of thermal contrast or the likelihood of detecting an object. The recognition properties of the thermal image will be maximally possible at rational intervals of the day.

The results of experimental studies to determine the change in the distance of detection of an object (brick) on the background of the ground during the day. According to the results of experimental studies, it can be argued that using of new method allows to reveal rational time intervals when guaranteed to detect of object carried out with the

given value of the probability for a specific type of thermal imager, from distance, that in 4,2 times higher than the same value at irrational time intervals.

REFERENCES

- [1] A. Lukov, *The theory of thermal conductivity*, Moscow, Vischaya Shkola, 1967.
- [2] V. Lykanin, *Heat engineering*, Moscow, Vischaya Shkola, 2000.
- [3] A. Dumnich, *Thermal conductivity*, Donetsk, DNU, 2003.
- [4] J. Lloyd, "Systems of thermal insulation", Moscow, Mir, pp. 13–231, 1978.
- [5] G. Gossorg, "Infrared thermography. Fundamentals, technique, application", Moscow, Mir, pp. 21–51, 1988.
- [6] A. Givichin, and V. Sokolov, *Deciphering of images*, Moscow, Nedra, 1980.
- [7] Yu. Rebyn, "Infrared icons systems", Kiev, KVVAIU, pp. 3–18, 1985.
- [8] M. Dmitriev, O. Panchenko, O. Derkachov, and I. Rutcovska, "Definition of regional condition at the airfield surface" Kiev, VNAU. no. 1, pp. 161–164, 2008.
- [9] V. Kaplan, "Experimental studies of the radiation contrasts of military personnel in the middle and far infrared ranges of the electromagnetic spectrum", Kazan, Applied physics, no. 5, pp. 108–113, 2006.
- [10] M. V. Korobchinskiy, M. U. Slonov, and A. A. Mariliv, "Mathematic model of forecasting daily cycle temperature changes of agricultural lands", Kiev, Artificial Intelligence, no. 72, pp. 124–132, 2016.
- [11] H. Gunzler, *IR Spectroscopy*, Weinheim, Germany, Wiley-VCH, 2002.
- [12] M. J. Riedl, *Optical Design Fundamentals for Infrared Systems*, Washington, USA, SPIE Press, 2001.

Bioinspired Approaches to the Selection and Processing of Video Information

Vitaliy Boyun

Department of Intelligent Real Time Video Systems

V.M.Glushkov Institute of Cybernetics NASU

Kyiv, Ukraine

vboyun@gmail.com

Abstract— Significant expansion of the range of applications of real-time video systems requires further improvement in their productivity, efficiency and intelligence. Therefore, researchers are increasingly turning to the human eye analyzer as a prototype to create more sophisticated systems of technical vision. The paper proposes a number of approaches and methods for the selection and processing of video information inspired by the human visual analyzer. In particular: the method of hierarchical selective perception of video information; dynamic models of processes for finding objects, tracking them, panning the scene and the mechanisms of attention and allocation of the essence, which, by managing the parameters of reading information from a video sensor, provide reading of a part of the image relevant to the task; information measure of the dynamic image (δ -entropy), which characterizes its spatial frequencies and is an effective information feature for the search and recognition of objects; methods of expanding the dynamic range of perception of brightness; the principles of circular organization neurons of the central fovea, which provide increased contrast and the allocation of informative features; circular organization of the retinal neurons with a summation of signal sticks that contribute to increased sensitivity in conditions of insufficient lighting; specialization of neurons and organization multilayer neural network.

Keywords— video system, human visual analyzer, bio-inspired approaches, relevant information, neural network

I. INTRODUCTION

The tasks that arise in transport, industry, robotics, medical-biological research, defense-military sphere, etc., require further increase of productivity, efficiency and level of intelligence of computer video systems. Therefore, researchers are increasingly turning to the visual system of man, as the most perfect prototype for the construction of computer vision systems.

The human eye system has been improving for millions of years and has reached an extremely high level of organization. A generalized model of the human visual system is multifunctional and consists of several dozen or even hundreds of local models that describe a range of structural, physical, geometric and psychophysical mechanisms and processes. The process of perceiving visual information by a person is dynamic, with many parameters that change in the process of perception, with many feedback bonds. We not only see, but we react, that is, such process is active. Therefore, the phenomenon of vision provides a lot of versatile elegant solutions for computer vision systems.

Given the perfection of the human visual analyzer, it is advisable to study and distinguish its elements for use in modern technical systems. It is not necessary to exactly copy them, but, conversely, understanding their functioning, implement them taking into account the statement of a specific technical problem and the capabilities of the level of technology.

II. ORGANIZATION OF THE HUMAN VISUAL SYSTEM

Approaches to the analysis of perception and processing of information are based on ascending and descending processes. Ascending processes, or processes for the transfer of information, begin with simple (basic) elements - discrete sensory, derived from sensory receptors. These basic features include: the difference in the luminosity of fragments, spatial frequencies, or the position of elements in space. Incoming touch information is transmitted from the base (lower) level to higher and integrative levels. In this case, the visual system carries out the design and creation of identifiable patterns by combining the basic elements under the influence of the reflex mechanisms of the visual system and the brain, not controlled by human.

Descending processes, or processes for the conceptualization of information data, use global, abstract, and higher levels of analysis for implementing processes at lower levels. These processes of perception of the form are based on the knowledge previously obtained by the observer, his previous experience, comprehension and interpretation, as well as his expectations.

Both descending and ascending processes are a form of manifestation of the activity of the visual system and the brain, and in most cases they are performed jointly, complementing each other.

Sticks and retinal cones perceive the image of the surrounding world as a set of individual points, although this world consists of separate, differentiated objects and surfaces that have a definite shape and outlines. For such a generalized perception of the world, the efforts of the multi-level neuronal network of the retina and higher levels of the visual system are being applied. Despite the huge streams of video information, the human visual-analyzing system copes with them due to its extremely high selectivity, which ensures the selection of only relevant information for specific conditions with the help of a multi-level neural network, which from level to level increases the abstract presentation of information.

Much of the pre-processing of visual information is already at the retina level. The most important role in our perception of form, edges and boundaries of the regions is the contours that appear when the adjacent surfaces are illuminated differently (that is, taking into account the intensity and color). The contrast of the surfaces, acting on the visual receptors, causes their interaction and creates conditions for perception of contours by the visual system. Ganglion cells of the retina with their lateral bonds and mechanisms of excitation and lateral inhibition are of great importance in the perception of contours.

Changes in the brightness (color) of the image on the retina sticks cause the excitation of the neurons (the mechanism of attention), which controls the rapid movements of the eye (saccades). Excited areas of the image are consistently (with priority) reflexively transmitted to the central fovea for detailed consideration and allocation of local informative features (ascending processes of control). The central fovea and periphery of the retina are organized according to the ring principle. However, in the central fossa, each cone has a gateway to the ganglion cell, and on the peripheral retina the sticks are grouped together, summing up signals from larger regions of the receptor field and providing increased sensitivity in low light (in exchange for a decrease in spatial resolution).

Besides to the roughly-accurate perception of visual information on space (i.e., in X, Y coordinates), the human eye reacts not to the amount of luminance or chromaticity in the image, but to changes between the luminance values of neighboring receptors, or the luminance values of a given receptor in time, that is, on dynamics of this parameter.

An arbitrary part field of view with contrasting light and dark areas can be analyzed and transformed into its spatial frequency - the number of variations of luminosity in a certain area of space, or the number of cycles of alternation of dark and light bands in a given field of view. It is proved that in the visual system there are detectors of spatial frequency - specialized cells, are most sensitive to certain spatial frequencies.

Thus, spatial frequencies are a simple and reliable way to describe and generalize the structural details of various visual objects.

The human eye system provides an extremely high range of light perception (10^{10}), which is ensured by the logarithmic perception of brightness, high sensitivity of the sticks and the lower sensitivity of the cones, as well as a number of organizational principles of the neural network of the retina. In particular, this is facilitated by: summing signals from the sticks of the peripheral retina to increase sensitivity in conditions of insufficient illumination, circular organization of neurons of peripheral retina and of central fovea, and the like.

From the ganglion cells of the central fossa and the peripheral retina, a whole series of dynamic images is sent to the corpus geniculatum laterale, each of which displays only one aspect of the overall visual picture. Each video stream is transmitted over its group of optic nerve fibers. In particular, streams of local features (orientational, colors, movement, etc.) are transmitted to the corpus geniculatum laterale.

The organization of the corpus geniculatum laterale is similar to the ring organization of the retina, but it covers and analyzes large portions of the receptor field.

From the corpus geniculatum laterale, signs of a higher level enter the visual cortex, which has a line organization (lines, bands, rectangles are allocated, their length, width, orientation is determined), that is, signs of a higher level of abstraction are formed - signs of the essence of the image. These signs in the human brain are compared with models that were acquired from human experience. They control the conscious movements of the eyes (downward control processes), determined by the cognitive process of perception of information. In this case, the brain, which has a complete retina model and acquired experience in perceiving images of objects, receiving information from the retina and comparing it with the models of objects, in accordance with the goal and characteristics of objects determines high-frequency automatic modes of controlling eye movements and directs low-frequency regimes [1-4].

Thus, the use of functions, principles of construction and adaptive mechanisms of the human visual analyzer will contribute to the creation of computer vision systems of the new generation.

III. USE OF ELEMENTS OF THE VISUAL ANALYZER IN THE SYSTEM OF COMPUTER VISION

The most important features of the human visual analyzer for computer vision systems are high selectivity of perception of video information and wide parallelization of information processing on layers of neurons of the retina and higher levels of the brain.

Taking into account the knowledge of high-level control of the movements of the eye of the visual system, the Institute of Cybernetics of the National Academy of Sciences of Ukraine developed dynamic models of processes for finding objects in the image, tracking them, panning, etc., which contribute to the selection of relevant information, greatly reducing the redundancy of its presentation. In particular, unlike pyramidal perception Burt [5], the method of *hierarchical selective perception* [6-8] is proposed, which, by analogy with the human visual analyzer, is based on reading a scene with a low resolution (an analogue of the peripheral retina) for a quick search of object of the given features (the mechanism of attention) and the next reading (analog of the saccades) with high resolution (analogue of the central fovea) of the part of the image with the object for further consideration, measurement, recognition. This can significantly reduce the amount of information processed, increase the efficiency and effectiveness of systems of technical vision.

Dynamic models of object tracking processes are based on the sequential reading of the location of an object, taking into account the direction and speed of its movement, and possibly also changes in overall dimensions (analogue - follow-up eye movements). Dynamic models of panning processes (analogous to processes of eye movement or head rotation) are based on reading an additional image that appears due to these movements, and "pasting" it with the previous image.

To determine the amount of information in the video sequence, a potential estimate based on the amplitude-spatial and temporal resolution is usually used

$$C_{s.n.} = \frac{X}{\Delta x} \cdot \frac{Y}{\Delta y} \cdot \log_2 \left(\frac{Z}{\delta z} + 1 \right) \frac{1}{\Delta t},$$

where X and Y are the size of the image field; Z is the brightness coordinate of the image; Δx , Δy , δz , Δt – the discreteness of representing the corresponding coordinates of the image.

The values of X, Y, and Z in the formula are usually taken to be fixed and equal to the maximum value; the values of Δx , Δy , δz , Δt are also fixed, so this approach gives an upper estimate of the amount of information that is very overestimated.

The considered dynamic models, by changing the parameters in the formula, allow you to allocate useful (dynamic) information relevant to the mode of perception, greatly reducing the redundancy of the representation of the image and the cost of its transmission and processing. These dynamic models have become the basis for the development of methods for dynamically managing the parameters of reading information from a video sensor. Modern K-MON-video sensors have in their composition, in addition to the sensor matrix, several hundred registers for adjusting the readout parameters of the sensor matrix and up to ten specialized processors for the preliminary technological preparation of the image prior to use. Effective realization of these possibilities ensures time alignment of the processes of input and processing of information, as a result of which there is an opportunity to obtain parameters for managing the reading of the next frame with minimal delay information after the processing of the current frame [9].

Since the human eye reacts not to the amount of luminance or chromaticity in the image, but to changes between the luminance values of neighboring receptors, or the luminance values of a given receptor in time, that is, to the dynamics of this parameter, as the dynamics of the image are proposed, it is proposed to select the concepts of δ -entropy, the average value of the derivative of the rows and columns of the image [6,10].

In the discrete form δ -entropy the image is defined as

$$H_\delta = \frac{1}{MN} \sum_{j=1}^M \sum_{i=1}^N \left(\left| \frac{\Delta Z_{ij}}{\delta z} \right| \right),$$

where M, N - the size of the image field,

ΔZ – the differences between the brightness of the pixels in rows or columns.

In contrast to the method of averaging the transverse sections of the image brightness profile by rows and columns [11] to determine the location of the object with the background in the image, it is suggested to use δ -entropy. That is, averaging modulo transverse slices of the image with brightness differences between adjacent pixels in rows and columns [9]:

$$H\delta_i = \frac{1}{M} \sum_{j=1}^M \left| \frac{\Delta Z_{ij}}{\delta z} \right|; \quad H\delta_j = \frac{1}{N} \sum_{i=1}^N \left| \frac{\Delta Z_{ij}}{\delta z} \right|.$$

Such a dynamic amount of information can be effectively used to segment the image on a high and low dynamic domain, segmentation of textual information in the image, search and classification of textures (Fig.1), search of car numbers (Fig.2), bar codes, DMX codes (Fig.3), fingerprints, character recognition (Fig.4), control of the shooting frequency of the camcorder, and the like.

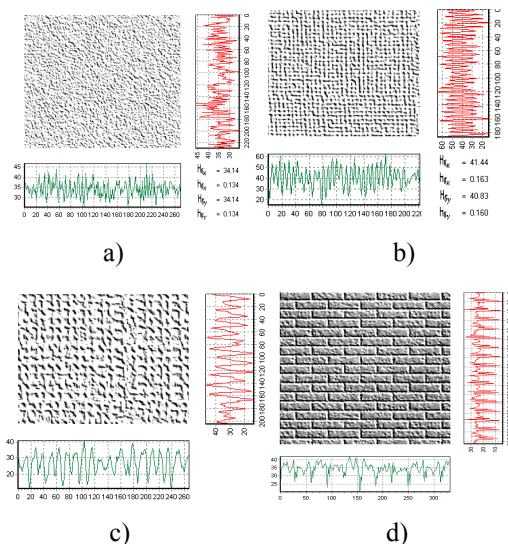


Fig. 1. Evaluation of texture parameters: a) sand, b) linen, c) sackling, d) brickwork.

The δ -entropy makes it easier to compare of the texture, the size of the grain with respect to spatial frequencies, the contrast of the amplitude of the differences, to reveal defects in texture images (violation of regularity, Fig. 1,c).

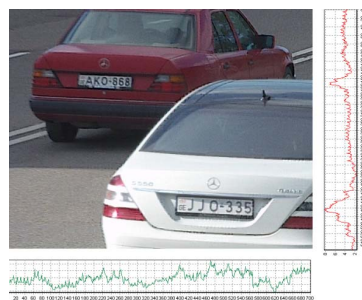


Fig. 2. Searching of the car number in the car image.

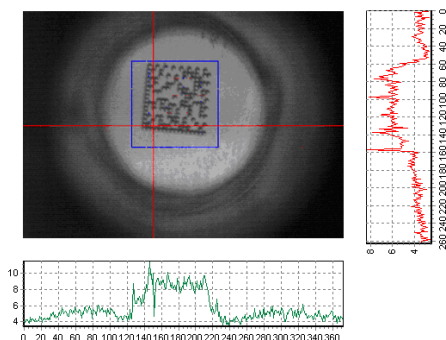


Fig. 3. Searching of the DMX-code in the micro image.

For the recognition of symbols (Fig. 4), a modification of the concept of δ -entropy is used, in particular, the number of swings from "0" to "1" and vice versa. In this case, for noisy images, it is advisable to enter a check on the threshold of the values of the differences or to carry out a low-frequency filtration of the image. It will be effective to use the integral value of δ -entropy in rows and columns as a characteristic vector, or even their sum, that is, one value of the characteristic.

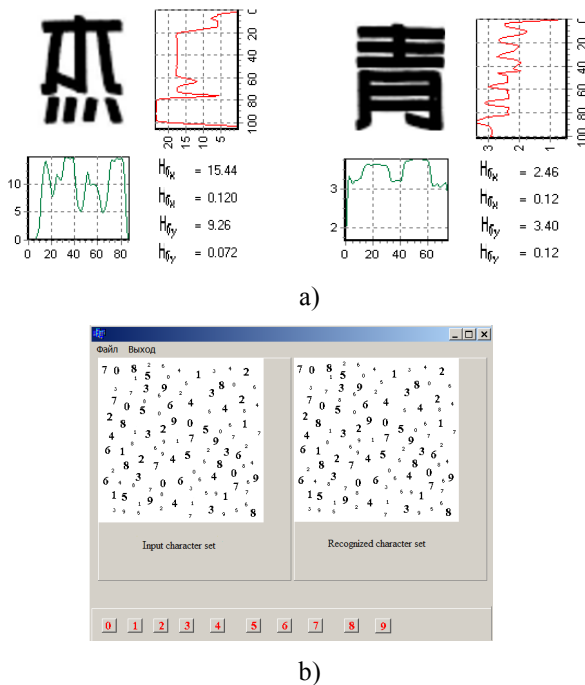


Fig. 4. Character recognition based on δ -entropy: a) hieroglyphs; b) characters with different scales.

Expansion of the dynamic range of computer vision systems can be achieved by non-linear perception of brightness in an analog-to-digital conversion, using as a prototype a visual analyzer of a person.

The neural network organization of calculations is extremely effective, there are already dozens of variants for solving various problems, but they are extremely complicated with hardware implementation and require a complex adjustment to the task. As a prototype for processing information in the video system at the lower level, it is suggested to use the principles of short-range interaction with the ring organization of the on- and off-centers of the human eye, specialization of the layers of neurons, feedback between cells for controlling perception, adapting the sizes and forms of receptor fields (so-called plasticity of neurons).

The central fovea of the retina is specialized for clear vision and is organized on cones, horizontal (HC), bipolar (BC) and ganglionic P-type cells (G_pC) on a ring basis ("on" - and "off" centers) (Fig.5) Horizontal cells are inhibitory. Feedback through the amakrinovi (AC) cells control the perception of contrast by changing the threshold or by building up layers of neurons around the central. Such organization of the neural network is in good agreement with the arsenal of methods for distinguishing various informative features using the masks of Laplace, Sobel, Previti, Roberts and others [2]. Thanks to this, it is possible to effectively

emphasize the contours, to highlight the features of the edge, to identify informative points, lines, their orientation, calculate gradients and etc. The ring organization of the neurons of the central fovea is considered to increase the contrast, which increases with the buildup of layers of neurons around the central element. The organization of connections between neurons is quite universal and enables the implementation of a convolution with matrices 3x3, 5x5, ... due to an increase in computation time on such a structure. All calculations on the structure are performed in parallel. It should also be noted that the calculation of the sums from the outputs of ring neurons and the calculation of the difference between the exciting and inhibitory neurons is carried out by a sequential code, it makes it possible to implement on this circuit not only the bit codes after the threshold limitation, but also the continuous codes. In this case, the computation time increases almost in proportion to the full-bit codes. The coefficients of the matrices can almost always be taken as numbers proportional to the power of the two (0, 1, 2, 4, 8, ...), which eliminates the need for multiplication.

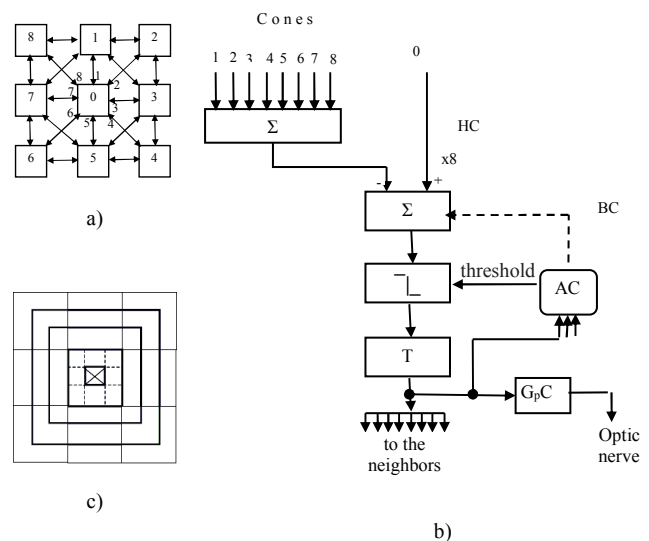


Fig. 5. Horizontal and bipolar cells with increasing of rings: a) circular organization of neurons in the central fovea ("on" - center): element "0" in the center - excitatory, elements "1-8" around - inhibitory; b) implementation of the "on" - centre of 3x3; c) rings around the central element to increase sensitivity to the perception of contrast.

The periphery of the retina is a specialized on high sensitivity and is organized on the sticks of the retina, diffuse (DC), bipolar (BC) and ganglionic M-type cells ($G_M C$) also on the ring principle (Fig. 6). Increased sensitivity in conditions of insufficient illumination is ensured by the summation of signals from a large number of rods and the action of inhibitory diffuse cells. At the same time, accordingly, the spatial resolution is reduced. Interphase-shaped (IPSC)-linked cells control thresholds or receptor field sizes.

Using these methods, a number of specialized technical solutions protected by patents for inventions have been developed to combine the processes of perception of video information with its processing directly on the sensor array. In particular, these are sensor matrices with parallel binarization of the image and determination of the location and parameters of the object, with the calculation of the first and second moments of inertia binarized image for rows, columns and the whole image, with morphological processing of binarized images, parallel analog-to-digital

conversion and the possibility of nonlinear perception of the brightness. As an example, Fig.7 shows a generalized block diagram of one layer of a sensor array with image processing. Each element of the matrix has connections only with neighboring elements of the matrix (locality principle), that is, by analogy with the human visual analyzer, can realize “on”- and “off”-centers with a ring organization.

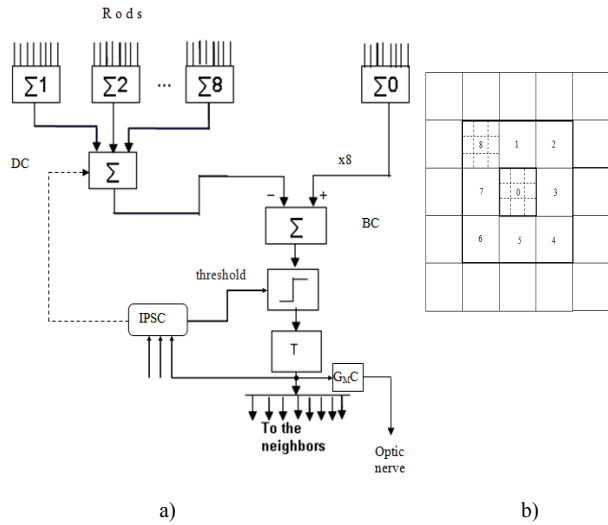


Fig. 6. Diffuse and bipolar cells with increasing rings: a) the implementation of the ring organization of the peripheral neurons of the retina; b) rings for increased sensitivity in low light conditions.

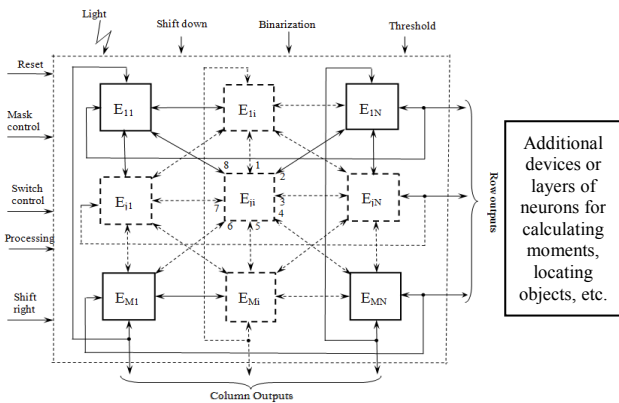


Fig. 7. Structure of the sensor matrix with image processing.

IV. CONCLUSIONS

Thus, the use of the principles of construction and operation of the human visual analyzer as a prototype of the computer vision system allows us to develop a whole arsenal of methods for increasing the efficiency of the processes of perception, processing and recognition of images.

REFERENCES

- [1] R. Schiffmann, Sensation and perception. 5-th ed. Piter, SPb., Russia, 2003. (in Russian)
- [2] R. Gonsales, and R. Woods, Digital image processing. Moscow, Russia, Technosphere, 2005. (in Russian)
- [3] S. Shan, and M. D. Levine, "Visual Information Processing in Primate Cone Pathways - Part 1: A Model. Part 11: Experiments," IEEE Trans. On Systems, Man, and Cybernetics - Part B: Cybernetics, vol.26, no.2, Apr. 1996.
- [4] D. Anderson, Cognitive psychology. 5-th ed. Piter, SPb., Russia, 2002. (in Russian)
- [5] P. J. Burt, "Smart Sensing within a Pyramid Vision Machine," IEEE, vol. 76, no. 8, pp. 175-185, 1988.
- [6] V. Boyun, "Intelligent selective perception of visual information," Informational aspects. Artificial intellect. no. 3. pp.16-24, 2011. (in Ukrainian)
- [7] V. Boyun, "Intelligent Selective Perception of Visual Information in Vision Systems," 6-th IEEE International Conference on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Application. (IDAACS'2011). Prague, Czech Republic, vol.1, pp. 412-416, 2011.
- [8] V. Boyun, "A human visual analyzer as a prototype for construction of the set of dedicated systems of machine vision," International conference "Artificial intelligence. Intelligent systems. II-2010, vol. 1, pp. 21-26, 2010. (in Ukrainian)
- [9] V. Boyun, "Directions of Development of Intelligent Real Time Video Systems," Application and Theory of Computer Technology, [S.l.], vol. 2, no. 3, pp. 48-66, 2017. ISSN 2514-1694. Available at: <<http://www.archyworld.com/journals/index.php/atct/article/view/65>>. Date accessed: 26 sep. 2017. doi: <https://doi.org/10.22496/atct.v2i3.65>.
- [10] V. Boyun, The dynamic theory of information. Fundamentals and applications. Institute of Cybernetics of NASU, Kyiv, Ukraine, 2001, (in Russian)
- [11] W. Prett, Digital Image processing. vol. 2, Mir, Moscow, USSR, 1982., (in Russian)

Model and Training Methods of Autonomous Navigation System for Compact Drones

Viacheslav Moskalenko
dept. of computer science
Sumy State University
Sumy, Ukraine

v.moskalenko@cs.sumdu.edu.ua

Olha Boiko
dept. of computer science
Sumy State University
Sumy, Ukraine

o.shulyma@ssu.edu.ua

Alona Moskalenko
dept. of computer science
Sumy State University
Sumy, Ukraine

a.moskalenko@cs.sumdu.edu.ua

Serhii Martynenko
dept. of computer science
Sumy State University
Sumy, Ukraine

serg.martynenko@gmail.com

Artem Korobov
dept. of computer science
Sumy State University
Sumy, Ukraine

a.korobov@cs.sumdu.edu.ua

Oleksandr Borovenskyi
dept. of computer science
Sumy State University
Sumy, Ukraine

aleks.borovensky@gmail.com

Abstract — The paper presents a novel model of convolutional neural network for visual feature extraction, support vector machine for position prediction and information-extreme classifier for obstacle prediction with new training methods to build decision rules of autonomous navigation system for compact drones are presented in the paper. Sparse-coding neural gas algorithm for unsupervised training of the convolution filters, supervised incremental learning method for training the regression model and particle swarm optimization algorithm for training the classifier model are proposed. The complex criterion for choosing parameter of feature extractor model is considered. Simulation results with optimal model on test open datasets confirm the suitability of proposed algorithms for practical usage.

Keywords—navigation, visual odometry, convolutional neural network, neural gas, information criterion, support vector regression

I. INTRODUCTION

Unmanned aerial vehicles (UAV) are widely used in precision agriculture, search and rescue operations, transport and aerial filming. Development information technology which lowers the demands on UAV hardware resources and improves the reliability of autonomous decision-making under constantly changing environmental conditions and variability of objects of interest allows to reduce the system's weight and cost whilst simultaneously expanding the functionality of the onboard system. One of the ways to improve the functional efficiency of the UAV system is to use machine vision and machine learning to build data analysis models based on visual and inertial sensor data [1-3].

The use of functional navigation systems based on the comparison of visual features requires the availability of a database of reference images. This makes is potentially unsuitable for situations requiring a rapid response to changes in the environment [4]. Visual odometry and Simultaneous Localization And Mapping (SLAM) methods are less efficient in poorly textured scenes and in the presence of non-static elements in the field of view [5, 6]. Moreover, the deployment of these technologies requires significant computational resources, which limits their use in autonomous compact UAVs.

Today, the convolutional neural network, consisting of a multilayer feature extractor based on convolution filters and decisive rules in the form of fully connected neural layers, is an undisputed leader among the image analysis models [7, 8]. However, the essential disadvantages of traditional convolutional neural networks lay in their inability to analyze the processes occurring in time, as well as the high computational complexity of the backpropagation-based learning algorithm, which makes adapting to changes in operating conditions difficult. Conversely, using unsupervised learning methods based on sparse-coding neural gas to train neural networks shows promise. It reduces both the required quantity of labeled observations and computational load [9].

This paper proposes a model of convolutional neural network for analysis of spatial-temporal patterns to be used in autonomous navigation and identification of obstacles under computational resource constraint. We also propose a training method for such network based on unsupervised learning combined with decision rules based on support vector machines [10] and intellectual information-extreme technology [11]. The results of parameter optimization and testing of proposed algorithms on real-life open source data sets are considered.

II. MATERIALS AND METHODS

Let an annotated set of video frames be formed $\{c_t = \langle v_t, x_t, y_t, z_t, a_t \mid t = \overline{1, n} \rangle\}$, where v_t – frame image at time t and x_t, y_t, z_t – camera coordinates obtained from the Global Positioning System (GPS) and converted to the North East Down (NED) local coordinate system, $a_t \in \{A_r^o \mid r = \overline{1, R}\}$ – operator's response to the obstacle, where A_r^o denotes a recognition class that characterizes the obstacle.

A structured vector of space-time parameters of the UAV navigation system operation in general has a structure :

$$g = \langle e_1, \dots, e_{\xi_1}, \dots, e_{\xi_2}, f_1, \dots, f_{\xi_2}, \dots, f_{\xi_2} \rangle, \Xi_1 + \Xi_2 = \Xi, \quad (1)$$

where $\langle e_1, \dots, e_{\xi_1}, \dots, e_{\xi_1} \rangle$ – genotype model parameters which affect the parameters of the feature extraction algorithms; $\langle f_1, \dots, f_{\xi_2}, \dots, f_{\xi_2} \rangle$ – phenotypic model parameters which influence the decision rules.

At the same time, known limitations on the corresponding model parameters are:

$$R_{\xi_1}(e_1, \dots, e_{\xi_1}, \dots, e_{\xi_1}) \leq 0; R_{\xi_2}(f_1, \dots, f_{\xi_2}, \dots, f_{\xi_2}) \leq 0.$$

The process of machine learning of the navigation system is focused on determining the optimal coordinate values of the vector (1), which provide the maximum of the complex criterion

$$J = \frac{\bar{E}}{E_{\max}} \cdot \frac{\varepsilon_{\min}}{\varepsilon} \cdot \frac{C_{\min}}{C}, \quad (2)$$

$$g^* = \arg \max_g \{J(g)\}, \quad (3)$$

where \bar{E} – information criterion of learning effectiveness for the recognition of obstacle averaged by the set of classes; ε – the value of the mean square error of regression when determining the change of camera coordinates in space; C – the criterion of computational complexity of feature extraction algorithms; E_{\max} , ε_{\min} , C_{\min} – the maximum possible value of the informational criterion of classifier training efficiency and the minimum allowable values of regression model error and the criterion of computational complexity of the system's algorithms, respectively.

For the formation of the input mathematical description of the intelligent information system, KITTI Vision Dataset [8] training kits a containing both the frame sequence of the image from the moving video camera and the movement data along three coordinate axis reported by GPS and LiDaR [8]. To train the model, movement data is converted to the local NED coordinate system and the relative movements of the camera $\Delta x, \Delta y, \Delta z$ are determined between adjacent video frames.

The schematic of the intelligent navigational system for a compact UAV is shown in Fig. 1. In order to extract the feature representation of visual observations, it is proposed to use a convolutional neural network, using a multichannel image formed by a series sampling of successive video footage in grayscale format as input. The convolutional neural network has a multilayered structure to form a high-level feature representation of observation results, with convolutional filters trained in unsupervised manner successively layer by layer. An information-extreme classifier trained in supervised mode on the training samples encoded by the corresponding high-level features is used for obstacle prediction and output of the corresponding reaction. The regression model in the framework of the support vector machine is used to map the visual features and the data from inertial sensors into the corresponding estimation of the displacement of the video camera in space.

Fig. 2 shows the 4-layer architecture of the convolutional neural network, in the first layer of which there are 3D-filters of different scales: $5 \times 5 \times K_1$, $3 \times 3 \times K_1$ and $1 \times 1 \times K_1$. The number of filters is regulated by the parameter K_2 . To preserve the same size of character maps formed by multiple-scale filters, the technique of padding with zeros is used [8]. In the second and third layers, stride parameter of scanning a feature map with multiple-scale filters is 3 and 2, respectively.

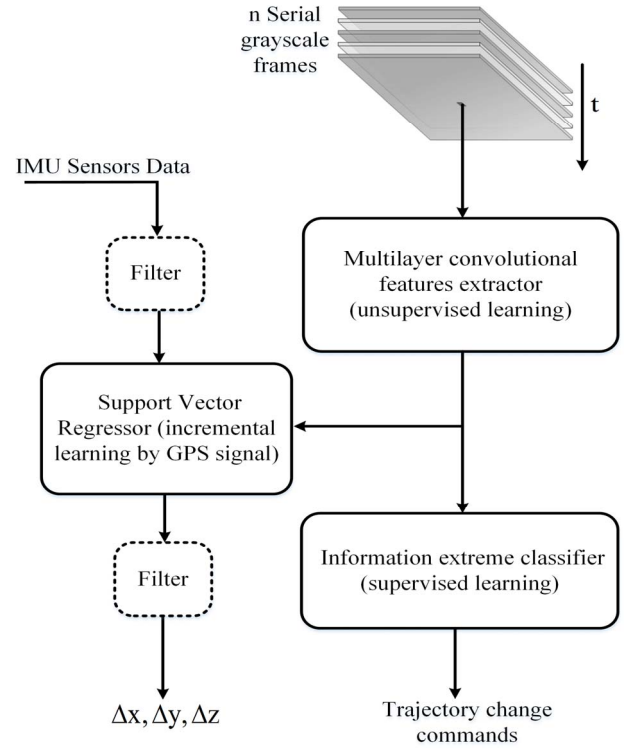


Fig. 1. A generalized scheme of intelligent navigational system of small size UAV

Fig. 2 does not show the activation function applied to each feature map. We propose to use the Orthogonal Matching Pursuit algorithm [9] for calculating a response on each feature map and rectifier function $y = \max(0, x)$, however, to avoid information loss, we can double the feature map using the following function: $y = \{\max(0, x), \max(0, -x)\}$.

An important step of data analysis is a preliminary normalization with the view to removing linear correlation of components of observation and the unification of primary feature representation. Data whitening with the use of the method of ZCA (Zero-phase Component Analysis) is one of the most common methods of preliminary data normalization. ZCA method implies performance of the following steps:

- 1) calculation of mean selected value of features $\mu = \text{mean}(X)$;
- 2) calculation of co-variative matrix of selected observations $\Sigma = \text{cov}(X)$;
- 3) singular decomposition of co-variative matrix $\Sigma \approx VDTT$;

4) whitening of each observation by formula $x_j := VD^{-1/2}V^T(x_j - \mu)$.

Unsupervised learning of convolutional filters is proposed to be carried out in accordance with the algorithm of sparse-coding neural gas, which was considered and studied in [9]. The input data for the algorithm of sparse-coding neural gas is the power of the set of the basis vectors M , the dimension of feature space N , $\lambda_0, \lambda_{final}$ – the initial and final value of the neighborhood size, η_0, η_{final} – the initial and final values of the learning rate.

Consider the main steps of the algorithm.

1) Initialization of the dictionary of basis vectors $D = (d_1, \dots, d_M)$ by random numbers with uniform distribution;

2) Initialization of the counter of training vectors $t := 1$.

3) Choosing a random vector x from the set of training vectors X .

4) The normalization of vectors from the dictionary $D = (d_1, \dots, d_M)$ by bringing it to a unit length.

6) The similarity calculation of the input vector x to the basis vectors $d_{l_k} \in D$ for their sorting

$$-(d_{l_0}^T x)^2 \leq \dots \leq -(d_{l_k}^T x)^2 \leq \dots \leq -(d_{l_{M-1}}^T x)^2$$

7) Update the coordinates of the main vectors $d_{l_k} \in D$ according to the Oja's learning rule [9]

$$d_{l_k} := d_{l_k} + \eta_t \exp(-k / \lambda_t) y (x - y d_{l_k}), \quad y := c_{l_k}^T x, \\ k = \overline{0, M-1}.$$

8) If $t < t_{max}$, then the increment of the counter $t := t + 1$ and go to the step 3.

The information-extreme classifier for evaluation of the obstacle performs the adaptive discretization of the feature representation of dataset $\{x_{r,i}^{(j)} \mid i = \overline{1, N}; j = \overline{1, n_r}; r = \overline{1, R}\}$ on the basis of the coarse binary coding algorithm. This involves comparing the value of the i -th feature with the corresponding lower $T_{L,l,i}$ and upper $T_{U,l,i}$ thresholds of the asymmetric receptive field l , which are calculated by the formulas

$$T_{L,l,i} = x_{i,\max} \left[1 - \frac{\delta_{l,i}}{\delta_{\max}} \right], \quad T_{U,l,i} = x_{i,\max}, \quad l = \overline{1, L}$$

The formation of a binary training set $\{b_{r,i}^{(j)} \mid i = \overline{1, N * L}; j = \overline{1, n_r}; r = \overline{1, R}\}$ is carried out according to the rule

$$b_{r,i}^{(j)} = \begin{cases} 1, & \text{if } T_{L,l,i} \leq x_{r,i}^{(j)} \leq T_{U,l,i}; \\ 0, & \text{else.} \end{cases}$$

The calculation of the values of the coordinates of the binary support vector x_m , relative to which container classes are constructed on a radial basis, is carried out according to the rule

$$b_{r,i}^{(j)} = \begin{cases} 1, & \text{if } \frac{1}{n_r} \sum_{j=1}^{n_r} b_{r,i}^{(j)} > \frac{1}{n} \sum_{r=1}^R \sum_{j=1}^{n_r} b_{r,i}^{(j)}; \\ 0, & \text{else.} \end{cases}$$

Normalized modification of S. Kullback's information measure is used as a criterion of our classifier's machine learning efficiency [11]:

$$E_r = \frac{1 - (\alpha_r + \beta_r)}{\log_2(2 + \zeta) + r \log_2 10} \cdot \log_2 \left[\frac{2 - (\alpha_r + \beta_r) + \zeta}{(\alpha_r + \beta_r) + \zeta} \right], \quad (4)$$

where α_r, β_r – false-positive and false-negative rates of classification decisions regarding the affiliation of the input

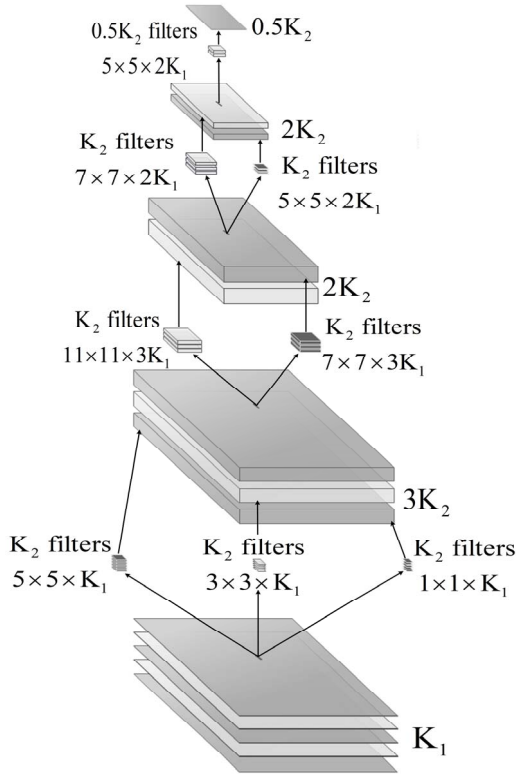


Fig. 2. The architecture of the convolutional neural network for visual feature extraction in the UAVs navigation system

5) Calculation of the current values of the neighborhood size λ_t and learning rate η_t :

$$\lambda_t := \lambda_0 (\lambda_{final} / \lambda_0)^{t/t_{max}};$$

$$\eta_t := \eta_0 (\eta_{final} / \eta_0)^{t/t_{max}}.$$

vectors to the class A_r^o ; ς – any small positive number entered to avoid uncertainty when dividing by zero.

The complexity of information-extreme machine learning increases faster than the square of the number of training vectors. Therefore, a reduction of multi-class classifier to a series of two-class classifiers is used to speed up training. The classifiers are constructed on the principle of "one-against-one", with a total of $M \cdot (M-1) / 2$ two-class classifiers constructed [11].

In the exam mode, the decision on the affiliation of the observation x to one of the classes of set $\{A_r^o\}$ is taken according to the geometric membership function [11]

$$\mu_r^*(x) = \max_{\{r\}} \{\mu_r(x)\},$$

where $\mu_k(x)$ is the membership function of vector x to the container of class $\{A_r^o\}$ which is calculated by the rule:

$$\mu_r(x) = \exp \left(- \frac{\sum_{i=1}^{N \cdot L} (x_{r,i}^* \oplus x_i)}{radius_r^*} \right),$$

where $radius_r^*$ is the optimal radius of class container A_r^o .

To train the regression model $y = f(x)$, output variable $y^{(t)} \in R$ of which corresponds to the change of the coordinates of the camera $\Delta x, \Delta y$, or Δz , a set of $(x^{(t)}, y^{(t)})_{t=1}^n$ training data, consisting of visual features and measurements of inertial sensors, is used, where $x^{(t)} \in R^N$. The regression function is linear in the secondary feature space and has the following form

$$f(x) = (\omega, \varphi(x)) + b, \quad (5)$$

$$\varphi: R^n \rightarrow H, \omega \in H. \quad (6)$$

where ω and b are empirical coefficients which can be obtained through training; H – multidimensional space of secondary features.

The coefficients ω and b can be found by minimizing the following formula:

$$\min R(\omega, \xi, \xi^*) = \frac{1}{2} \|\omega\|^2 + \Psi \sum_{i=1}^n (\xi_i^* + \xi_i)$$

$$\begin{aligned} y^{(t)} - (\omega, \varphi(x)) - b &\leq \varepsilon + \xi_i^* \\ (\omega, \varphi(x)) + b - y^{(t)} &\leq \varepsilon + \xi_i \\ \xi_i^*, \xi_i &\geq 0, i = 1, 2, \dots, n, \varepsilon \geq 0. \end{aligned} \quad (7)$$

where Ψ – coefficient of regularization; ξ, ξ^* – slack variable, which measures the measurement uncertainty from below and above, respectively; ε – the insensitivity of the loss function, which means that if $f(x)$ is in the range $y^{(t)} \pm \varepsilon$, then the measurement uncertainty is not taken into account.

The optimization problem (7) is a quadratic programming problem with linear constraints, which can be solved by introducing Lagrange multipliers and applying the Karush-Kuhn-Tucker conditions to solve a dual problem. [10]:

$$\begin{aligned} \min R(v, v^*) &= \sum_{t=1}^n (v_t^* - v_t)(v_j^* - v_j) K(x^{(t)}, x^{(j)}) + \\ &+ \varepsilon \sum_{t=1}^n (v_t^* + v_t) - \sum_{t=1}^n y^{(t)} (v_t^* - v_t) \\ &\sum_{t=1}^n (v_t^* + v_t) = 0 \\ 0 \leq v_t, v_t &\leq \frac{\Psi}{l}, t = 1, 2, \dots, n \end{aligned} \quad (8)$$

where v_t and v_t^* are Lagrange multipliers associated with constraints (8);

$$K(x^{(t)}, x^{(j)}) = \varphi(x^{(t)}) \cdot \varphi(x^{(j)}).$$

A typical example of a kernel function is a polynomial kernel and a Gaussian kernel. In general, the regression function has the form

$$f(x) = \sum_{t=1}^n (v_t^* - v_t) K(x^{(t)}, x) + b$$

Not all training samples can become support when support vectors are used for training the regression model. Only vectors on the boundary have the probability of becoming support. The incremental training of a regression model on support vectors can be realized by determining the convex border of discrete points when choosing a set of boundary vectors as a set of training ones. In this case, the convex border of discrete points is the border, which can surround all the discrete points, formed by the outermost point through connections. Therefore, after processing of the first sub-sample, the formed support vectors are compared with the vectors of the following subclasses at an angle of inclination, to form a plurality of boundary vectors, as vectors of the maximum inclination. Each step of the supplement may be accompanied by retraining. (7)

III. SIMULATION RESULTS AND DISCUSSION

To train the feature extractor, both the training and test video sequences of the KITTI Vision Dataset set are used, without taking annotations into account. To reduce the computational complexity of the algorithms, the images are compressed to a resolution of 200x200 pixels. In this case, the procedure is repeated for different values of parameters K_1 and K_2 , which affect both the informative nature of the

feature representation and the computational complexity. We propose to measure complexity by the quantity of Mul and Add operations performed during the convolutional operations with an image or a feature map. For the network architecture shown in Fig. 2, the complexity can be calculated as

$$C = K_2(2706472K_1 + 4438784K_2). \quad (9)$$

For the classifier and regression model, the optimal configuration of the convolutional extractor may be different as they are responsible for different tasks. Therefore, a complex criterion (2) offers a compromise from the point of view of the accuracy of the decision rules and the computational complexity of the extractor of visual features.

For our support vectors of our regression model, we propose to use a Radial basis kernel in the following form

$$K(x^{(i)}, x^{(j)}) = \exp(\gamma \|x^{(i)} - x^{(j)}\|^2), \gamma \geq 0$$

where γ – the kernel coefficient, the default value of which is $\gamma = 1/N$.

The set of recognition classes $\{A_r^o\}$ is describing the characteristic obstacles and the corresponding reaction commands, and has a power $R=5$. The first class of recognition A_1^o characterizes the normal state of following a prescribed trajectory. The classes A_2^o and A_3^o correspond to the left turn of 45 and 90 degrees respectively. The classes A_4^o and A_5^o correspond to the right turn of 45 and 90 degrees respectively. The volume of the training samples of each class is $n_r = 500$.

The optimization of the parameters of the receptive field $\{\delta_{m,i}\}$ and other genotype parameters for the information-extreme classifier amounts to finding the extremum of the criterion function (4) in the hyperspace of solutions. For the purposes of this it is suggested to use a Particle Swarm Optimization algorithm (PSO) [11]. The effectiveness of each particle of a population algorithm, which lies in its proximity to the global optimum, is measured by means of a predetermined fitness function. This role is fulfilled in our case by the training efficiency criterion (4). In this case, the following parameters of the population algorithm configuration are specified: maximum particle speed $V_{\max,i} = 2$, particles acceleration constants $c_1 = c_2 = 1$, the number of swarm agents $n_a = 100$, the coefficient of inertia $w = 0,95$ and the number of iterations $K_{ITER} = 3000$.

The optimization of the phenotypic parameters of the decision rules (radii of container classes) can be carried out by the direct search with a given step, since the number of steps for such a search is relatively small. To identify the tendency of changing in average values of the partial and complex criteria during the growth of parameters K_1 and K_2 , which affect the size of the convolution extractor (Fig. 2), a simulation was performed for the three fixed values of each of these parameters (Table I).

TABLE I. DEPENDENCE OF PARTIAL AND COMPLEX CRITERIA ON EXTRACTOR PARAMETERS OF A FEATURES DESCRIPTION K_1 AND K_2

K_1	K_2	\bar{E} / E_{\max}	$\epsilon_{\min} / \epsilon$	C_{\min} / C	J
3	4	0,083	0,112	1,000	0,009296
5	4	0,101	0,188	0,827	0,015703
7	4	0,098	0,200	0,705	0,013818
3	8	0,28	0,688	0,297	0,057214
5	8	0,29	0,756	0,264	0,057879
7	8	0,29	0,775	0,238	0,053491
3	16	0,39	0,968	0,082	0,030957
5	16	0,55	1,000	0,077	0,04235
7	16	0,51	1,000	0,072	0,03672

The analysis of table 1 shows that an increase in parameter values K_1 and K_2 in general leads to an increase in the reliability and computational complexity (9) of the decision rules of the classifier and the regression model. At the same time, the increase of the parameter K_1 has little effect on the efficiency of the classifier due to the decrease in the efficiency of the swarm search with a significant increase in the size of the feature space, while the regression error is equally sensitive to the value of parameters K_1 and K_2 .

However, given that with growth in K_1 and K_2 the reliability of decision rules grows more slowly than the computational complexity use of complex criterion J offers a suitable compromise. That is, we consider the following parameter values to be optimal $K_1^* = 5$ and $K_2^* = 8$.

In the optimal configuration of the feature extractor, the average value of the information criterion of functional efficiency is $\bar{E} = 0,29$. This corresponds to accuracy of 95,2% for the training set, and 94% for the test dataset. The number of receptive fields per primary feature is $L = 3$, chosen as the minimum value at which the information criterion (4) ceases to grow on the test dataset. Fig. 3 shows a graph of the change of the average information efficiency criterion (4) in relation to the number of iterations of the particle swarm search algorithm.

The analysis of Fig. 3 shows that after a 1000th iteration growth of the information criterion (4) has begun to slow down, and after 2500th iteration remained virtually unchanged. Such a change in the criterion indicates that the further increase in the information criterion is achievable only with the increase in the informative nature of the features by increasing values of K_1 and K_2 or improving the structure of the extractor (Fig. 2).

For a visual assessment of the effectiveness of the machine learning of the navigation system, a reference trajectory measured using GPS and LiDaR can be compared with a reconstructed trajectory obtained using a trained model.

Fig. 4a shows the reference trajectory (dashed line) and the reconstructed trajectory (solid line) created by the proposed algorithms on the basis of the test data from the KITTI database [8]. Fig. 4b shows the results of a similar experiment, but using the model proposed in [8].

The analysis of Fig. 4 shows that the accuracy of reconstruction of the trajectory in both cases is acceptable for practical use and does not differ significantly. Notably, however, the proposed model has much fewer parameters and allows the use of an unsupervised training instead of a computationally intensive gradient descent algorithm.

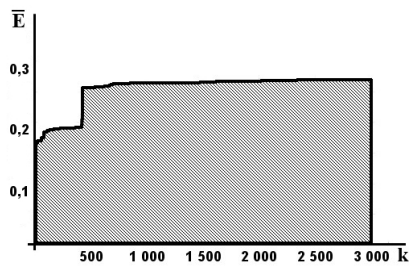


Fig. 3. A graph of the change of the average information efficiency criterion (4) in dependence from the number of iterations of the optimization swarm search algorithm

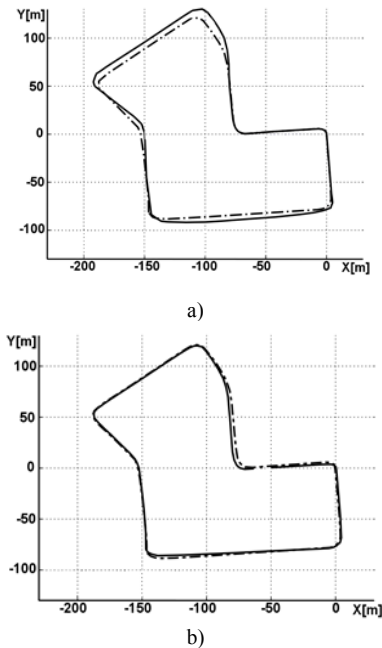


Fig. 4. Reference and reconstructed trajectory: a – the developed model; b – the model proposed in the work [9]

IV. CONCLUSION

1. The scientific novelty of the results is as follows:

- a new model for the autonomous navigation system of a compact UAV is proposed for the first time. The constituent parts of the model are a feature extractor trained without supervision, a support vector regression model, which can be incrementally trained under supervision on the visual and inertial sensor data, and an information-extreme obstacle classifier, which learns to react to obstacles under supervision, which in turn reduces the computational resource requirements;

- a model of a 4-layer convolutional network using as inputs a series of successive frames which are interpreted as channels of one image and scanned by multiple-scale filters is proposed for the first time;

- a method of unsupervised training of the convolution filters based on sparse-coding neural gas, which allows training simultaneous with direct propagation of the signal without using the back error propagation is proposed for the first time;

- a method for evaluating the effectiveness of the data analysis model in navigation problems was improved with

the application of multiplicative convolution of partial criteria. This allows to select the optimal system parameters in the information and computational cost sense.

2. The practical value of the obtained results for unmanned aviation lies in the formation of a modern scientific and methodological basis for designing compact autonomous navigation systems for UAVs operating under resource and information constraints, and capable of learning. At the same time, the results of the simulation model confirm the high efficiency of the resulting decision rules for determining the coordinates in space and recognition of obstacles based on the video stream and inertial sensor data.

ACKNOWLEDGMENT

The work was performed in the laboratory of intellectual systems of the computer science department at Sumy State University with the financial support of the Ministry of Education and Science of Ukraine in the framework of state budget scientific and research work of DR No. 0117U003934.

REFERENCES

- [1] S. Wang, Z. Deng, and G. Yin, "An Accurate GPS-IMU/DR Data Fusion Method for Driverless Car Based on a Set of Predictive Models and Grid Constraints," Basel, Switzerland: Multidisciplinary Digital Publishing Institute, vol. 16(3), pp. 280–293, 2016.
- [2] B. Suwandi, T. Kitasuka, and M. Aritsugi, "Low-cost IMU and GPS fusion strategy for apron vehicle positioning," TENCON 2017, IEEE Region 10 Conference, Penang, Malaysia, pp. 449–454, Nov. 2017.
- [3] B. A. Mary, and P. H. Gerhard, "Pose Estimation of a Mobile Robot Based on Fusion of IMU Data and Vision Data Using an Extended Kalman Filter," Basel, Switzerland: Multidisciplinary Digital Publishing Institute, vol. 17(10), pp. 2164, 2017.
- [4] J. Folkesson, J. Leederkerken, R. Williams, and A. Patrikalakis, "A Feature Based Navigation System for an Autonomous Underwater Robot," In: Laugier C., Siegwart R. (eds) Field and Service Robotics. Springer Tracts in Advanced Robotics. Springer: Berlin/Heidelberg, 2008. vol. 42, pp. 105-114.
- [5] C. Cadena, L. Carlone, H. Carrillo, Y. Latif, D. Scaramuzza, J. Neira, I. Reid, and J. J. Leonard, "Past, Present, and Future of Simultaneous Localization And Mapping: Towards the Robust-Perception Age," IEEE Transactions on Robotics, vol. 32, no. 6, pp. 1309–1332, 2016.
- [6] G.-L. Dorian, S. Marta, D. T. Juan, and J.M.M. Montiel, "Real-time monocular object SLAM. Robotics and Autonomous Systems," North-Holland Publishing Co.: Amsterdam, Netherlands, 2016. vol.75, i. PB, pp. 435–449.
- [7] Th. Ayoul, T. Buckley, and F. Crevier. UAV Navigation above Roads Using Convolutional Neural Networks Available from: <http://cs231n.stanford.edu/reports/2017/pdfs/553.pdf>.
- [8] V. Mohanty DeepVO: A Deep Learning approach for Monocular Visual Odometry. Computer Vision and Pattern Recognition, 2016. Available from: <https://arxiv.org/pdf/1611.06069.pdf>
- [9] K. Labusch, E. Barth, and T. Martinez, "Sparse Coding Neural Gas: Learning of Overcomplete Data Representations," Neurocomputing, Elsevier Science Publishers B. V.: Amsterdam, Netherlands, vol. 72, is. 7–9, pp. 1547–1555, 2009.
- [10] H. Xu, R. Wang, and K. Wang. A New SVR Incremental Algorithm Based on Boundary Vector. 2010. Available from: https://www.researchgate.net/publication/238519783_A_New_SVR_Incremental_Algorithm_Based_on_Boundary_Vector
- [11] V. V. Moskalenko, and A. G. Korobov, "Information-extreme algorithm of the system for recognition of objects on the terrain with optimization parameter feature extraction," Radio Electronics, Computer Science, Control. Zaporizhzhya National Technical University: Zaporizhzhya, Ukraine, no. 2, pp. 38–45, 2017.

Braille Character Recognition Based on Neural Networks

Kirill Smelyakov

Department of Electronic Computers
Kharkiv National University of Radio
Electronics
Kharkiv, Ukraine
kirill.smelyakov@gmail.com

Anastasiya Chupryna

Department of Program Engineering
Kharkiv National University of Radio
Electronics
Kharkiv, Ukraine
anastasiya.chupryna@nure.ua

Dmytro Yeremenko

Department of Electronic Computers
Kharkiv National University of Radio
Electronics
Kharkiv, Ukraine
dmytro.yeremenko@nure.ua

Anton Sakhon

Department of Electronic Computers
Kharkiv National University of Radio
Electronics
Kharkiv, Ukraine
anton.sakhon@nure.ua

Vitalii Polezhai

Department of Electronic Computers
Kharkiv National University of Radio
Electronics
Kharkiv, Ukraine
vitalii.polezhai@nure.ua

Abstract—Braille is the most popular system used for interaction between visually-impaired and sighted people using tactile means. Optical Braille character recognition (OBCR) includes two main steps: Braille cells' recognition (image acquisition, preprocessing, Braille dots' recognition, Braille cells' recognition and segmentation) and Braille cells' transcription to corresponding natural language characters. System example has been created using image processing methods and artificial neural networks approach. These methods allow to achieve high speed and recognition accuracy level. System can adapt to factors like quality of input patterns and differences between them dynamically. In this paper, artificial neural network is developed to identify letter's images of Cyrillic alphabet in Braille representation system. Network will be trained and tested for identifying of scanned Cyrillic letters in Braille. Some of the letters are noised with some type of noise to simulate the real-world environment.

Keywords—alphabet, artificial neural network, Braille, character, image processing, image recognition.

I. INTRODUCTION AND PROBLEM STATEMENT

As stated by World Health Organization in 2017, there are an estimated 253 million people around the world live with vision impairment: 36 million of them are blind and 217 million have moderate to severe vision impairment [1]. These people still continue to contribute efficiently to the society nevertheless their disabilities. However, they face with challenging problem about expressing their contributions as they use different scripting language, which makes information transferring between them and the sighted people more difficult. Braille was founded by a French teacher, Louis Braille, in 1824. It represents binary tactile code system used by blind and visually impaired people instead of usual printed reading and writing methods, so they can feel raised dots with tips of their fingers on Braille page. The Braille script represents of cells itself; each of them contains of six raised dots arranged in three rows and two columns as shown in Figure 1. These six dots can be raised or flat according to corresponding Braille character. They are interpreted as series of symbols ranging between 0 and 63, so these dots are combined to give $2^6 = 64$ different sets of

combinations (including the empty Braille character "space"). Mappings (sets of numbers and symbols) are different between languages in Braille alphabets. For example, there are 3 types of characters encoding in English language's Braille alphabet: *Grade 1* is basic literacy-used with letter-by-letter transcription; *Grade 2* is complemented by abbreviations and contractions; *Grade 3* varies from basics with non-standardized personal short hands.



Fig. 1. Braille cell template

Braille paper's size can be also used not just for cell representing the character, but for word or even sentence. Therefore, Braille language consists of three grades: *Grade 1* – grade where each Braille cell represents a single language character and word is constructed of Braille cells' combinations. *Grade 2* is similar to Grade 1 but differs with some abbreviations and contractions. *Grade 3* is the most difficult Braille grade combining complex phrases and sentences. Figures 2, 3 show Grade 1 Braille alphabet for English and Ukrainian languages.

• ○	• ○	• •	• •	• ○	• •	• •	• ○	○ •	○ •
○ ○	• ○	○ ○	○ •	○ •	• ○	• •	• •	• •	• •
○ ○	○ ○	○ ○	○ ○	○ ○	○ ○	○ ○	○ ○	○ ○	○ ○
a/1	b/2	c/3	d/4	e/5	f/6	g/7	h/8	i/9	j/0
• ○	• ○	• •	• •	• ○	• •	• •	• ○	○ •	○ •
○ ○	• ○	○ ○	○ •	○ •	• ○	• •	• •	• •	• •
• ○	• ○	• ○	• ○	• ○	• ○	• ○	• ○	• ○	• ○
k	l	m	n	o	p	q	r	s	t
• ○	• ○	• •	• •	• ○					○ •
○ ○	• ○	○ ○	○ •	○ •					• •
• •	• •	• •	• •	• •					○ •
u	v	x	y	z					w

Fig. 2. Braille alphabet for English language

а	б	в	г	ґ	д	е	є	ж	з	и
•○ ○○ ○○	•○ ○○ ○○	○• •• ○•	•• •• ○•	•• •• ○•	•• ○• ○•	○• ○• ○•	○• •• ○•	○• •• ○•	•• •• ••	○• ○• ○•
і	ї	й	к	л	м	н	о	п	р	с
•• ○• ••	•• ○• ○•	•• •• ••	○• ○• ••	○• ○• ••	•• ○• ○•	•• ○• ••	○• ○• ○•	•• •• ○•	•• •• ○•	○• ○• ○•
т	у	ф	х	ц	ч	ш	щ	ь	ю	я
○• •• ○•	○• ○• ••	•• ○• ○•	○• •• ○•	○• ○• ○•	•• •• ○•	○• ○• ○•	○• ○• ○•	○• ○• ○•	○• ○• ○•	○• ○• ○•

Fig. 3. Braille alphabet for Ukrainian language

This paper reviews Braille system, common techniques used to read and write it by blind people; contains a brief introduction of artificial neural networks; its structure, types of learning and Backpropagation algorithm; network's training, testing and results of these both processes.

Artificial neural network (ANN) will be designed to identify Cyrillic letters' images in this research. Some of them will be noised with some type of noise to simulate somehow the real-world environment. ANN will be trained and tested to be used for identifying them.

II. OPTICAL BRAILLE CHARACTER'S RECOGNITION METHODOLOGY

Different sets of techniques used for optical Braille characters recognition system creating will be described in this section. Optical Braille characters recognition methodology is shown in Figure 4.

Optical Braille character recognition process in general can be split to next steps: 1) image acquisition; 2) image preprocessing; 3) image segmentation [2]; 4) Braille dot recognition [3]; 5) Braille cell translation [4].

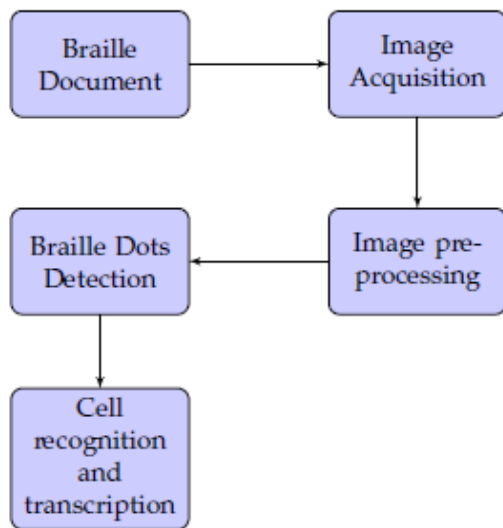


Fig. 4. General methodology

A. Image acquisition:

Braille paper was captured using a scanner to take characters' images with a resolution of 41 pixels x 60 pixels. Capturing process' results are generated as JPEG-type image files and have black and white color scale.

B. Image processing:

This method is used to prepare the picture for the next step, to make the process of Braille characters recognition easier. It includes image cropping, grayscaling, thresholding, erosion and dilation [5].

C. Image segmentation:

Braille character segmentation's area correctness can improve recognition's accuracy level. Segmentation is done by making a small segment as many as 8 areas of segmentation, which consist of 2 columns and 3 rows. Value of pixels will be read from 8 small segmentations and equal to value between 0 (black) or 255 (white). So, each small segmentation area will get one data input. Therefore, process result will be 8 data. It will play a role of data source for artificial neural network process.

III. ARTIFICIAL NEURAL NETWORKS

There was a huge growth of hybrid intelligent systems' successful using at different areas in recent years. Increased neural networks' using for pattern recognition, classification and optimization tasks has played the main contributing factor's role for the development of hybrid systems [6].

Artificial Neural Network (ANN) is computer technique designed to simulate the way of human brain to do different tasks. ANNs can do that by a lot of parallel distributed processing units – "nodes". These units are mathematical models called "neurons". They have ability to process and store information as same as biological neurons do. ANN consists of groups of interconnected artificial neurons which process the information [7].

A. ANN's structure

The most commonly used structure contains many layers – the first layer is acts as input layer, then one or more play the role of hidden layers, and, finally, there is output layer. Each of them consists of at least one or more neurons. These neurons are connected by connection line, which indicates the flow of information from one node to another and from the input layer to the output via network [8].

B. ANN's learning

ANNs have impressive features in ability to learn by adjusting weighted connections between neurons in network layers. There are different types of learning. The objective of learning process is to find a set of weight matrices which should map any input to a correct output when applied to the network. Below list of the most used learning types [9]:

1) Supervised learning:

Desired output for network is also provided with input in this type of learning while training the network. It is possible to calculate an error based on the differences between the target output and actual output of network, which is used to make corrections of network weights.

2) Unsupervised Learning:

Only the set of inputs without output is given in case of neural network's learning in this type and its responsibility is to find a pattern with inputs provided.

3) Reinforcement Learning:

Reinforcement learning is quite similar to supervised learning, because some feedback is given instead of providing of reward to target output, looking at system performance. The reinforcement learning's main goal is to maximize the reward which system receives during trial-and-error process.

C. Backpropagation algorithm

Backpropagation (BP) [9] stands for "backward propagation of errors". This algorithm is the most common training algorithm of ANNs. The method calculates the gradient of loss function with respect to all the weights in the network. The gradient is fed to the optimization method, which, in turn, uses it to update the weights in attempt to minimize error function.

Desired output for each input value must be known in order to calculate error function gradient in Backpropagation algorithm. It usually considered as learning method with supervising Backpropagation learning algorithm can be split into two phases:

1) Propagation:

First of all, training pattern's input's forward propagation through ANN in order to generate the output signal, then backward propagation of output signal through ANN using deltas (difference between targeted and actual output values) of all output and hidden neurons.

2) Weight update:

Secondly, after calculating deltas from first phase, multiply its output delta and input activation to get the gradients of weights. Weight gradients' signs indicate where the error increases – this is why the weights should be updated in the opposite direction. Phase 1 and 2 are repeated until network's performance becomes satisfactory and error is minimized.

IV. SYSTEM DESIGN

ANN will be designed to identify Cyrillic letters written in Braille representation system. Neural network should be able to classify noisy letters as well as letters without noise. Some of number images are collected from the Internet. Some of them are noisy with a type of filter to insure that neural network should be able to identify the true letters and generate an output according to each letter. Network output will be used to identify the corresponding Braille letter [10].

System includes next steps:

- Image acquisition;
- Image processing (filtering and normalization);
- Image segmentation;
- ANN training;
- ANN testing.

Image Acquisition technique plays the role of starting optical Braille character recognition phase. Input files were standardized to JPEG format. JPEG format's usage can free this procedure from using scanners etc. The size of images are 41x60 pixels. The original folder contains 33 images, representing 33 Cyrillic letters (A-Я).

Acquired Braille images are still not very good and need some improvements to be processed to the next stage of optical Braille character recognition that could be achieved using the defined set of image preprocessing techniques, e.g. converting of image to gray scale for easier future processing, because all color components will be compressed into one. Figure 4 shows an example of training images.

Matrix with fixed dots identified through gray level, according to shade ranging from black to white, is passed to matrix with normalized dots which may correspond only to one of possible three values: black, white or paper – during this operation. "Dot" itself will mean "result of optical scanning" in this paper. Noise removal algorithms were applied after converting of images to black and white scale [11]. Figure 5 shows the set of example training images.

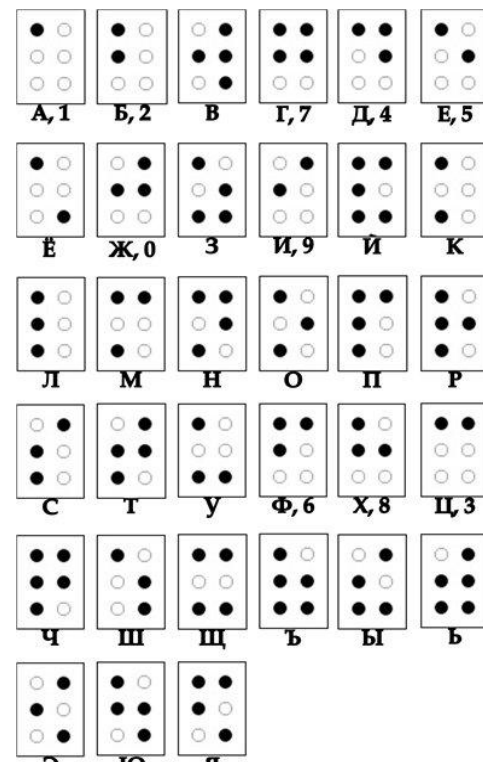


Fig. 5. Set of example training images

In testing process there was one image per each letter to test identification accuracy for neural network, as shown in Figure 6 below.



Fig. 6. Examples of testing images

Several parameters were used for artificial neural network training – like learning rate, momentum factor, minimal error, maximal iterations' number, input layer's neurons' number, hidden layers' number and output layer's neurons number, as shown in Figure 7 below [12]. Six neurons were used in the output layer to have the possibility to make binary combination of 6 bits.

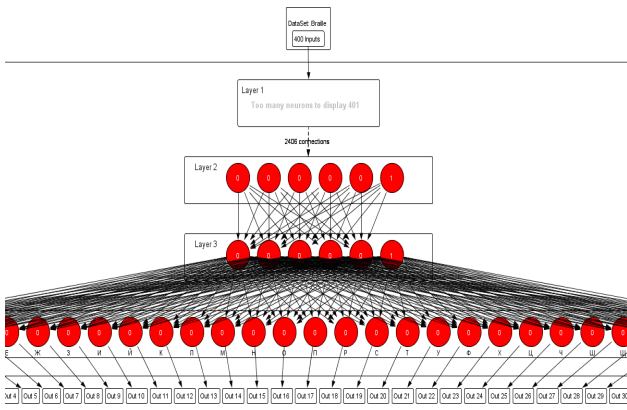


Fig. 7. Used ANN structure topology

V. RESULTS

Threshold value used to differentiate between the identified and not identified numbers was 90% in result of experiment. Artificial neural network (ANN) system designed uses Backpropagation algorithm at training stage. Accuracies of training and testing were very high. They depend on training algorithms of artificial neural network. The number of images were used in the experiment was 33 images for training and 8 images for testing per test word.

A lot of different parameters were used in the experiment to achieve the goal of the system to identify the number that was written in Braille representation. Table I below shows some parameters that were used in experiment:

TABLE I. TRAINING PARAMETERS

No.	Parameter	Value
1	Learning Rate	0,005
2	Momentum Factor	0,2
3	Minimum Error	0,0037
4	Number of Iterations	2713
5	Maximum Iterations	4000
6	Number of Hidden layers	5

Changes in any table value listed above would be affect on results of training and testing process, where these values affect training of neural network and time spent on this process and also training and testing accuracy. Figure 8 shows gradient of training which represents performance of neural network.

Identification results using training image set gained 97.1% accuracy which represented average of all letters in training; testing image set has 95% accuracy.

According to system threshold, result obtained from experiment was successful. Tables II, Table III show neural network's training and testing accuracy.

TABLE II. NEURAL NETWORK TRAINING AND TESTING ACCURACY

Phase	Matching	Percentage
Training	Accuracy	97.1%
Testing	Accuracy	95%

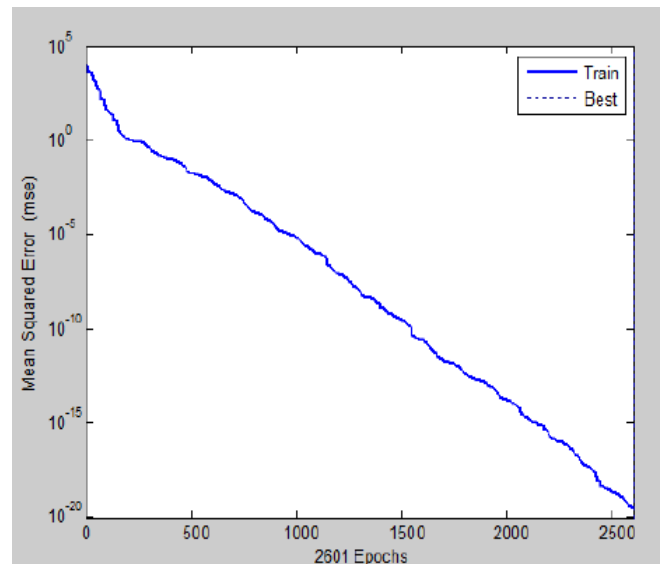


Fig. 8. Neural network performance graph

TABLE III. RESULTS OF BRAILLE RECOGNITION EXPERIMENT

No.	Input data	Error	Accuracy
1	400	0,0057	94,75 %
2	400	0,0043	94,67 %
3	400	0,0034	95,34 %
4	400	0,0039	94,77 %

System's accuracy level was determined during the testing of it on the set of Braille characters by performing the recognition on 10 different images data.

Output window produced by Braille recognizer and error network graph are shown in Figures 4, 5. The user can check the interpretation correctness manually and, therefore, can fix it even without any Braille system knowledge. Depending on scanned image file's quality, optical Braille character recognition using artificial neural networks can reach in mean about 98-100% of automatic correctness [13].

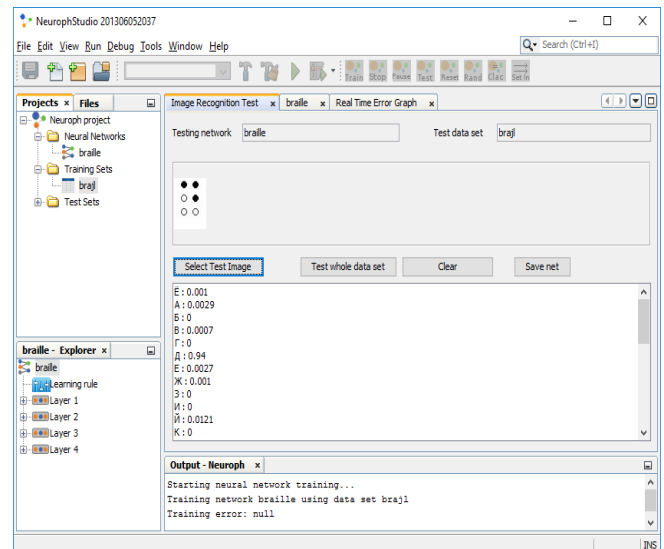


Fig. 9. Recognition step: Neuroph output window

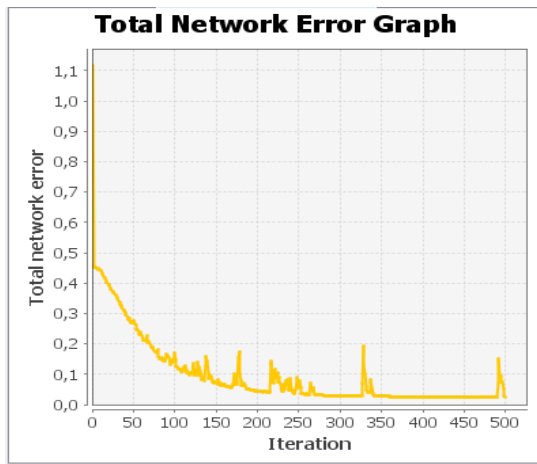


Fig. 10. Error network graph

VI. CONCLUSION

Optical Braille character recognition system using artificial neural networks (ANN) method was researched in this paper.

Optical Braille character recognition (OBCR) system is able to build the bridges between blind, visually impaired and sighted people and can convert Braille characters to natural language characters they correspond to. It can help anyone who doesn't know Braille scripting language, but needs to deal with blind people.

System uses multilayer perceptron at its basics, that was implemented using modified Backpropagation algorithm, which allows to reduce the convergence time, and shows very good performance rate.

Designing and implementation of system that identifies letters written in Braille character representation using artificial neural network; practical results has proven success of this system, where many experiments was carried out and result was very high.

According to experiment, result of the identification of characters written in Braille representation using artificial neural network the training accuracy was 95.7% and testing accuracy was 95%.

Based on the experiment, it can be concluded that using of artificial neural network in identification process is successful and very useful because of the easiest way to programming the network architecture and processing time that takes to training and testing with any number of images, it can obtain high identification rate and accuracy.

As a future works it might be to increase training and testing accuracy of ANN up to ~100%.

REFERENCES

- [1] W. H. Organization. Visual impairment and blindness. 2017. Available: <http://www.who.int/mediacentre/factsheets/fs282/en/>
- [2] A. Al-Salman, A. El-Zaart, S. Al-Salman, and A. Gumaiei, "A novel approach for Braille images segmentation," *Multimedia Computing and Systems (ICMCS)*, International Conference on. IEEE, pp. 190–195, 2012.
- [3] T. Shreekanth, and V. Udayashankara, "A review on software algorithms for optical recognition of embossed Braille characters," *International Journal of Computer Applications*, vol. 81, no.3, pp. 25–35, 2013.
- [4] A. Al-Salman, A. El-Zaart, Y. Al-Suhaibani, K. Al-Hokail, and A. A. Al-Qabbany, "An efficient Braille cells recognition," *Wireless Communications Networking and Mobile Computing (WiCOM)*, 6th International Conference on. IEEE, pp. 1–4, 2013.
- [5] M. Wajid, M. W. Abdullah, and O. Farooq, "Imprinted Braille character pattern recognition using image processing techniques," *Image Information Processing (ICIIP)*, International Conference on IEEE, pp. 1–5, 2011.
- [6] I. Aleksander, and H. Morton, *An introduction to neural computing*, 2nd edition. London: International Thomson Computer Press. 1995.
- [7] J. M. Zurada, *Introduction to Artificial Neural Networks*, 2nd edition. Published by Jaico Publishing House, India, 1996.
- [8] S. S. Haykin, *Neural Networks: A Comprehensive foundation*. Prentice Hall, 1999.
- [9] E. V. Bodyanskiy, and O. G. Rudenko. *Artificial neural networks: architectures, обучение, applications*. Kharkiv, «Teletech», 2004.
- [10] G. Morgavi, and M. Morando, "A neural network hybrid model for an optical Braille recognizer," *International Conference on Signal, Speech and Image Processing (ICOSSIP)*, 2014.
- [11] J. Li, and Y. Xiaoguang, "Optical Braille Character Recognition with Support-Vector Machine Classifier," *International Conference on Computer Application and System Modeling (ICCSM)*, 2010.
- [12] Zhang Namba, "Cellular Neural Network for Associative Memory and Its Application to Braille Image Recognition", *International Joint Conference on Neural Networks*, BC, Canada, pp. 2409 – 2414, 2006.
- [13] *Java Neural Network Framework Neuroph Manual*, v2.93, December 2017..

Mathematical Model for the Probabilistic Minutia Distribution in Biometric Fingerprint Images

Sergey Rassomakhin

V. N. Karazin Kharkiv National University
Kharkiv, Ukraine
rassomakhin@karazin.ua

Alexandr Kuznetsov

V. N. Karazin Kharkiv National University
Kharkiv, Ukraine
kuznetsov@karazin.ua

Vladimir Shlokin

V. N. Karazin Kharkiv National University
Kharkiv, Ukraine
vshlokin@ukr.net

Ivan Belozertsev

V. N. Karazin Kharkiv National University
Kharkiv, Ukraine
ivanbelozertsev.jw@gmail.com

Roman Serhiienko

National Army Academy named after
Hetman Petro Sahaidachnyi
Lviv, Ukraine

Abstract—The research involves development of a mathematical model for the probabilistic minutia distribution in biometric fingerprint images. The suggested model is based on heuristic analysis of the fingerprint scanning results with account for the nature of the potential errors.

Keywords—biometric authentication, fingerprint images, minutiae

I. INTRODUCTION

Dactyloscopy occupies a special place among the known methods of biometric authentication [1-8]. Biometric fingerprint image minutiae processing which underlines the above method, enables robust and efficient identification of individuals.

At the same time, minutiae distribution of certain implementations may be described with rather complex dependencies. This is explained by the significant differences in the number of minutiae and their placement [9-12]. Error types and their distribution functions are also rather ambiguous due to the multiple nature of possible causes.

The choice between simplicity and adequacy of the models, describing minutiae placement and errors, is a compromise option. However, the closed nature of existing fingerprint recognition algorithms makes it impossible to collect an amount of statistics enough for solving the problem in a straightforward way. Therefore, the research and development of mathematical models for the probabilistic minutiae distribution in biometric fingerprint images is an important and relevant scientific problem. The models represented in this paper were obtained through heuristic analysis of the fingerprint scanning results with account for the nature of the potential errors.

II. THE ANALYSIS OF BIOMETRIC FINGERPRINT IMAGES

Let us use database DB1_1 [13] for the analysis of characteristic and error distributions which may occur during fingerprint image processing. This database contains 8 images (files 101_1.tif – 101_8.tif) of the same fingerprint. The goal of the analysis is to make a preliminary conclusion about the nature of the errors typical for minutia recognition.

The original fingerprint images are shown on figure 1. Figure 2 shows the processing results of the given samples using SourceAFIS.FingerprintAnalysis [14-15].



Fig. 1. Biometric images of a single finger

The results represent detected minutiae which correspond to the endings and bifurcations of the ridges. The arrows represent the angles of given minutiae.

Variation of fingerprint orientations and their displacement, as well as the changes in contrast and brightness, cause the significant differences of processing results which translates to the variation of the number of minutiae and their positioning. The following figure shows the circles that correspond to the same area of the fingerprint, but displaced and rotated during the scanning process.



Fig. 2. Fingerprint minutiae extraction results

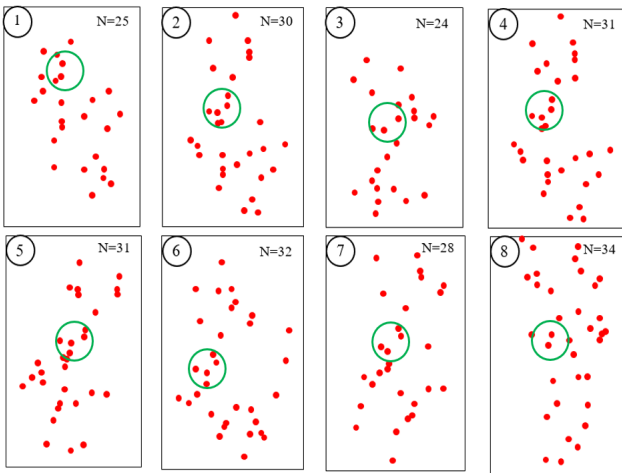


Fig. 3. Plain portraits of the extracted minutiae distribution

Figure 3 represents plane portraits of minutiae placement for the given fingerprint images. As we can see, the degree of similarity of the given portraits is rather low. Visual similarity takes place only in case of similar scanning conditions (on Fig. 3 it is the pairs 2 and 4 or 3 and 7). Apparently, the cause of the problem mentioned above is not only the complexity of the procedure itself, but also the imperfection of the used recognition algorithm implemented in SourceAFIS.FingerprintAnalysis [14-15].

III. MATHEMATICAL MODEL FOR THE PROBABILISTIC MINUTIAE DISTRIBUTION

The analysis of the portraits mentioned above indicates the following *features* that we can base our empiric choice of the type of minutiae distribution on.

- the density of distribution of points along the horizontal (X) and vertical (Y) axes is roughly uniform in the central part of the frame and slightly decreases to its edges;
- linear displacements of the center of the fingerprint horizontally and vertically do not imply the appearance of zones free of minutiae at the edges of the frame (new points may enter the scanning area);
- The distribution of minutiae angles is approximately uniform in the range of $[0, 2\pi]$.

Let us use the following assumptions to construct a model of minutiae distribution according to the features mentioned above:

- the portrait coordinates of the fingerprint X, Y , as well as the minutiae angles values are normalized in the range $[-0.5; +0.5]$, while the geometric center of the image has zero coordinates on the plane $[0; 0]$, and the portrait itself is placed in a unit square area covering all 4 quadrants of the image plane;
- for the primary generation of random numbers necessary to obtain the distribution of minutiae coordinates on the fingerprint image portraits, a uniformly distributed (continuous) random number generator in range $[0; 1]$: $f(x_i, y_i):unif[0,1]$,

$i \in 1 \dots N$, where N - the number of minutiae in the portrait, a random value that does not go out of range $[15; 60]$ with a mathematical expectation $m_N = 25 \div 35$ and unimodal distribution.

The analysis of features which were discussed before, as well as taking into account the assumptions made above, allow us to use the dependency shown on Figure 4 to describe the probability density function (PDF) $f(x)$ and $f(y)$ Cartesian coordinates of the minutiae on the plain portraits. This type of PDF provides a uniform points distribution in the central part of the unit square and the decreasing probability of point appearance at the edges of the square area of the portrait. The area of non-zero PDF values $[-0.75; +0.75]$ is 0.25 in both directions beyond the unit square, which provides a non-zero probability of the point appearance in the border areas of the portrait. The errors appear in the form of possible geometric center drifting. The choice of this PDF is, of course, not the only one possible, however, in our opinion, is acceptable, considering the tradeoff between the simplicity and features mentioned above.

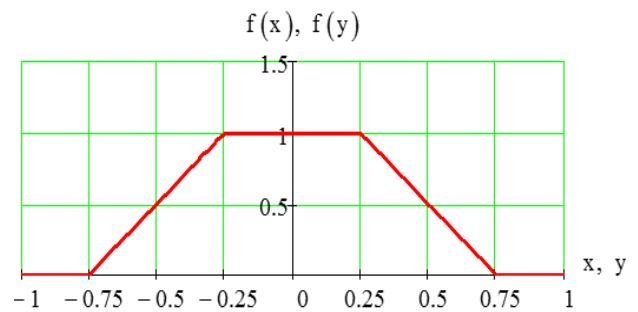


Fig. 4. Probability density function of the minutiae coordinates

To obtain test samples of minutia placement portraits the generator of random numbers, distributed according to PDF $f(x)$, $f(y)$ (Fig. 4), is required. Considering the identity of the distributions along the coordinates of the plane when using the normalized unit square of the portrait, in the following we shall consider only the function $f(x)$:

$$f(x) = \begin{cases} 2x + 1.5 & \text{if } -0.75 \leq x < -0.25; \\ 1 & \text{if } -0.25 \leq x \leq 0.25; \\ -2x + 1.5 & \text{if } 0.25 < x \leq 0.75; \\ 0 & \text{if } |x| > 0.75. \end{cases} \quad (1)$$

To generate a random variable subject to distribution (1), one can use the functional result transformation of the standard for most programming systems of a random number generator located continuously uniformly in the range $[0, 1]$.

We use the inverse function method: if $z:unif[0,1]$ then the random variable x obtained by a functional transformation z in the form of

$$x = \begin{cases} \sqrt{z} - 0.75 & \text{if } 0 \leq z < 0.25; \\ z - 0.5 & \text{if } 0.25 \leq z \leq 0.75; \\ -\sqrt{1-z} + 0.75 & \text{if } 0.75 < z \leq 1; \end{cases} \quad (2)$$

will have a PDF (1).

Figure 5 shows the histogram of the statistical tests of the functional transformation (2) from $unif[0,1]$ the number of trials equal to 30,000 and dividing the interval $[-0.75, +0.75]$ into 100 equal subintervals. The dashed line in Fig. 5 shows the envelope (1).

The resulting algorithm for random number generation will be used later to obtain the coordinates of the characteristic points of the normalized square fingerprint portraits.

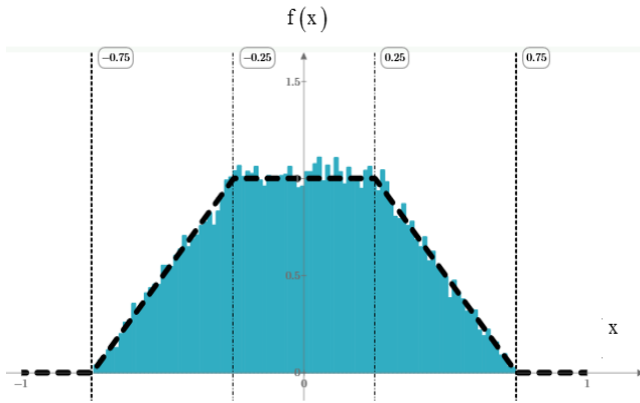


Fig. 5. Result of statistical tests of the random coordinate sensor

Let us choose the use of a discrete (integer) random value N from the range of integers $[15, 45]$ with a discrete normal truncated distribution and the following numerical characteristics to generate a random variable – number of minutiae in a fingerprint portrait sample:

- mathematical expectation $m_N \approx 30$;
- standard deviation $\sigma \approx 2 \div 5$.

Let us again use the functional transformation data of the generator $unif[0,1]$ to obtain a random number of minutiae on a fingerprint portrait sample . We simulate the samples of a random variable N based on the central limit theorem. We proceed to the discrete form of uniformly distributed numbers using the operation of integer rounding and centering:

$$z' = round(z) - 0.5, \text{ where } z:unif[0,1]. \quad (3)$$

Then, limiting the number of terms to $m_N = 30$, the random number of minutiae in the portrait can be determined as the sum

$$N = \sum_{i=1}^{30} z' + 30. \quad (4)$$

A discrete random variable N can take integer values from a range $[15, 45]$. The truncated normal function of the PDF of this random variable is approximated by weighted binomial coefficients:

$$Q(N_i) = \binom{i}{30} \cdot \left(\frac{1}{2}\right)^{30}, i \in [0, 30], N_i \in [15, 45], \quad (5)$$

where $Q(N_i)$ is the probability that the number of minutiae on a portrait (taking into account points masked outside the unit square) will be a value N_i .

The form and numerical characteristics of the distribution (5) are shown in Fig. 6.

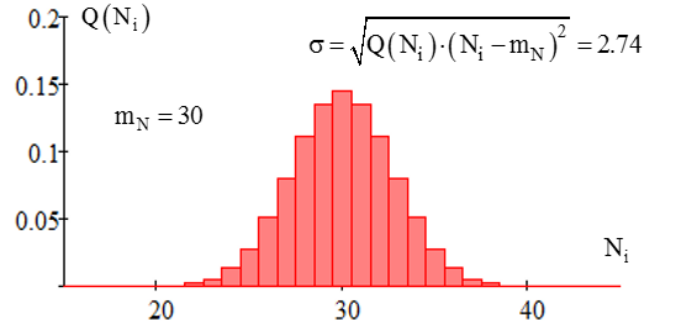


Fig. 6. PDF of the number of minutiae on a normalized fingerprint portrait

To simulate the random values of the minutiae angles normalized in the unit square, it is expedient to use a random variable $z:unif[0,1]$ uniformly distributed over a unit interval:

$$\varphi = z. \quad (6)$$

The true minutia angle is determined on the basis of the normalized value (6): $\Phi = 2\pi \cdot \varphi$.

Table I presents the results of modeling a normalized portrait based on the distributions (1), (4), and (6).

The highlighted rows in Table I correspond to points that did not fall into a unit square. Therefore, in spite of the fact that during the experiment we obtained $N = 28$, only 21 points were found in the unit square (Fig. 7). "Masked" points can appear in case the shifts along the axes X and Y or the image rotation occur.

In case of necessity, it is possible to consider a three-dimensional space for point placing by adding a third coordinate for the normalized angle φ in the corresponding processing algorithm.

IV. CONCLUSION

The analysis of the fingerprint scanning results shows an extremely small degree of similarity among the obtained images. Considering the nature of the possible errors, visual similarity takes place only in case of similar scanning conditions. The cause of the problem mentioned above is not only the complexity of the procedure itself, but also the imperfection of the used recognition algorithms. The

research involves development of a mathematical model for the probabilistic minutiae distribution in biometric fingerprint images. The suggested model is based on heuristic analysis of the fingerprint scanning results with account for the nature of the potential errors. As the result, we were able to model a typical minutiae behavior in the biometric fingerprint images. This research might be useful for the improvement of various biometric methods of information security, as well as other practical use [16-54].

TABLE I. PORTRAIT MATRIX

№	X	Y	φ
1	0.21	-0.07	0.6
2	0.28	0.18	0.58
3	0.12	-0	0.49
4	0.19	0.31	0.74
5	0.07	-0.68	0.62
6	0.05	-0.12	0.8
7	-0.25	0.33	0.58
8	-0.01	0.42	0.91
9	0.17	0.15	0.73
10	0.12	0.23	0.67
11	0.3	0.54	0.32
12	0.64	-0.14	0.31
13	0.13	0.44	0.11
14	0.09	-0.05	0.85
15	0.01	0.39	0.15
16	-0.05	-0.55	0.08
17	0.51	0.5	0.64
18	-0.25	-0.34	0.55
19	0.23	-0.41	0.41
20	0.13	-0.41	0.47
21	0.14	0.08	0.15
22	0.06	0.23	0.74
23	-0.05	0.17	0.83
24	0.69	0.31	0.87
25	0.1	0.7	0.3
26	-0.33	-0.3	0.13
27	-0.17	-0.11	0.78
28	0.15	0.08	0.61

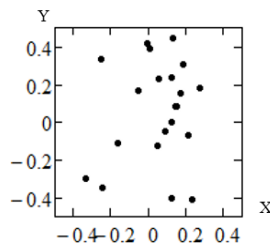


Fig. 7. Random points distribution sample

REFERENCES

[1] Xudong Jiang and Wei-Yun Yau, "Fingerprint minutiae matching based on the local and global structures," 15th Int. Conf. on Pattern Recognition. ICPR-2000, Barcelona, vol. 2, pp. 1038-1041, 2000.

[2] D. Maltoni, D. Maio, A. K. Jain, and S. Prabhakar, Handbook of Fingerprint Recognition, Springer, New York, 2003.

[3] "Privacy Enhancing Technologies for Biometric Data", 2015. [Online]. Internet: <http://www.cs.haifa.ac.il/~orrd/PrivDay/2015/>

[4] "Privacy Enhancing Technologies for Biometric Data", 2016. [Online]. Internet: <http://www.cs.haifa.ac.il/~orrd/PrivDay/>

[5] ISO/IEC 19794-3. Information technology – Biometric data interchange formats – Part 3: Finger pattern spectral data.

[6] ISO/IEC 19794-2. Information technology – Biometric data interchange formats – Part 2: Finger minutiae data.

[7] ISO/IEC 19794-4. Information technology – Biometric data interchange formats – Part 4: Finger image data.

[8] N. K. Ratha, K. Karu, Shaoyun Chen and A. K. Jain, "A real-time matching system for large fingerprint databases," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 18, no. 8, pp. 799-813, Aug 1996.

[9] I. Craw, N. P. Costen, T. Kato, and S. Akamatsu, "How should we represent faces for automatic recognition?," IEEE Trans. Pat. Anal. Mach. Intel., vol. 21, pp. 725-736, 1999

[10] B. A. Draper, K. Baek, M. S Bartlett, and J. R. Beveridge, "Recognizing faces with PCA and ICA", Computer Vision and Image Understanding, vol.91, pp. 115-137, 2003.

[11] ISO/IEC 19794-5. Information technology – Biometric data interchange formats – Part 5: Face image data.

[12] C. Xiang, X. A. Fan, and T. H. Lee. "Face recognition using recursive Fisher linear discriminant." Communications, Circuits and Systems.– vol.2., pp. 27-29, 2004.

[13] "FVC2004. Fingerprint Verification Competition. Databases". [Online]. Internet: <http://bias.csr.unibo.it/fvc2004/databases.asp>

[14] "SourceAFIS for Java and .NET". [Online]. Internet: <https://sourceafis.machinezoo.com/>

[15] "SourceAFIS Fingerprint recognition library for .NET and experimentally for Java". [Online]. Internet: <https://sourceforge.net/projects/sourceafis/>

[16] T. Lavrovskaya and S. Rassomahin, "Physical model of pseudorandom codes in multidimensional Euclidean space," Third Int. Scien.-Pract. Conf. Problems of Infocommunications Science and Technology (PIC S&T), Kharkiv, pp. 67-70, 2016.

[17] G. Aggarwal, N. K. Ratha and R. M. Bolle, "Biometric Verification: Looking Beyond Raw Similarity Scores," Conf. on Computer Vision and Pattern Recognition Workshop (CVPRW'06), pp.31-31, 2006.

[18] A. A. Kuznetsov, A. A. Smirnov, D.A. Danilenko, and A. Berezovsky. "The statistical analysis of a network traffic for the intrusion detection and prevention systems." Telecommunications and Radio Engineering, vol. 74, pp. 61-78, 2015.

[19] Jinyu Zuo, N. K. Ratha and J. H. Connell, "A new approach for iris segmentation," IEEE Computer Society Conf. on Computer Vision and Pattern Recognition Workshops, Anchorage, AK, pp.1-6, 2008.

[20] Yu.V. Stasev, and A. A. Kuznetsov. "Asymmetric code-theoretical schemes constructed with the use of algebraic geometric codes," Kibernetika i Sistemnyi Analiz, no. 3, pp. 47-57, May-June 2005.

[21] S. Biswas, N. K. Ratha, G. Aggarwal and J. Connell, "Exploring Ridge Curvature for Fingerprint Indexing," IEEE Second Int. Conf. on Biometrics: Theory, Applications and Systems, Arlington, VA, pp. 1-6, 2008.

[22] O. Kuznetsov, Y. Gorbenko and I. Kolovanova, "Combinatorial properties of block symmetric ciphers key schedule," Third Int. Scien.-Pract. Conf. Problems of Infocommunications Science and Technology (PIC S&T), Kharkiv, pp. 55-58, 2016.

[23] A. Kuznetsov, I. Kolovanova and T. Kuznetsova, "Periodic characteristics of output feedback encryption mode," 4th Int. Scient.-Pract. Conf. Problems of Infocommunications. Science and Technology (PIC S&T), Kharkov, pp. 193-198, 2017.

[24] I. Gorbenko, A. Kuznetsov, M. Lutsenko and D. Ivanenko, "The research of modern stream ciphers," 4th Int. Scient.-Pract. Conf. Problems of Infocommunications. Science and Technology (PIC S&T), Kharkov, pp. 207-210, 2017.

[25] X. Lin and S. Chen, "Color image segmentation using modified HSI system for road following," IEEE International Conference on Robotics and Automation, Sacramento, CA, vol. 3, pp. 1998-2003, 1991.

[26] A. Yanko, S. Koshman and V. Krasnobayev, "Algorithms of data processing in the residual classes system," 4th Int. Scient.-Pract. Conf. Problems of Infocommunications. Science and Technology (PIC S&T), Kharkov, pp. 117-121, 2017.

[27] V. M. Grachev, V. I. Esin, N. G. Polukhina, S. G. Rassomakhin. "Technology for developing databases of information systems." Bulletin of the Lebedev Physics Institute, vol. 41(5), pp. 119-122, May 2014.

[28] A. Kuznetsov, I. Svatovskij, N. Kiyan and A. Pushkar'ov, "Code-based public-key cryptosystems for the post-quantum period," 2017 4th Int. Scient.-Pract. Conf. Problems of Infocommunications. Science and Technology (PIC S&T), Kharkov, pp. 125-130, 2017.

[29] O. Potii, O. Illiashenko, and D. Komin. "Advanced Security Assurance Case Based on ISO/IEC 15408," Theory and Engineering

- of Complex Systems and Dependability Advances in Intelligent Systems and Computing, vol. 365, pp. 391-401, 2015.
- [30] S. Prabhakar, A. K. Jain, Jianguo Wang, S. Pankanti and R. Bolle, "Minutia verification and classification for fingerprint matching," Proceedings 15th International Conference on Pattern Recognition. ICPR-2000, Barcelona, vol.1, pp. 25-29, 2000.
- [31] A. V. Potii, A. K. Pesterev, "A System Approach to Certification of Pseudorandom Numbers Generators Used in Information Protection Systems," Telecommunications and Radio Engineering, vol. 52, iss. 4, pp 97-102, 1998.
- [32] V. A. Krasnobayev, S. A. Koshman, and M. A. Mavrina, "A Method for Increasing the Reliability of Verification of Data Represented in a Residue Number System" Cybernetics and Systems Analysis, vol. 50, iss. 6, pp. 969-976, Nov. 2014.
- [33] I. D. Gorbenko, A. A. Zamula, and Ye. A. Semenko, "Ensemble and correlation properties of cryptographic signals for telecommunication system and network applications." Telecommunications and Radio Engineering, vol. 75, iss. 2, pp. 169-178, 2016.
- [34] John Juyang Weng and Shaoyun Chen, "Incremental learning for vision-based navigation," Proceedings of 13th Int. Conf. on Pattern Recognition, Vienna, vol. 4, pp. 45-49, 1996.
- [35] I. Gorbenko, and V. Ponomar. "Examining a possibility to use and the benefits of post-quantum algorithms dependent on the conditions of their application." EasternEuropean Journal of Enterprise Technologies, vol 2, no 9 (86), pp. 21-32, 2017.
- [36] I. Gorbenko, and R. Hanzia, "Examination and implementation of the fast method for computing the order of elliptic curve," EasternEuropean Journal of Enterprise Technologies, vol 2, no 9 (86), pp. 11-21, 2017.
- [37] N. Sun, N. Haas, J. H. Connell and S. Pankanti, "A model-based sampling and sample synthesis method for auto identification in computer vision," Fourth IEEE Workshop on Automatic Identification Advanced Technologies (AutoID'05), pp. 160-165, 2005.
- [38] I. Gorbenko, M. Yesina, and V. Ponomar. "Anonymous electronic signature method," Third Int. Scient.-Pract. Conf. Problems of Infocommunications Science and Technology (PIC S&T), Kharkiv, pp. 47-50, 2016.
- [39] S. Kavun. "Conceptual fundamentals of a theory of mathematical interpretation." Int. J. Computing Science and Mathematics, vol. 6, no. 2, pp. 107-121, 2015.
- [40] S. Pankanti, S. Prabhakar and A. K. Jain, "On the individuality fingerprints," IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001, vol. 1, pp. I-805 - I-812, 2001
- [41] Y. Gorbenko, I. Svatovskiy, and O. Shevtsov, "Post-quantum message authentication cryptography based on error-correcting codes," Third Int. Scient.-Pract. Conf. nce Problems of Infocommunications Science and Technology (PIC S&T), Kharkiv, pp. 51-54, 2016.
- [42] O. Kazymyrov, R. Oliynykov, and H. Raddum. "Influence of addition modulo $2n$ on algebraic attacks." Cryptography and Communications, vol. 8, iss. 2, pp. 277-289, April 2016.
- [43] A. Kuznetsov, R. Serhiienko and D. Prokopovych-Tkachenko, "Construction of cascade codes in the frequency domain," 4th Int. Scient.-Pract. Conf. Problems of Infocommunications. Science and Technology (PIC S&T), Kharkov, pp. 131-136, 2017.
- [44] M. Rodinko, R.Oliynykov, and Y. Gorbenko, "Improvement of the high nonlinear S-boxes generation method," Third Int. Scient.-Pract. Conf. Problems of Infocommunications Science and Technology (PIC S&T), Kharkiv, pp. 63-66, 2016.
- [45] J. J. Weng and Shaoyun Chen, "Autonomous navigation through case-based learning," Proceedings of International Symposium on Computer Vision - ISCV, Coral Gables, FL, pp. 359-364, 1995.
- [46] A. Kuznetsov, Y. Gorbenko, A. Andrushkevych and I. Belozershev, "Analysis of block symmetric algorithms from international standard of lightweight cryptography ISO/IEC 29192-2," 4th Int. Scient.-Pract. Conf. Problems of Infocommunications. Science and Technology (PIC S&T), Kharkov, pp. 203-206, 2017.
- [47] V. M. Grachev, V. I. Esin, N. G. Polukhina, and S. G. Rassomakhin. "Data security mechanisms implemented in the database with universal model." Bulletin of the Lebedev Physics Institute, vol. 41, Issue 5, pp. 123-126, May 2014.
- [48] V. A. Krasnobayev, A. S. Yanko, and S. A. Koshman. "A Method for arithmetic comparison of data represented in a residue number system." Cybernetics and Systems Analysis, vol. 52, iss. 1, pp. 145-150, Jan. 2016.
- [49] O. Kuznetsov, M. Lutsenko and D. Ivanenko, "Strumok stream cipher: Specification and basic properties," Third Int. Scient.-Pract. Conf. Problems of Infocommunications Science and Technology (PIC S&T), Kharkiv, pp. 59-62, 2016.
- [50] A. Andrushkevych, T. Kuznetsova, I .Bilozertsev, and S. Bohucharskyi, "The block symmetric ciphers in the post-quantum period," Third Int. Scient.-Pract. Conf. Problems of Infocommunications Science and Technology (PIC S&T), Kharkiv, pp. 43-46, 2016.
- [51] M. Rodinko and R. Oliynykov, "Open problems of proving security of ARX-based ciphers to differential cryptanalysis," 4th Int. Scient.-Pract. Conf. Problems of Infocommunications. Science and Technology (PIC S&T), Kharkov, pp. 228-231, 2017.
- [52] V.I. Dolgov, I.V.Lisitska, and K.Ye. Lisitskiy. "The new concept of block symmetric ciphers design," Telecommunications and Radio Engineering, vol. 76, iss. 2, pp. 157-184, 2017.
- [53] I. D. Gorbenko, V. I. Dolgov, V. I. Rublinetskii, and K. V. Korovkin. "Methods of Information Protection in Communications Systems and Methods of Their Cryptoanalysis." Telecommunications and Radio Engineering, vol. 52, iss. 4, pp. 89-96, 1998.
- [54] A. Nagar, H. Choi and A. K. Jain, "Evidential Value of Automated Latent Fingerprint Comparison: An Empirical Approach," IEEE Transactions on Information Forensics and Security, vol. 7, no. 6, pp. 1752-1765, Dec. 2012.

Deep 2D-Neural Network and its Fast Learning

Yevgeniy Bodyanskiy
Control Systems Research Laboratory
Kharkiv National University of Radio Electronics
Kharkiv, Ukraine
yevgeniy.bodyanskiy@nure.ua

Daria Kopaliani
beat.no
Norway
daria.kopaliani@gmail.com

Iryna Pliss
Control Systems Research Laboratory
Kharkiv National University of Radio Electronics
Kharkiv, Ukraine
iryna.pliss@nure.ua

Olena Boiko
Control Systems Research Laboratory
Kharkiv National University of Radio Electronics
Kharkiv, Ukraine
olena.boiko@nure.ua

Abstract—A deep 2D-neural network and its learning algorithm are proposed. This system is based on the 2D analogue of the elementary Rosenblatt's perceptron, the error backpropagation procedure, and the matrix analogue of the Kaczmarz–Widrow–Hoff algorithm. The distinguishing feature of the proposed system is that it saves information between the rows of the processed 2D observations.

Keywords—deep learning, artificial neural networks, image processing, 2D-neural networks, multilayer networks

I. INTRODUCTION

Nowadays artificial neural networks (ANN) have found their application in many domains, connected with information processing due to their approximation capabilities and ability to learning by tuning either both their synaptic weights and their architecture. In the recent years, the attention of researchers is attached to the deep neural networks (DNN) [1–3] that have significantly improved approximation capabilities as compared with traditional shallow ANN. The payment for this is the complication of the architecture due to the need for significantly larger number of hidden layers and increasing of the learning time. Wherein the nodes of the shallow and the deep ANN are usually Rosenblatt's perceptrons with sigmoidal and piecewise activation functions. The learning process is based on the error backpropagation with gradient or quasi-Newton procedures.

In the most cases the information for processing is presented as a sequence of vector observations and the output signal also has the form of a vector. Since the most developed class of DNN – convolutional neural networks – is used for image processing, the input image after convolution and pooling is needed to be presented as a vector that is further processed by a sequence of layers, that is in general a multilayer perceptron [4]. At the network output the processed vector signal is devectorized, i.e. transformed to the form of image matrix.

Such transformations lead to an increasing of the number of tunable synaptic weights, a decrease in the speed of the learning process, a significant loss of information contained between the rows of the processed image.

In that regard, studies have appeared related to two-dimensional matrix neural networks [5-8], whose input and output signals have a matrix form.

It should be noted that the learning process of 2D matrix ANNs seems too cumbersome from the computational point of view and it is characterized by a low learning speed, because it is based on gradient procedures with a constant learning rate parameter.

In that regard, it seems expedient to develop an architecture and learning algorithm for 2D-DNN, characterized by the simplicity of computational implementation and high (within the optimal one in the gradient procedures class) rate its parameters adjustment.

II. ADAPTIVE 2D-MODEL AND ITS LEARNING ALGORITHM

To describe 2D-signals to be processed, it is convenient to use a matrix bilinear form [9, 10]

$$Y(k) = AX(k)B \quad (1)$$

where $X(k)$, $Y(k)$ are $(n_1 \times n_2)$ initial and processed matrix signals at a discrete point in time $k=1,2,\dots$, A , B are $(n_1 \times n_1)$, $(n_2 \times n_2)$ -parameter matrices to be estimated.

To find the transformation parameters (1), an adaptive 2D model is introduced [11, 12] in the form

$$\hat{Y}(k) = A(k-1)X(k)B(k-1), \quad (2)$$

whose parameter matrices $A(k-1)$, $B(k-1)$ are tuned in online learning process (identification process).

In principle, the description (1) and model (2) can be rewritten in the traditional vector form

$$\begin{cases} \bar{Y}(k) = (B^T \otimes A) \bar{X}(k) = C \bar{X}(k), \\ \tilde{Y}(k) = (B^T(k-1) \otimes A(k-1)) \bar{X}(k) = C(k-1) \bar{X}(k) \end{cases} \quad (3)$$

where $\bar{Y}(k)$, $\tilde{Y}(k)$, $\bar{X}(k)$ are $(n_1 n_2 \times 1)$ -vectors, \otimes – the symbol of the tensor product (the Kronecker product). However, the use of the representation (3) is inexpedient for two reasons. First, the dimension of the parameters matrix C

being evaluated increases sharply, since with $n_1 > 2$, $n_2 > 2$ it turns out that $(n_1 n_2)^2 > n_1^2 + n_2^2$. Second, on each iteration k there is a need to solve a system of bilinear equations

$$B^T(k-1) \otimes A(k-1) = C(k-1),$$

which in the general case is uniquely unresolvable. It should be noticed, that $(n_1 n_2)^2$ is the number of matrix parameters C to be evaluated, $n_1^2 + n_2^2$ is the number of matrix parameters A and B to be evaluated.

To organize the process of adaptation (training), three types of errors that arise during the learning process are introduced:

$$\begin{cases} E(k) = Y(k) - A(k-1)X(k)B(k-1) = Y(k) - \hat{Y}(k), \\ E_A(k) = Y(k) - A(k)X(k)B(k-1) = Y(k) - \hat{Y}_A(k), \\ E_B(k) = Y(k) - A(k)X(k)B(k) = Y(k) - \hat{Y}_B(k) \end{cases}$$

and two learning criteria

$$\begin{cases} E_A^*(k) = \frac{1}{2} \text{Tr} E(k) E^T(k), \\ E_B^*(k) = \frac{1}{2} \text{Tr} E_A(k) E_A^T(k). \end{cases} \quad (4)$$

The minimization of the criteria (4) based on the gradient optimization procedure leads to the parameters adaptation algorithm

$$\begin{cases} A(k) = A(k-1) + \eta_A(k) E(k) B^T(k-1) X^T(k), \\ B(k) = B(k-1) + \eta_B(k) X^T(k-1) A^T(k) E_A(k) \end{cases} \quad (5)$$

where $\eta_A(k)$, $\eta_B(k)$ are learning rate parameters.

The optimization of algorithm (5) by speed leads to the result [11]

$$\begin{cases} A(k) = A(k-1) + (\text{Tr} E(k) B^T(k-1) X^T(k) X(k) \times \\ \quad \times B(k-1) E^T(k)) (\text{Tr} E(k) B^T(k-1) X^T(k) X(k) \times \\ \quad \times B(k-1) B^T(k-1) X^T(k) X(k) B(k-1) E^T(k))^{-1} \times \\ \quad \times E(k) B^T(k-1) X^T(k), \\ B(k) = B(k-1) + (\text{Tr} E_A(k) A(k) X(k) X^T(k) \times \\ \quad \times A^T(k) E_A(k)) (\text{Tr} A(k) X(k) X^T(k) A^T(k) E_A(k) \times \\ \quad \times E_A^T(k) A(k) X(k) X^T(k) A^T(k))^{-1} \\ \quad \times X^T(k) A^T(k) E_A(k), \end{cases} \quad (6)$$

which is the generalization of the Kaczmarz-Widrow-Hoff algorithm for the matrix case.

It is possible to give the algorithm (6) additional filtering properties by introducing its modification in the form (5) with

$$\begin{cases} \eta_A^{-1}(k) = r_A(k) = \beta r_A(k-1) + \text{Tr} (E(k) B^T(k-1) \times \\ \quad \times X^T(k) X(k) B(k-1) B^T(k-1) X^T(k) X(k) \times \\ \quad \times B(k-1) E^T(k)), \\ \eta_B^{-1}(k) = r_B(k) = \beta r_B(k-1) + \text{Tr} (A(k) X(k) X^T(k) \times \\ \quad \times A^T(k) E_A(k) E_A^T(k) A(k) X(k) X^T(k) A^T(k)), \end{cases}$$

where $0 \leq \beta \leq 1$ is the forgetting factor.

III. 2D ROSENBLATT'S PERCEPTRON

Adaptive 2D-model (2) can serve as a basis for synthesizing a matrix multilayer perceptron, while a nonlinear transformation realized by a single layer can be written in the form

$$\begin{aligned} \hat{Y}(k) &= \{\hat{y}_{j_1 j_2}(k)\} = \Psi \odot (A(k-1) X(k) B(k-1)), \\ j_1 &= 1, 2, \dots, n_1; j_2 = 1, 2, \dots, n_2 \end{aligned}$$

where $(n_1 \times n_2)$ -matrix of nonlinear activation functions Ψ transforms the outputs of the model (2) elementwise so that

$$\hat{y}_{j_1 j_2}(k) = \psi(A_{j_1 j_2}(k-1) X(k) B_{j_1 j_2}(k-1)) \quad (7)$$

where $A_{j_1 j_2}(k-1)$, $B_{j_1 j_2}(k-1)$ are $(1 \times n_1)$ and $(n_2 \times 1)$ - vectors of synaptic weights correspondingly.

Eq. (7) is the generalization of the elementary Rosenblatt's perceptron [13] for the case under consideration.

Introducing three types of errors similarly to the above

$$\begin{cases} e_{j_1 j_2}(k) = y_{j_1 j_2}(k) - \psi(A_{j_1 j_2}(k-1) X(k) B_{j_1 j_2}(k-1)) = \\ \quad = y_{j_1 j_2}(k) - \psi(u_{j_1 j_2}(k)) = y_{j_1 j_2}(k) - \hat{y}_{j_1 j_2}(k), \\ e_{A_{j_1 j_2}}(k) = y_{j_1 j_2}(k) - \psi(A_{j_1 j_2}(k) X(k) B_{j_1 j_2}(k-1)) = \\ \quad = y_{j_1 j_2}(k) - \psi(u_{A_{j_1 j_2}}(k)) = y_{j_1 j_2}(k) - \hat{y}_{A_{j_1 j_2}}(k), \\ e_{B_{j_1 j_2}}(k) = y_{j_1 j_2}(k) - \psi(A_{j_1 j_2}(k) X(k) B_{j_1 j_2}(k)) = \\ \quad = y_{j_1 j_2}(k) - \psi(u_{B_{j_1 j_2}}(k)) = y_{j_1 j_2}(k) - \hat{y}_{B_{j_1 j_2}}(k) \end{cases}$$

and two learning criteria

$$\begin{cases} E_{A_{j_1 j_2}}^*(k) = \frac{1}{2} e_{j_1 j_2}^2(k), \\ E_{B_{j_1 j_2}}^*(k) = \frac{1}{2} e_{A_{j_1 j_2}}^2(k), \end{cases}$$

it is easy to write the algorithm for tuning the synaptic weights of the 2D-Rosenblatt's perceptron in the form

$$\begin{cases} A_{j_1 j_2}(k) = A_{j_1 j_2}(k-1) + \eta_A(k) e_{j_1 j_2}(k) \times \\ \quad \times \psi'(u_{j_1 j_2}(k)) B_{j_1 j_2}^T(k-1) X^T(k), \\ B_{j_1 j_2}(k) = B_{j_1 j_2}(k-1) + \eta_B(k) e_{A j_1 j_2}(k) \times \\ \quad \times \psi'(u_{A j_1 j_2}(k)) X^T(k) A_{j_1 j_2}^T(k). \end{cases} \quad (8)$$

Introducing into consideration δ -errors

$$\begin{cases} \delta_{j_1 j_2}(k) = e_{j_1 j_2}(k-1) \psi'(u_{j_1 j_2}(k)), \\ \delta_{A j_1 j_2}(k) = e_{A j_1 j_2}(k-1) \psi'(u_{A j_1 j_2}(k)), \end{cases}$$

it is possible to get final δ -learning rule

$$\begin{cases} \Delta A_{j_1 j_2}(k) = \eta_A(k) \delta_{j_1 j_2}(k) B_{j_1 j_2}^T(k-1) X^T(k), \\ \Delta B_{j_1 j_2}(k) = \eta_B(k) \delta_{A j_1 j_2}(k) X^T(k) A_{j_1 j_2}^T(k) \end{cases}$$

where

$$\begin{cases} \Delta A_{j_1 j_2}(k) = A_{j_1 j_2}(k) - A_{j_1 j_2}(k-1), \\ \Delta B_{j_1 j_2}(k) = B_{j_1 j_2}(k) - B_{j_1 j_2}(k-1). \end{cases}$$

It should be noted here that equation (7), unlike the conventional Rosenblatt's perceptron, does not contain a threshold component, which, naturally, reduces the functionality of the considered system.

It is possible to eliminate this disadvantage by introducing a $((n_1 + 1) \times (n_2 + 1))$ matrix of input signals

$$\tilde{X}(k) = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & x_{11}(k) & \cdots & x_{1n_2}(k) \\ \vdots & \vdots & \ddots & \vdots \\ 0 & x_{n_1 1}(k) & \cdots & x_{n_1 n_2}(k) \end{pmatrix}$$

and synaptic weights vectors

$$\begin{cases} A_{j_1 j_2}(k) = (a_{j_1 j_2 0}(k), a_{j_1 j_2 1}(k), \dots, a_{j_1 j_2 n_1}(k)), \\ B_{j_1 j_2}(k) = (b_{j_1 j_2 0}(k), b_{j_1 j_2 1}(k), \dots, b_{j_1 j_2 n_2}(k))^T \end{cases}$$

of dimensionality $(1 \times (n_1 + 1))$ and $((n_2 + 1) \times 1)$ correspondingly.

Then the relations (8) can be rewritten in the form

$$\begin{cases} \Delta A_{j_1 j_2}(k) = \eta_A(k) e_{j_1 j_2}(k) \psi'(u_{j_1 j_2}(k)) B_{j_1 j_2}^T(k-1) \tilde{X}^T(k), \\ \Delta B_{j_1 j_2}(k) = \eta_B(k) e_{A j_1 j_2}(k) \psi'(u_{A j_1 j_2}(k)) \tilde{X}^T(k) A_{j_1 j_2}^T(k) \end{cases} \quad (9)$$

or elementwise

$$\begin{cases} \Delta a_{j_1 j_2 i_1}(k) = \eta_A(k) e_{j_1 j_2}(k) \psi'(u_{j_1 j_2}(k)) \times \\ \quad \times \sum_{i_2=0}^{n_2} b_{j_1 j_2 i_2}(k-1) x_{i_1 i_2}(k), \\ \Delta b_{j_1 j_2 i_2}(k) = \eta_B(k) e_{A j_1 j_2}(k) \psi'(u_{A j_1 j_2}(k)) \times \\ \quad \times \sum_{i_1=0}^{n_1} a_{j_1 j_2 i_1}(k) x_{i_1 i_2}(k). \end{cases} \quad (10)$$

By introducing the auxiliary notations

$$\begin{cases} \sum_{i_2=0}^{n_2} b_{j_1 j_2 i_2}(k-1) x_{i_1 i_2}(k) = \hat{x}_{i_1}(k), \\ \sum_{i_1=0}^{n_1} a_{j_1 j_2 i_1}(k) x_{i_1 i_2}(k) = \hat{x}_{i_2}(k), \end{cases}$$

it is possible to write the learning algorithm in a compact form

$$\begin{cases} \Delta a_{j_1 j_2 i_1}(k) = \eta_A(k) e_{j_1 j_2}(k) \psi'(u_{j_1 j_2}(k)) \hat{x}_{i_1}(k), \\ \Delta b_{j_1 j_2 i_2}(k) = \eta_B(k) e_{A j_1 j_2}(k) \psi'(u_{A j_1 j_2}(k)) \hat{x}_{i_2}(k). \end{cases} \quad (11)$$

If the traditional hyperbolic tangent function

$$\psi(u(k)) = \tanh u(k)$$

is used as activation functions $\psi(u(k))$, and equations

$$\begin{cases} E_{A j_1 j_2}^*(k) = \frac{1}{2} (1 + y_{j_1 j_2}) \ln \frac{1 + y_{j_1 j_2}(k)}{1 + \hat{y}_{j_1 j_2}(k)} + \\ \quad + \frac{1}{2} (1 - y_{j_1 j_2}) \ln \frac{1 - y_{j_1 j_2}(k)}{1 - \hat{y}_{j_1 j_2}(k)}, \\ E_{B j_1 j_2}^*(k) = \frac{1}{2} (1 + y_{j_1 j_2}) \ln \frac{1 + y_{j_1 j_2}(k)}{1 + \hat{y}_{A j_1 j_2}(k)} + \\ \quad + \frac{1}{2} (1 - y_{j_1 j_2}) \ln \frac{1 - y_{j_1 j_2}(k)}{1 - \hat{y}_{A j_1 j_2}(k)}, \end{cases}$$

accepted in image recognition tasks [14], are used, then the learning algorithm (9) takes the form

$$\begin{cases} \Delta A_{j_1 j_2}(k) = \eta_A(k) e_{j_1 j_2}(k) B_{j_1 j_2}^T(k-1) \tilde{X}^T(k), \\ \Delta B_{j_1 j_2}(k) = \eta_B(k) e_{A j_1 j_2}(k) \tilde{X}^T(k) A_{j_1 j_2}^T(k), \end{cases} \quad (12)$$

practically coinciding with (5), that allows to optimize the learning rate as in the algorithm (6).

Thus, the learning process of the matrix 2D-Rosenblatt's perceptron can be optimized by speed in image recognition problems.

IV. DEEP 2D-NEURAL NETWORK LEARNING

Let's introduce into consideration a multilayer neural network, each layer of which realizes a transformation of the form

$$O^{[l]}(k) = \Psi \odot (A^{[l]}(k-1)O^{[l-1]}(k)B^{[l]}(k-1))$$

where $O^{[l]}(k)$, $O^{[l-1]}(k)$ are $(n_1 \times n_2)$ -matrices of the outputs of the l th and the $(l-1)$ th layers correspondingly, $l=1,2,\dots,L$, L is the total number of layers in the network.

Then the general transformation realized by the network as a whole can be written in the form

$$\hat{Y}(k) = \Psi \odot (A^{[L]}(k-1)(\Psi \odot (A^{[L-1]}(k-1) \times (\dots \Psi \odot (A^{[1]}(k-1)\tilde{X}(k)B^{[1]}(k-1))\dots) \times B^{[L-1]}(k-1)))B^{[L]}(k-1)). \quad (13)$$

To train the neural network (13), an error backpropagation procedure can be used using the tuning algorithms (10), (11).

In this, the learning process of the output L th layer is realized according to the relations

$$\begin{cases} \Delta a_{j_1 j_2 i_1}^{[L]}(k) = \eta_A(k) \delta_{j_1 j_2}^{[L]}(k) \hat{o}_{i_1}^{[L-1]}(k), \\ \Delta b_{j_1 j_2 i_2}^{[L]}(k) = \eta_B(k) \delta_{A j_1 j_2}^{[L]}(k) \hat{o}_{i_2}^{[L-1]}(k) \end{cases}$$

where components are $\hat{o}_{i_1}^{[L-1]}(k) = \sum_{i_2=0}^{n_2} b_{j_1 j_2 i_2}^{[L]}(k-1) o_{i_2}^{[L-1]}(k)$,

$$\delta_{j_1 j_2}^{[L]}(k) = \psi'(u_{j_1 j_2}^{[L]}(k)) e_{j_1 j_2}(k), \quad \hat{o}_{i_2}^{[L-1]}(k) = \sum_{i_1=0}^{n_1} a_{j_1 j_2 i_1}^{[L]}(k) o_{i_1}^{[L-1]}(k),$$

$$\delta_{A j_1 j_2}^{[L]}(k) = \psi'(u_{A j_1 j_2}^{[L]}(k)) e_{A j_1 j_2}(k).$$

To tune the l th hidden layer ($1 < l < L$), equations

$$\begin{cases} \Delta a_{j_1 j_2 i_1}^{[l]}(k) = \eta_A(k) \delta_{j_1 j_2}^{[l]}(k) \hat{o}_{i_1}^{[l-1]}(k), \\ \Delta b_{j_1 j_2 i_2}^{[l]}(k) = \eta_B(k) \delta_{A j_1 j_2}^{[l]}(k) \hat{o}_{i_2}^{[l-1]}(k) \end{cases}$$

are used, where $\delta_{j_1 j_2}^{[l]}(k) = \psi'(u_{j_1 j_2}^{[l]}(k)) \sum_{i_1=0}^{n_1} \delta_{j_1 j_2}^{[l+1]}(k) a_{j_1 j_2 i_1}^{[l+1]}(k)$,

$$\hat{o}_{i_1}^{[l-1]}(k) = \sum_{i_2=0}^{n_2} b_{j_1 j_2 i_2}^{[l]}(k-1) o_{i_2}^{[l-1]}(k),$$

$$\delta_{A i_2}^{[l-1]}(k) = \sum_{i_1=0}^{n_1} a_{j_1 j_2 i_1}^{[l]}(k) o_{i_1}^{[l-1]}(k)$$

$$\delta_{A j_1 j_2}^{[l]}(k) = \psi'(u_{A j_1 j_2}^{[l]}(k)) \sum_{i_2=0}^{n_2} \delta_{A j_1 j_2}^{[l+1]}(k) b_{j_1 j_2 i_2}^{[l+1]}(k).$$

And, finally, the parameters of the first layer of the network are tuned using the procedure

$$\begin{cases} \Delta a_{j_1 j_2 i_1}^{[1]}(k) = \eta_A(k) \delta_{j_1 j_2}^{[1]}(k) \hat{o}_{i_1}^{[0]}(k), \\ \Delta b_{j_1 j_2 i_2}^{[1]}(k) = \eta_B(k) \delta_{A j_1 j_2}^{[1]}(k) \hat{o}_{i_2}^{[0]}(k), \end{cases}$$

where $\delta_{j_1 j_2}^{[1]}(k) = \psi'(u_{j_1 j_2}^{[1]}(k)) \sum_{i_1=0}^{n_1} \delta_{j_1 j_2}^{[2]}(k) a_{j_1 j_2 i_1}^{[2]}(k)$,

$$\hat{o}_{i_1}^{[0]}(k) = \sum_{i_2=0}^{n_2} b_{j_1 j_2 i_2}^{[1]}(k-1) x_{i_2}(k), \quad \hat{o}_{i_2}^{[0]}(k) = \sum_{i_1=0}^{n_1} a_{j_1 j_2 i_1}^{[1]}(k) x_{i_1}(k),$$

$$\delta_{A j_1 j_2}^{[1]}(k) = \psi'(u_{A j_1 j_2}^{[1]}(k)) \sum_{i_2=0}^{n_2} \delta_{A j_1 j_2}^{[2]}(k) b_{j_1 j_2 i_2}^{[2]}(k).$$

V. COMPUTATIONAL EXPERIMENTS

The efficiency of the proposed system was demonstrated on the classification task. The experiment was carried out on the hand-written digits dataset from the UCI repository [15].

In general the dataset has 10 classes (digits from 0 to 9).

Some examples of the images from this dataset are presented in Fig. 1.

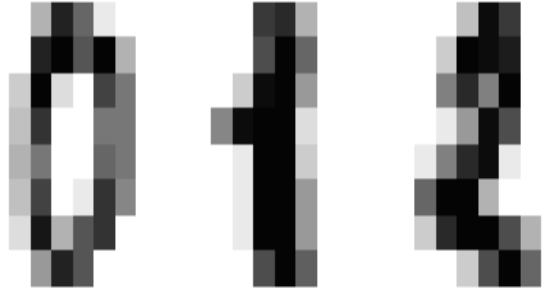


Fig. 1. Examples of the images from the hand-written digits dataset

The dataset contains 5620 observations of digits 0–9. Each observation is presented as a 8×8 matrix of digits that represents pixel values. These values were preprocessed before training using normalization.

The results of the experiment depending on the number of hidden layers in the network are presented in Table I.

TABLE I. RESULTS OF THE EXPERIMENT

Number of hidden layers	Testing error, %
1	45
2	40
3	35
4	34
5	32
10	20

CONCLUSION

A deep 2D-neural network and its learning algorithm are proposed. They are based on the bilinear matrix model, the 2D analogue of the elementary Rosenblatt's perceptron, the error backpropagation procedure, and the matrix analogue of the Kaczmarz–Widrow–Hoff algorithm. The considered DNN possesses an increased speed due to the reduction of the number of adjusted synaptic weights and the learning algorithm optimized by speed.

REFERENCES

- [1] Y. LeCun, Y. Bengio, and G. Hinton, "Deep Learning," *Nature*, vol. 521, pp. 436-444, 2015.
- [2] J. Schmidhuber, "Deep Learning in neural networks: An overview," *Neural Networks*, vol. 61, pp. 85-117, 2015.
- [3] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016.
- [4] Y. Guo, Y. Liu, A. Oerlemans, S. Lao, S. Wu, and M. S. Lew, "Deep learning for visual understanding: A review," *Neurocomputing*, vol. 187, pp. 27-48, 2016.
- [5] P. Stubberud, "A vector matrix real time backpropagation algorithm for recurrent neural networks that approximate multi-valued periodic functions," *Int. J. on Computational Intelligence and Application*, 8(4), pp. 395-411, 2009.
- [6] P. Daniušis and P. Vaitkus, "Neural networks with matrix inputs," *Informatica*, 19, №4, pp. 477-486, 2008.
- [7] M. Mohamadian, H. Afarideh, and F. Babapour, "New 2D Matrix-Based Neural Network for Image Processing Applications," *IAENG International Journal of Computer Science*, 42(3), pp. 265-274, 2015.
- [8] J. Gao, Y. Guo, and Z. Wang, "Matrix neural networks," in *Proceedings of the 14th International Symposium on Neural Networks (ISNN), Part II*, Sapporo, Sapporo, Japan, pp. 1-10, 2017.
- [9] V. M. Kuntsevych and M. M. Lychak, *Synthesis of optimal and adaptive control systems. The game approach* (В. М. Кунцевич, М. М. Лычак, Синтез оптимальных и адаптивных систем управления. Игровой подход). Kyiv : Naukova dumka, 1985.
- [10] V. M. Kuntsevych, "On a solving of the problem of two-dimensional discrete filtration (synthesis of matrix filters)," *Automatica i Telemekhanika*, no. 6, pp. 68-78, 1987.
- [11] Ye. V. Bodyanskiy, and I. P. Pliss, "On a solving of the problem of a matrix object controlling under uncertainty conditions," *Automatica i Telemekhanika*, no. 2, pp. 175-178, 1990.
- [12] Ye. Bodyanskiy, I. Pliss, and V. A. Timofeev, "Discrete adaptive identification and extrapolation of two-dimensional fields," *Pattern Recognition and Image Analysis*, vol. 5, no. 3, pp. 410-416, 1995.
- [13] S. Haykin, *Neural Networks: A Comprehensive Foundation*. Upper Saddle River, N. J. : Prentice Hall, Inc., 1999.
- [14] C. M. Bishop, *Neural Networks for Pattern Recognition*. Oxford : Clarendon Press, 1995.
- [15] <http://archive.ics.uci.edu/ml/datasets/Optical+Recognition+of+Handwritten+Digits>

F-transform 3D Point Cloud Filtering Algorithm

Andriy Yerokhin
Computer Science Faculty
Kharkiv National University of Radio
Electronics
Kharkiv, Ukraine
andriy.yerokhin@nure.ua

Valerii Semenets, Alina Nechyporenko
Biomedical Engineering Department
Kharkiv National University of Radio
Electronics,
Kharkiv, Ukraine
alina.nechyporenko@nure.ua

Oleksii Turuta, Andrii Babii
Department of software engineering
Kharkiv National University of Radio
Electronics
Kharkiv, Ukraine
{oleksii.turuta, andrii.babii}@nure.ua

Abstract — The paper proposes a new 3D point cloud filtering approach using F-transform. We achieve this by usage of uniform fuzzy partitioning and applying direct and inverse discrete F-transform on a point cloud data. The point cloud was filtered in a depth domain. The performance of developed approach was compared with several common-used statistical-based point cloud filtering methods.

Keywords — point cloud, filtering, fuzzy methods, F-transform

I. INTRODUCTION

A technology of 3D scanning have been in active development over past few decades. Various hardware can be used to obtain the point cloud data. This includes implementations based on usage of lasers, structural light, infrared light, stereo and time-of-signal measurement methods [1].

Acquired data can be used in various applications, including medical, environmental, engineering and entertainment. Low-cost 3D sensors become more accessible over past few years [2], but point clouds obtained with such sensors, however, contain noise component and outliers due to different reasons, related with the technology limitation, nature of the measured surfaces and lighting [3]. Therefore, 3D point cloud filtering is an important pre-processing operation of raw data registered by sensor. Filtering approaches can be divided into the following seven groups of algorithms for filtering the 3D point cloud [4]:

- Filtering techniques based on statistic. These methods include various static parameters for filtering: local median filtering, local averaging, kernel-based clustering approaches, different variant of principal component analysis, Bayesian statistics usage for denoising, variations of Least Squares approach, optimization techniques on normal estimation, growing neural gas networks.
- Filtering based on neighborhood analysis. This approach estimates results using similarity metric for a point and its neighborhood. This wide group of methods such as: bilateral filters and its variations, 3D mean shift filter, 3D non-local means filter, filters based on relative deviation of the local neighborhood metrics and the average local neighborhood
- Projection-based approaches. This is a set of methods based on adjusting the position of each point using different projection strategies. It includes a locally optimal projection operator and its modifications, moving least squares and its modifications, plane fitting operation.

- Filtering based on a signal processing methodology. It includes usage of Laplacian operator and its modifications, Fourier transformation based filters, combination of Laplace–Beltrami operator and Wiener filtering.
- Filtering based on partial differential equations (PDE). PDE can be applied to pre-processed neighborhoods, curvatures, weighted arbitrary graphs to filter point clouds.
- Hybrid filtering technique. Usually this is a mix of two or more methods from different groups to filter point clouds.
- Other methods. This class includes groups of methods which are not related to previous points. It includes voxel grid, oriented particles and similar filtering approaches.

In this paper we will focus on F-transform application to the point cloud processing. F-transform is an approximation method based on fuzzy partitioning proposed by Perfilieva in [5] and it has various practical applications. Fuzzy sets are used to deal with uncertain information [6] usually, but fuzzification of the crisp data and usage of fuzzy theory may give additional benefits for data processing. In paper [7] the main principles of F-transform usage for noise filtering were described. Application of fuzzy transformation to image processing were published in paper [8]. Approaches based on usage of fuzzy relations and choosing of optimal granulation level (which were represented by a fuzzy rule) were described in [9]. F-transform were used for image processing in the work [10].

The aim of this work is to apply the F-transform to 3D point cloud filtering.

Section II describes the input data sources and methodology of the point cloud processing. Section III describes an application of the F-transform to 3D point cloud filtering and some aspects related to usage of this transformation. Results of filtering and comparing with several statistical-based 3D filtering methods are demonstrated in section IV. Conclusions and some of further research directions are described in Section V

II. POINT CLOUD DATA PREPROCESSING

A. Data sources

In the current paper, we used the Stanford Models dataset [11]. These models are scanned by a Cyberware 3030 MS scanner. There are 3D models, such as Stanford Bunny, Drill bit, Happy Buddha, Dragon, Armadillo and

etc. For example Fig. 1, the Stanford Bunny model consists of 10 scans and total size is 362,272 points (about 725,000 triangles). The model was aligned by the modified ICP algorithm [12]. The reconstruction size is 35947 vertices and 69451 triangles. The 3D model is zipping and smoothing merged volumetric range of image of the manifold surfaces. The point cloud was processed after 3D scanning. The reconstructed model is point cloud which stored in a PLY file format (ASCII format).

The Stanford Bunny model (Fig. 1) is used in pre-processing section.



Fig. 1. Stanford Bunny

B. Point cloud data pre-processing

A high-level overview of fuzzy filtering of a 3D point cloud includes a pre-processing irregular point cloud, preparing patches, fuzzy filtering for noise reduction of patches as a result point clouds. Overview of pre-processing stages is given in Fig. 2.

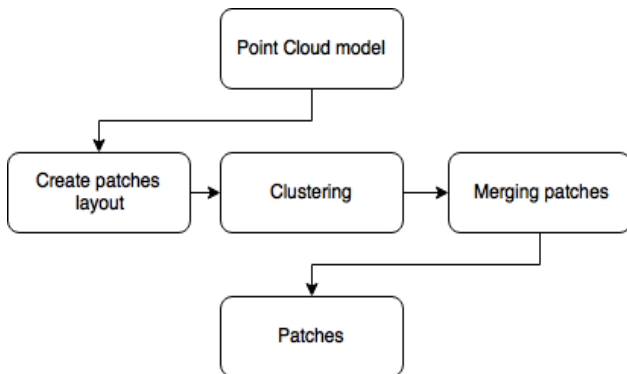


Fig. 2. High-level stages of pre-processing point cloud.

First of all, the irregular point cloud model is divided into a number of points. A set of closed points is collected into raw patches. Criteria of collection points are defined in [12 GEOMETRY]. Such a patches layout describe the surface of this 3D objects presented by point clouds.

Secondly, we cluster points of patches layout into clusters. A binary space partition (BSP) tree samples the set along size of bounding box. We merge the leaves of the tree into clusters with a common parent in the BSP tree. There are potentially many other neighbors so we stop clustering stage when cluster achieves need size (20-100 samples, depending of model) [12].

Thirdly, created patches layout has local neighborhood information. An optimization approach [12] is used for merging patches. Let a local quality metrics Φ and P_i, P_j be

neighboring patches such as $P = (P_i, P_j)$ which merges candidate patches.

Φ is defined as:

$$\Phi(P) = \Phi_{Size}(P) \cdot \Phi_{NC}(P) \cdot \Phi_B(P) \cdot \Phi_{Reg}(P) \quad (1)$$

where, $\Phi_{Size}(P)$ - estimation of high priority of small patches; $\Phi_{NC}(P)$ - estimation of penalty of increasing width of patch; $\Phi_B(P)$ - estimation of boundary of the patches; $\Phi_{Reg}(P)$ - is regularization of the patch distribution.

The raw patches are merged into a connected region of surface.

Finally, we have got the merged patches of point clouds. Visualization of merged patch presented in Fig. 3. We create a mesh of point clouds. It is a part of nose of the Stanford Bunny. In next section we will apply fuzzy transform (FT) approach for smoothing this patch.

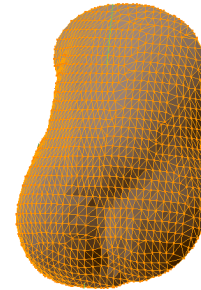


Fig. 3. Visualization of merged patch

III. 3D POINT CLOUD FT-SMOOTHING

Filtering of noise in the patch with point cloud, received in the previous section will be next step in point cloud data processing. Each point cloud element will store information about X-coordinate, Y-coordinate and Z-coordinate - depth. We assume that points with noise in depth domain should be modified. Output will be filtered patch with depth component in Z-coordinate. We propose to apply F-transform [7] method to solve this problem.

A. Fuzzy partition

F-transform technique requires defining of fuzzy partitions. It can be partitions according to Ruspini condition [12], called Ruspini partitions. Each center of this partition is placed uniformly. As a result, we obtain a regular grid of partition centers.

In the paper [13] uniform fuzzy partitions were modified to parametric generalized fuzzy partitions, with some relaxations of Ruspini condition. In this paper we use non-modified uniform fuzzy partitions.

In order to define two-dimension fuzzy partitions we use a generalization of single-dimension partitioning:

You can find an example of a triangle membership function in Fig. 4.

Let X be vector containing fixed nodes on interval $[a, b]$ so that $x_1 = a, x_m = b$ and $m \geq 2$. The membership

functions A_1, \dots, A_m of the fuzzy sets should fulfill the following conditions [5] for $i = 1, \dots, m$:

1. A_i is continuous;
2. A_i strictly increases on $[x_{i-1}, x_i]$ and strictly decreases on $[x_i, x_{i+1}]$;
3. $A_i : [a, b] \rightarrow [0, 1], A_i(x_i) = 1$;
4. $A_i(x) = 0$ if $x \notin (x_{i-1}, x_{i+1})$, where we set $x_0 = a$ and $x_{m+1} = b$;
5. For all $x \in [a, b]$,

$$\sum_{i=1}^m A_i(x) = 1 \quad (2)$$

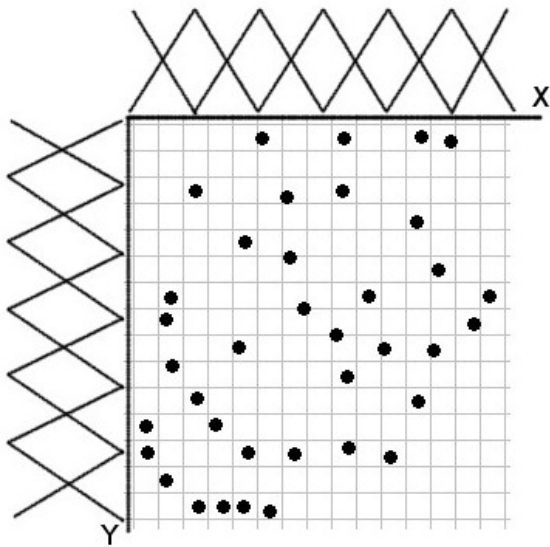


Fig. 4. Triangle membership functions for fuzzy uniform partitioning of 2D projection of point cloud.

The membership function can have different shapes. Shape of the function should be chosen by user. In this paper we use uniform fuzzy partitions.

Shape of the membership function and set of centers should be chosen at the beginning of computation.

B. FT-smoothing point cloud filtering

The membership function should be defined for each axis in order to create F-transform for two dimensions

We consider a set of points in the point cloud which can be represented as a function f , defined on $[M_a, M_b] \times [N_a, N_b]$, where M_a, M_b are boundaries of the patch on X-axis, and N_a, N_b - boundaries on Y-axis.

It is assumed, that values in coordinates (x, y) are the depth values, measured by Z-coordinate axis in point cloud belonging to the set P.

Let A_1, \dots, A_m and B_1, \dots, B_n be membership functions for partitions on X and Y axes. A point of the point cloud is

covered by membership function for two dimensional fuzzy partition with membership functions A_i and B_j if $A_i(x) > 0$ and $B_j(y) > 0$.

Let p_l will be value of coordinate at x-axis, q_k -value of coordinate on y-axis, $f(p_l, q_k)$ will be dense (Z) value at (p_l, q_k) .

We define matrix $[F_{kl}]$ as a $m \times n$ matrix of real numbers, are called F-components.

F-transform of f with respect to $\{A_1, \dots, A_m\}$ and $\{B_1, \dots, B_n\}$ for all $i = 1, \dots, m, j = 1, \dots, n$ will be based on equations proposed in [15] and can be re-formulated as follows:

$$F_{ij} = \frac{\sum_{l=M_a}^{M_b} \sum_{k=N_a}^{N_b} f(p_l, q_k) A_i(p_l) B_j(q_k)}{\sum_{l=M_a}^{M_b} \sum_{k=N_a}^{N_b} A_i(p_l) B_j(q_k)} \quad (3)$$



Fig. 5. Patch of 3D point cloud (mesh added for visualization) with noise.



Fig. 6. Filtered by FT-smoothing patch of 3D point cloud (mesh added for visualization).

The inverse transformation is applied to receive smoothed values:

$$\hat{f}(k, l) = \sum_{i=1}^m \sum_{j=1}^n F_{ij} A_i(p_l) B_j(q_k) \quad (4)$$

Example of point cloud filtering represented in Fig. 5, 6.

This function \hat{f} will be approximation of dense component represented in two-dimensional image f .

IV. IMPLEMENTATION AND RESULTS

The experiments were carried on several point clouds from Stanford Models dataset. The input point clouds are corrupted by simulated Gaussian noise.

The performance of the filter is measured using peak signal to noise ratio (PSNR).

$$PSNR = 10 \cdot \log_{10} \left(\frac{MAX^2}{MSE} \right) \quad (5)$$

Where MAX – maximum depth value in point cloud, and MSE – mean squared error. Table I contains the results of experiments.

The experiments show that FT-smoothing filter is better for Gaussian, Uniform, Rayleigh noise types. Quality of the FT-filter is average for the Poisson noise type. So, we observed that the proposed approach yielded satisfactory results and outperforms the other methods presented in the literature.

TABLE I. EXPERIMENT RESULTS

Type of noise	Denoised point cloud PSNR				
	<i>FT-smoothing filter</i>	<i>Median filter</i>	<i>Average filter</i>	<i>Min filter</i>	<i>Max filter</i>
Gaussian noise	23.16	22.72	22.68	14.44	18.07
Uniform noise	25.20	24.59	24.38	16.23	16.83
Poisson noise	24.50	26.69	24.86	27.18	8.47
Rayleigh noise	22.78	22.33	22.35	18.01	14.68

V. CONCLUSION AND FUTURE DEVELOPMENTS

In this paper various point cloud filtering methods have been discussed and analyzed. The proposed FT-smoothing approach was compared with the most common statistical filters. The current study can be extended by including of more numbers of noise types and complex filters from different filtering groups.

The proposed implementation of FT-smoothing filter uses–uniform fuzzy partitions. Further developments should include directions related to usage of non-uniform and generalized fuzzy partitions and its optimal parameter selection.

Optimal settings of membership functions selection for measured scene, using small sample from point cloud can be also defined as a perspective development in this domain.

REFERENCES

- [1] J. Fu, D. Miao, W. Yu, S. Wang, Y. Lu, and S. Li, "Kinect-like depth data compression," IEEE Transactions on Multimedia, vol. 15, no. 6, pp. 1340–1352, 2013.
- [2] J. Han, L. Shao, D. Xu, and J. Shotton, "Enhanced computer vision with microsoft kinect sensor: A review," IEEE transactions on cybernetics, vol. 43, no. 5, pp. 1318–1334, 2013.
- [3] A. Nurunnabi, G. West, and D. Belton, "Outlier detection and robust normal-curvature estimation in mobile laser scanning 3D point cloud data," Pattern Recognition, vol. 48, no. 4, pp. 1404–1419, 2015.
- [4] X.-F. Han, J. S. Jin, M.-J. Wang, W. Jiang, L. Gao, and L. Xiao, "A review of algorithms for filtering the 3D point cloud," Signal Processing: Image Communication, vol. 57, pp. 103–112, 2017.
- [5] I. Perfilieva, "Fuzzy transforms: Theory and applications," Fuzzy sets and systems, vol. 157, no. 8, pp. 993–1023, 2006.
- [6] L. A. Zadeh, "Fuzzy sets," in Fuzzy Sets, Fuzzy Logic, And Fuzzy Systems: Selected Papers by Lotfi A Zadeh, World Scientific, 1996, pp. 394–432.
- [7] I. Perfilieva and R. Valášek, "Fuzzy transforms in removing noise," in Computational Intelligence, Theory and Applications, Springer, 2005, pp. 221–230.
- [8] Y. Nie and K. E. Barner, "The fuzzy transformation and its applications in image processing," IEEE Transactions on Image processing, vol. 15, no. 4, pp. 910–927, 2006.
- [9] K. Hirota and W. Pedrycz, "Fuzzy relational compression," IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics), vol. 29, no. 3, pp. 407–415, 1999.
- [10] F. Di Martino, V. Loia, I. Perfilieva, and S. Sessa, "An image coding/decoding method based on direct and inverse fuzzy transforms," International Journal of Approximate Reasoning, vol. 48, no. 1, pp. 110–131, 2008.
- [11] M. Levoy, J. Gerth, B. Curless, and K. Pull, "The Stanford 3D scanning repository," URL <http://www-graphics.stanford.edu/data/3dscanrep>, 2005.
- [12] G. Turk and M. Levoy, "Zippered polygon meshes from range images," in Proceedings of the 21st annual conference on Computer graphics and interactive techniques, pp. 311–318, 1994.
- [13] E. H. Ruspini, "A new approach to clustering," Information and control, vol. 15, no. 1, pp. 22–32, 1969.
- [14] L. Stefanini, "F-transform with parametric generalized fuzzy partitions," Fuzzy Sets and Systems, vol. 180, no. 1, pp. 98–120, 2011.
- [15] P. Vlačánek and I. Perfilieva, "Image reconstruction with usage of the F-Transform," in International Joint Conference CISIS'12-ICEUTE'12-SOCO'12 Special Sessions, pp. 507–514, 2013.

Software for Visual Insect Tracking Based on F-transform Pattern Matching

Petr Hurtik
IRAFM, CEIT4I University
of Ostrava Ostrava, Czech
Republic petr.hurtik@osu.cz

David Číž
Department of informatics and computers
University of Ostrava
Ostrava, Czech Republic
davidciz95@gmail.com

Oto Kakáb
Department of Biology and Ecology
University of Ostrava
Ostrava, Czech Republic
kalab.oto@gmail.com

David Musiolek
Department of Biology and Ecology
University of Ostrava
Ostrava, Czech Republic
david.musiolek@osu.cz

Petr Kocárek
Department of Biology and Ecology
University of Ostrava
Ostrava, Czech Republic
petr.kocarek@osu.cz

Martin Tomis
Department of Telecommunications
VSB-TU Ostrava
Ostrava, Czech Republic
martin.tomis@osu.cz

Abstract—We introduce a problem of tracking small animals, especially insects. To solve this problem, we focus on visual tracking in recorded movies, propose our pattern tracking mechanism based on F-transform, and implement a user-friendly software to handle the movies. The tracking core is compared with five state-of-the-art tracking algorithms: KCF, MIL, TLD, Boosting and MedianFlow from processing time and algorithm failure rate point of views. Based on the results computed from 1000 movie frames, we observed that the proposed F-transform tracking core is the fastest and the most reliable method.

Index Terms—Gryllus Assimilis, insect tracking, visual tracking, F-transform, pattern matching, 4k movie

I. INTRODUCTION

In Zoology field, transmitters placed onto animals are used to track animal paths and therefore help to understand the behavior of the tracked object. In the case of big animals such as wolfs, tigers etc., there are many publications dealing with such tracking, see e.g., [1] and [2]. In those cases, transmitters are small enough to be placed on such big animals without affecting their behavior or ability to move. Problems ensue when biologists want to track insects because the transmitters weight and size can affect the insect's behavior and therefore invalidate the research. In our work, we investigate the impact of various transmitters on insects, namely on a field cricket, *Gryllus assimilis* (Orthoptera), in order to define a safe weight an insect can carry without it influencing its movement. To formulate such statement, we need to record statistically big enough quantity of insects with various transmitters and track their movement. Because of the recorded data size, it is not possible to track them manually, therefore we have to design automated tracking software. The design of such application is the topic and the main goal of this paper. The paper structure is following: at first, we briefly describe visual tracking process in Section II and then we recall state-of-the-art software and algorithms for pattern tracking in Section III. Our own pattern

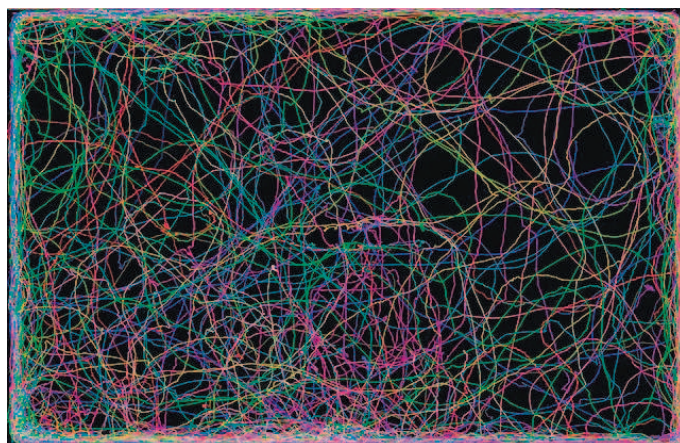


Fig. 1. Illustration of the task goal: tracked movement of insects in an arena.

tracking algorithm is described in Section IV. The detailed description of the considered insect tracking task with its obstacles, application design and benchmarking are the aims of Section V. Finally, conclusions are formulated in Section VI.

II. PATTERN TRACKING

Visual pattern tracking (tracing) [3] can be viewed as a special case of pattern matching technique [4]. Generally, pattern matching is searching and checking given object(s) (image where an insect is captured, in our case) for a presence of given patterns (images) in order to find and mark the patterns locations (if any) within the given objects. The matching can be exact or approximate. In the exact one, the pattern appears in its exact form while in the approximate one, it is allowed some freedom between the pattern and its found match. Formally, we expect computer two-dimensional image $f : D \rightarrow L$, $D = \{1, 2, \dots, W\} \times \{1, 2, \dots, H\}$

and $L = \{0, 1, \dots, 2^n - 1\}^c$, where W and H is width and height respectively, n denotes considered bit depth (amount of intensities, standardly $n = 8$) and c denotes number of color channels, i.e., $c = 1$ for a grayscale image and $c = 3$ for, e.g., RGB color model. Such image function is a space (usually called as a database) where we search pattern for. As a pattern, we consider image $f' : D' \rightarrow L$, which is a proper inclusion of f and $|D'| \leq |D|$ holds. The goal is to browse f and find inside it a sub-area f_s such $DistDist(f_s, f_p)$, where $Dist$ is a metric is minimal for all possible considered sub-areas.

The pattern tracking instead of one image f consider a database of functions $\mathbf{f} = \{f_1, f_3, \dots, f_t\}$, where the image functions are captured in various time moments. As a typical representative, we can mention a movie from a camera decomposed into particular images, called frames. The point is that a frame has not be matched fully in all frames. Because of the images in a database are supposed to be time-ordered and we know previous pattern location, we can search for the pattern in a Δ -neighborhood of the previous pattern location in an actual frame, where Δ is maximal considered pattern movement between frames. Further, because the previous pattern locations are known, we can determine pattern trajectory and approximate it in future frames in order to create a more precise match or to continue with tracking even if the pattern is covered by another object.

III. EXISTING SOFTWARE AND ALGORITHMS

In this paper, we will describe own made software for insect tracking, which will be benchmarked. In order to compare the software, we will describe current existing solutions and algorithms in this section.

A. Existing solutions in the form of software

During solving the problem, we searched for a full software dealing with the insect tracking problem at first. We omitted old, unsupported software and handled for different applications. The list of investigated applications is following.

*Ctrax*¹ is a software specialized in tracking walking flies. Unfortunately, we were not able to read the movie by the application even if it was converted from mp4 into avi file as is stated in the application manual.

*Bio-Records*² is aimed on general insect tracking. We were not able to use the application because it is not available for Windows operating system. The reason why we mention it is it is known in zoology field and because its web page and youtube presentation movies are impressing. On the other hand, even in their short demonstration movie³ there are visible a lot of cases when insect tracking fails.

*Noldus animal tracker*⁴ At least by the software price and its presentation, it is one-of-the-best software for animal tracking.

We tried to obtain free trial version, but we were not successful in a negotiation process with the company.

*SwisTrack*⁵ more than a complex tracking software, it is a package of basic graphical operations which can be stylized by a user into full algorithm serving as a tracker. We faced the same problem as in the case of *Ctrax* - the software was unable to open our movie, it shows "codec missing" even if we have had installed required codecs according to the reference manual.

*Winanalyze*⁶ is a software for a general tracking such as movement tracking etc. Our experience is the software critically crashed when a movie was opened.

As it is obvious, no one of the investigated software can be used to solve our task. The general problem is the applications are designed for one particular task without enough generality which is caused, e.g., by supporting minimum input file types.

B. Existing algorithms

To be our current state overview complete, we investigated also state-of-the-art pattern tracking algorithms which we used as a core in our implemented application described in Section V-B. The list of tested algorithms is following.

Adaptive color attributes tracker (KCF) [5]: Danelljan et al. propose to improve CSK tracker [6] which compares two bag-of-words of patterns, where words are shapes and colors. The improvement relies on adding more extracted color features. The benefit of the algorithm is it runs in real-time for a reasonable-big (small) movie resolution.

Tracking-Learning-Detection (TLD) [7]: the approach idea is not to accumulate error between initialized pattern and actual one in long-term tracking. This approach is similar to the idea of our algorithm and it realizes three steps: tracking, position improving and pattern re-learning. Pattern position is improved by performing pattern matching while pattern-relearning use actual detected area as the pattern when a distance between the pattern and the searched pattern is big enough.

Online Multiple Instance Learning (MIL) [8]: the algorithm works with a set of patches around the selected pattern, placed into bag-of-words in order to work as a weak classifier [9].

Real-Time Tracking via On-line Boosting [10]: the algorithm is based on the original on-line AdaBoost [11], i.e., it uses a lot of weak classifiers in order to establish a strong one. The benefit of the algorithm is that, at first, it can adapt itself to pattern changes online and, at second, the performance in comparison with the original work is improved because author designed so-called "global weak classifier" which allows updating all selectors in time.

Median Flow [12]: authors use Lucas-Kanade tracker [13] and improve it by tracking trajectory. In opposite to standard algorithms which process ascending frames, Median Flow tracks both ascending and descending frames, search for irregularities in searched locations given by Lucas-Kanade tracker and filter them out.

¹<http://ctrax.sourceforge.net/>

²<http://www.bio-tracking.org>

³<http://www.youtube.com/watch?v=T5W0iplroSg>

⁴<http://www.noldus.com/animal-behavior-research>

⁵<https://en.wikibooks.org/wiki/SwisTrack>

⁶<https://winanalyze.com>

IV. F-TRANSFORM PATTERN TRACKING

In this section, we will briefly describe how our tracking algorithm works. It is based on our general pattern matching algorithm [4] and because its speed is superior [14], we are able to follow the idea "tracking by matching" without significant loosing of performance and achieving high precision.

The algorithm is based on the idea of transforming images from their domain into a reduced domain using F-transform [15]. We can define the direct F-transform of a discrete function f of two variables defined on $[1, W] \times [1, H]$. Let $f : P \rightarrow \mathbb{R}$ where $P = \{(i, j) | i = 1, \dots, W; j = 1, \dots, H\}$ and let $\{A_1, \dots, A_m\} \times \{B_1, \dots, B_n\}$ establish a fuzzy partition of $[1, W] \times [1, H]$ such that $m < W$, $n < H$ and $\forall k, l \exists i, j; A_k(i)B_l(j) > 0$. Then the direct F-transform of f w. r. t. the chosen partition is a matrix $\mathbf{F}_{mn}[f] = (F[f]_{kl})$, $k = 1, \dots, m, l = 1, \dots, n$, of F-transform components where the components are defined for all $k = 1, \dots, m, l = 1, \dots, n$ and $(i, j) \in P$ as follows

$$F[f]_{kl} = c_{kl}(f \otimes C_{kl}),$$

where \otimes is a convolution, $C_{kl} = A_k B_l$, and

$$c_{kl} = \left(\sum_{i=1}^W \sum_{j=1}^H C_{kl}(i, j) \right)^{-1}.$$

Let us remark that in the following algorithm, we use the direct F-transform with respect to the h -uniform fuzzy partition of $[1, W] \times [1, H]$. The parameter h influences the number of basic functions in the fuzzy partition and thus also the size of the matrix $\mathbf{F}_{mn}[f] = (F[f]_{kl})$ representing the original function f . Specifically, the larger h the lesser components $F[f]_{kl}$ and thus, the bigger reduction of the original function. The algorithm consists of two phases. Both of them use the same-valued user-defined parameter h . The first one aims to pattern preparing as:

- 1) Take pattern f_p given on $[1, W_p] \times [1, H_p]$.
- 2) For f_p create two fuzzy partitions with respect to the same parameter h . The first one notated as $\langle 1 \rangle$ is created on $[1, W_p] \times [1, H_p]$, the second one $\langle 2 \rangle$ is on $[h/2, W_p] \times [h/2, H_p]$.
- 3) Compute the F-transform components for f_p , with respect to both fuzzy partitions, i.e., for f_p , we obtain two representations given by matrices $\mathbf{F}[f_p]\langle 1 \rangle$ and $\mathbf{F}[f_p]\langle 2 \rangle$, of the F-transform components.

The second part realizes the pattern matching as follows:

- 1) Take a particular movie frame (the actual image) and represent it by a discrete function f_D given on $[1, W_D] \times [1, H_D]$.
- 2) Create a fuzzy partition of $[1, W_D] \times [1, H_D]$ with respect to the same parameter h given in the first part.
- 3) Compute the F-transform components of f_D with respect to the fuzzy partition and obtain the matrix $\mathbf{F}[f_D]$.
- 4) Compare the matrices $\mathbf{F}[f_p]\langle j \rangle$, $j = 1, 2$, with $\mathbf{F}[f_D]$ by sliding windows comparison by computing distances between components $Dist_i(\mathbf{F}[f_p]\langle 1 \rangle, \mathbf{F}[f_D])$

and $Dist(\mathbf{F}[f_p]\langle 2 \rangle, \mathbf{F}[f_D])$, and remember the particular position of f_p in f_D where $Dist = Dist(\mathbf{F}[f_p]\langle 1 \rangle, \mathbf{F}[f_D]) + Dist(\mathbf{F}[f_p]\langle 2 \rangle, \mathbf{F}[f_D])$ is the smallest one.

- 5) Take the position $\mathbf{p} = \{p_x, p_y\}$ of f_p in f_D .

Such proposed algorithm realizes pattern matching. In order to realize pattern tracking, we propose following upgrades:

- 1) Instead of full frame domain $[1, W_D] \times [1, H_D]$, we consider small sub-area $[p_{x,t-1} - \Delta, p_{x,t-1} + \Delta] \times [p_{y,t-1} - \Delta, p_{y,t-1} + \Delta]$, where $t - 1$ denotes previous location, i.e., we use previous known pattern location and search in only its Δ -neighborhood. This restriction improves processing speed because searching space is reduced and also increases success rate from the same reason.
- 2) Let S is a spatial distance between pattern and its projection in searched sub-area and T_U is an threshold. When $S > T_U$, tracking is stopped. Such situation means pattern cannot be found.
- 3) When $S > T_L$, where T_L is a threshold and $T_L < T_U$ holds, algorithm replaces original pattern f_p by image of its located projection in actual searched sub-area. This pattern updating helps to handle pattern visual modification such as rotation.

V. EXPERIMENTS

In this section, we will formulate the exact conditions of our experiment at first, then design an application and finally, we will benchmark various tracking cores used inside the application.

A. Experiment setting and obstacles

In the experiment, we build a glass-side arena with dimensions 1200×800 mm. On the floor of the arena, a flat black cotton fabric is placed. The arena is recorded by a 4k camera, i.e., it is recorded with a resolution of 3840×2160 px in 24 frames per second. Further, we recorded simultaneously 20 pieces of insect at one time, where each one insect has placed a paper with a label on its back in order to unambiguously distinguish between them. As the label, we propose one set of characters $\{Z, X, B, O, M, H, 4, K, V, +\}$ in two colors: green and red. Because of the experiment has to be recorded during a night, the green and red colors are UV-reactive and over the arena is turned on a UV-light bulb. A part of a recorded movie is illustrated in Figure 2. Without any deeper exploration, the problem seems to be trivial to solve - a black background without any significant noise where highly-visible objects are moving.

Going into details, several problems appear. Even though a 4k movie is recorded, the insects and therefore the characters are really small - their size is approx 25×15 pixels so the possible information which can be extracted from such pattern is highly limited which may result in decreased success rate. The extracted patterns with all possible characters are visualized in Figure 3. Moreover, the labels are placed onto live insects which are moving in two dimensions, jump and

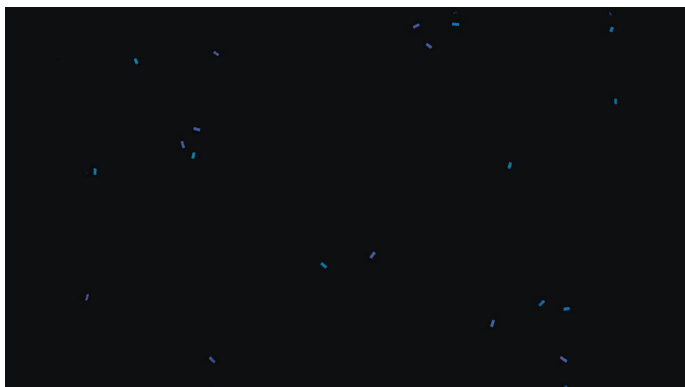


Fig. 2. Illustrative screenshot of a recorded movie

also rotate around its axes. From the reason, the same labeled pattern varies in time in its rotation, size and light reflection ability. An example is shown in Figure 4 where ten patterns with the same label "+" were arbitrarily extracted from one single movie. It is obvious that so big variety can lead to decreasing success rate when several insects are very close to each other and therefore a possible swap of different detected patterns may appear.

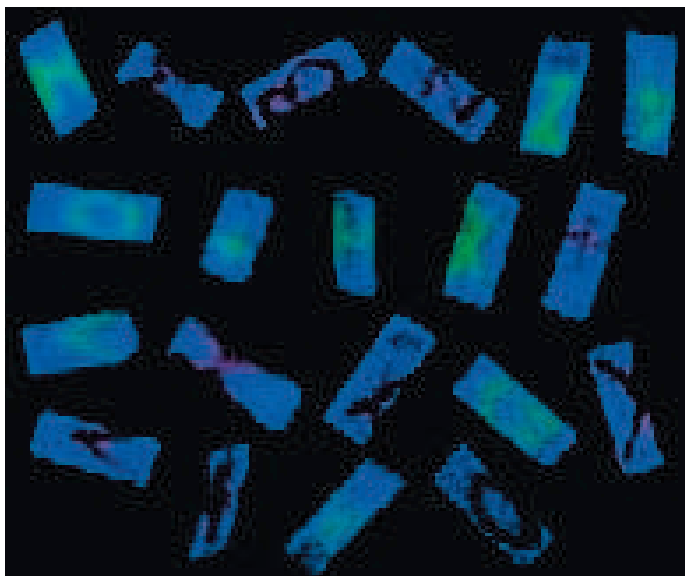


Fig. 3. Extracted objects - patterns from a single movie frame.

The next problem is processing speed. Because of the experiment design, 27 movies with the length of 600 seconds per movie were recorded. It results in $7.7 \cdot 10^6$ pattern positions to be tracked in 4k movie resolution. Just for illustration, if we consider general pattern tracking algorithm working in real-time (e.g., 30fps) for a full HD video, it means it would take 288 hours to process all the movies in our experiment setting excluding time for movie decoding/coding, handling failures etc.

The last obstacle is insect's behavior. Standardly, tracking algorithms include way, how a trajectory can be estimated



Fig. 4. Various rotations and light reflections of character "+".

when an object is lost - it can be done by, e.g., Kalman filter [16] and its derivatives. In our case, an insect can be lost for a while because one insect can cross another one by moving over its back. Unfortunately, Kalman filter etc. cannot be used for such tracking because an insect does not have a nice, smooth trajectory but its movement is *chaotic*. It can change direction randomly as same as speed - a non-moving insect can jump, i.e., it can achieve a massive acceleration between captured frames.

B. Application design

When handling a large number of movies, fast and reliable tracking core is useless without a user-friendly application interface. Therefore we implemented such interface using multi-platform Qt framework⁷ connected with C++ coding language. The application GUI is visualized in Figure 5 and offer following functionalities.

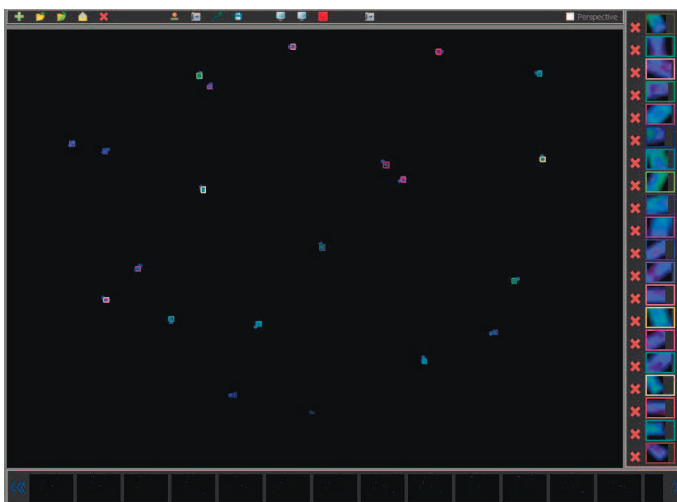


Fig. 5. GUI of the application for insect tracking.

Create project: a user can define project name and select the movie to be processed. After that, the movie is decoded into particular frames. Because FFmpeg⁸ is used for decoding, we

⁷<http://www.qt.io>

⁸<https://www.ffmpeg.org/>

can support a huge variety of movie formats and codecs. At the same time as the movie is decoded, frames thumbnails are created and saved to hard disk too. This step is time exhaustive (especially in the case of our 4k movie), but it is realized only once for the whole lifetime of the project. After that, the project can be opened using *Open project* button in a short time. When the project is loaded, a user can see the actual frame in the center of the window, and several actual thumbnails in the bottom. The thumbnails can be easily browsed using left/right buttons and after the click to a particular thumbnail, the big image is updated. In the big central image, a user can select (by a mouse drawing) patterns to be tracked. The selected patterns are visualized by a unique color and visualized in magnified form on the right side of the application window. If a pattern is not selected in a proper way, a user can click to the *delete* button and select it again.

The selected patterns can be stored using *save* button and also loaded back by *load* button. When the patterns are selected, a user can click on *process* button. With that, tracking is started, i.e., one-by-one frames are processed, stored into a hard disc and visualized in the application GUI. A user can also see actual progress in percents. The tracking process stops when one of the three following conditions is reached: 1) all patterns are processed; 2) user clicked on *stop* button; 3) application detects a possible error in tracking. In all the three cases, a user can click to *Show patterns* to see couples consisting of patterns with the same id from the first and the actual frame. This is necessary for easy comparison if some patterns have not been swapped. Also, by clicking on *Render movie*, a movie where all tracked positions are marked by the uniques colors is rendered using FFmpeg. When a user thinks that all fit, he can store positions and sizes of tracked patterns. Finally, there is button *show path* to visualize the complete paths of all patterns in one single image - the illustration of the output is shown in Figure 2.

C. Benchmark

To create fair conditions for our proposed F-transform (FT) based algorithm and other benchmarked ones, we use the same application functionalities as is described in Subsection V-B for all the algorithms, the only one thing which differs is which algorithm inside is called for tracking - if FT, KCF, MIL etc. To test the five state-of-the-art algorithms, we use OpenCV⁹ framework, where all the algorithms are implemented because, in this framework, we can be sure they are well coded. To be convinced, we tested one real-life movie where an object was set to be tracked and based on that, we can confirm that all the five algorithms work great.

From the qualitative point of view, we measured two variables - processing speed and failure rate. The processing speed is measured as the time needed for determining one single insect position; as the time needed for processing whole frame including twenty pieces of insect; and to be complete, as the estimated time to process all 27 movies. The three

measurements are dependent, we use them just to illustrate the scale of the problem. The times are measured on MacBook Air 2013 notebook and include time needed for loading images from a hard drive and storing them back. The measured times are shown in Table II. As we supposed, the fastest one is FT because it works over reduced space and realizes only one simple strong classifier which is by the principle faster than one strong classifier consisting of many weak ones. The second one is KCF - according to the original publication [5], KCF is able to process up to 100fps when one single pattern is tracked in a standard-resolution movie. Note, even if insect per second tracking seems to be quite similar to FT tracking algorithm, considering all 27 movies the difference is not negligible 121 hours of processing time. For the same reason, the rest of methods are unacceptable from the processing time point of view.

TABLE I
PROCESSING SPEED

Algorithm	Insects per sec	Frames per sec	27 movies [h]
F-transform	8.7	0.43	251
KCF	5.9	0.29	372
Boosting	3.5	0.17	635
Median Flow	3.5	0.17	635
MIL	1.0	0.05	2160
TLD	0.4	0.02	5400

The second measured variable, the failure rate is a percentage ratio how often is it necessary to stop the application (or the application has stopped itself) and manually repair the bad location. We designed test set which includes 1000 frames (i.e., we track 20000 positions in total), which were extracted from one randomly chosen movie and are time-sorted. The results are shown in Table II, where the best one - FT - achieved 15× lower failure rate than the runner-up - MIL. But in the case of MIL tracker, we have to point out there is one drawback. Even though the failure rate is pretty low, we cannot speak about precise detection, because the algorithm marks a location with a certain tolerance which results to oscillating between frames. According to that, the measured path differs from the true one. The measured length varies from 7 % to 53 % to be longer than the truth is. Such performance is not usable for further statistics processing.

TABLE II
ALGORITHM RELIABILITY

Algorithm name	Failure rate [%]
F-transform	0.005
MIL	0.075
Median flow	0.085
Boosting	> 1
TLD	> 1
KCF	N/A

The runner-up in processing speed benchmark, KCF tracker does not work in our experiment - when an insect is moving, KCF is remaining to marks the same area where the insect was originally located and it does not follow its trajectory. We

⁹http://docs.opencv.org/3.1.0/d0/d0a/classcv_1_1Tracker.html

tested the implementation on another real-life movie, where we observed it works when a big-enough area is marked to be tracked. Therefore, we tested not to mark the only insect, but also its surrounding in order to obtain bigger area. Unfortunately, such setting is not suitable because the algorithm fails when another insect came into the same area, which occurs very often.

In the case of TLD and Boosting, we tested only several frames - the algorithms failed for some of the twenty insects in almost all frames so we stopped the processing - it is not in human abilities to stop the algorithms almost every frame, fix the error and continue for the 1000 test images.

VI. CONCLUSION

In the paper, we have presented a real solved biologics problem - visual insect tracking. Because existing tested applications realizing tracking did not produce useful outputs, we design own user-friendly application allowing us to select patterns (insects) and which tracks them frame-by-frame. To realize the tracking, we developed and described own tracking core based on F-transform algorithm which serves as a strong classifier and follows idea "tracking by matching". This approach has been compared with five existing state-of-the-art algorithms (KCF, MIL, Boosting, TLD, MedianFlow) with the result that FT based algorithm is the fastest one and in the same time with the lowest failure rate. A part of a processed movie has been coded as a movie and uploaded into youtu.be/PYydVG6gjE0.

ACKNOWLEDGMENT

This research was supported by the project "LQ1602 IT4Innovations excellence in science".

We would like to thanks our colleague Michal Burda for his advice about UV light usage.

REFERENCES

- [1] R. Kays, M. C. Crofoot, W. Jetz, and M. Wikelski, "Terrestrial animal tracking as an eye on life and planet," *Science*, vol. 348, no. 6240, p. aaa2478, 2015.
- [2] M. Wikelski, R. W. Kays, N. J. Kasdin, K. Thorup, J. A. Smith, and G. W. Swenson, "Going wild: what a global small-animal tracking system could do for experimental biologists," *Journal of Experimental Biology*, vol. 210, no. 2, pp. 181–186, 2007.
- [3] A. W. Smeulders, D. M. Chu, R. Cucchiara, S. Calderara, A. Dehghan, and M. Shah, "Visual tracking: An experimental survey," *IEEE transactions on pattern analysis and machine intelligence*, vol. 36, no. 7, pp. 1442–1468, 2014.
- [4] P. Hurtik and P. Števílíáková, "Pattern matching: overview, benchmark and comparison with f-transform general matching algorithm," *Soft Computing*, vol. 21, no. 13, pp. 3525–3536, 2017.
- [5] M. Danelljan, F. Shahbaz Khan, M. Felsberg, and J. Van de Weijer, "Adaptive color attributes for real-time visual tracking," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Columbus, Ohio, USA, June 24-27, 2014*. IEEE Computer Society, 2014, pp. 1090–1097.
- [6] F. S. Khan, J. Van de Weijer, and M. Vanrell, "Modulating shape features by color attention for object recognition," *International Journal of Computer Vision*, vol. 98, no. 1, pp. 49–64, 2012.
- [7] Z. Kalal, K. Mikolajczyk, and J. Matas, "Tracking-learning-detection," *IEEE transactions on pattern analysis and machine intelligence*, vol. 34, no. 7, pp. 1409–1422, 2012.
- [8] B. Babenko, M.-H. Yang, and S. Belongie, "Visual tracking with online multiple instance learning," in *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*. IEEE, 2009, pp. 983–990.
- [9] C. Ji and S. Ma, "Combinations of weak classifiers," in *Advances in Neural Information Processing Systems*, 1997, pp. 494–500.
- [10] H. Grabner, M. Grabner, and H. Bischof, "Real-time tracking via on-line boosting," in *Bmvc*, vol. 1, no. 5, 2006, p. 6.
- [11] H. Grabner and H. Bischof, "On-line boosting and vision," in *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*, vol. 1. Ieee, 2006, pp. 260–267.
- [12] Z. Kalal, K. Mikolajczyk, and J. Matas, "Forward-backward error: Automatic detection of tracking failures," in *Pattern recognition (ICPR), 2010 20th international conference on*. IEEE, 2010, pp. 2756–2759.
- [13] B. D. Lucas, T. Kanade *et al.*, "An iterative image registration technique with an application to stereo vision," *IJCAI*, vol. 81, p. 674–679, 1981.
- [14] P. Hurtik, P. Hodáková, and I. Perfilieva, "Approximate pattern matching algorithm," in *International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems*. Springer, 2016, pp. 577–587.
- [15] I. Perfilieva, "Fuzzy transforms: Theory and applications," *Fuzzy sets and systems*, vol. 157, no. 8, pp. 993–1023, 2006.
- [16] D.-J. Jwo and S.-H. Wang, "Adaptive fuzzy strong tracking extended kalman filtering for gps navigation," *IEEE Sensors Journal*, vol. 7, no. 5, pp. 778–789, 2007.

Embedded Vision Modules for Text Recognition and Fiducial Markers Tracking

Ievgen Gorovyi
It-Jim
Kharkiv, Ukraine
ceo@it-jim.com

Vitalii Vovk
It-Jim
Kharkiv, Ukraine
ceo@it-jim.com

Maksim Shevchenko
It-Jim
Kharkiv, Ukraine
ceo@it-jim.com

Valerii Zozulia
It-Jim
Kharkiv, Ukraine
ceo@it-jim.com

Dmytro Sharapov
It-Jim
Kharkiv, Ukraine
ceo@it-jim.com

Abstract—In the paper, two examples of embedded vision modules are described. Firstly, it is demonstrated how fiducial marker tracking algorithm can be adopted for operation on Raspberry Pi. Usage of proposed ideas allows to achieve around 60fps speed of binary marker tracking. Secondly, we describe the problem of text detection and recognition in outdoor environment. Experimental results indicate on acceptable results and good potential to provide low-cost and efficient embedded vision system for this purpose. Technical details of both embedded vision modules are comprehensively discussed.

Keywords—computer vision, Raspberry Pi, fiducial markers, tracking, text recognition.

I. INTRODUCTION

Computer vision (CV) is a rapidly growing discipline making machines to percept and understand their surroundings as humans do [1]. There are a lot of practical applications of CV in medicine, industry, entertainment and many more [2]. CV algorithms can be run on different hardware: desktops, mobile phones, various digital signal processing (DSP) units. A particular interest is related with usage of low-power hardware such as Raspberry Pi [3]. Indeed, Raspberry is light weight, cheap and widely available in the market. Embedding of CV solutions transforms it into mobile autonomous intellectual system. Fig. 1 contains an example of Raspberry (Fig. 1a) and its setup with camera (Fig. 1b).

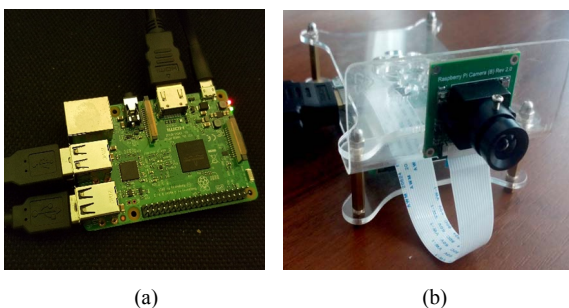


Fig. 1. Raspberry PI 3 and camera. (a) Raspberry PI Model 3B, (b) Raspberry with camera

Research community made a lot of CV experiments with Raspberry Pi. Example of multiple objects tracking can be found in [4]. In [5] a compact stereo-vision system is

described. A full stereo matching pipeline is constructed allowing to use the system as a depth-meter for outdoor scenarios. A specific use case is given in [6], when bees behavior is analyzed using embedded vision. Other examples include face recognition [7], license plates detection and recognition [8], autonomous cars applications [9], robotic assistants [10] and many more.

In the paper, we study two important problems. Firstly, it is demonstrated how to integrate and optimize fiducial marker recognition algorithm for Raspberry. Secondly, we describe the initial experimental results on number plate recognition as a separate embedded vision module.

Section II contains description of algorithm for fiducial marker recognition. Developed optimization steps are described as well. Section III contains information about text detection and recognition on Raspberry. Experimental examples are included in each section.

II. FIDUCIAL MARKERS RECOGNITION

In this section, we describe a pipeline for recognition of fiducial markers. Firstly, common steps for such process are explained. After that, some specifically developed improvements are discussed.

A. Main Steps of Marker Recognition Algorithm

Binary markers are used in many applications including robotic navigation [11]-[13], augmented reality [14] and logistics [15]-[17]. Examples of typically used binary markers are shown in Fig. 2.

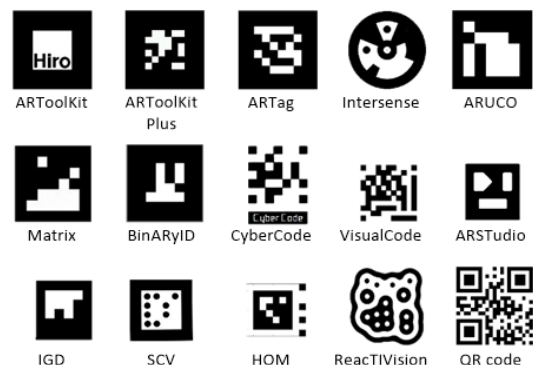


Fig. 2. Example of binary markers

The key advantage of fiducial markers usage is that we can properly build the marker structure in order to achieve high recognition rate. Also, in contrast to image markers [14], binary markers can be detected in the case more challenging geometry conditions. In our experiments, we use ArUco markers [18]. A typical marker from this collection is 7×7 blocks size containing 5×5 blocks inside (Fig. 2). A great advantage of such markers is possibility to generate a vocabulary of specific size.

A key idea of the marker recognition algorithm is to detect marker on input video frame and recognize it by comparison with previously formed database. Typically, this involves several basic steps, as shown in Fig. 3.

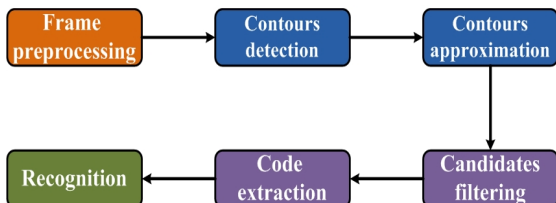


Fig. 3. Binary marker detection pipeline

Firstly, image thresholding [1] is applied to input frame. After that, contours are extracted from obtained binary image [1]-[2]. At the next step, all located contours are approximated by quadrangles. Finally, the binary codes are extracted from filtered marker candidates and then matched with existing markers database to find a proper coincidence [2].

The challenge is that the default solution is quite slow. Initial tests of performance on Raspberry give around 10 FPS. Fortunately, we have found the way to significantly boost the performance. Key proposed ideas are described in the next subsection.

B. Integration to Raspberry

It is common way to perform image thresholding before contour finding by usage of adaptive threshold (because of its robustness to image brightness changes). Since fiducial markers are binary images, such operation works as edge detector. We have found that usage of Canny edge detector [1] instead of adaptive threshold (see Fig. 4) provides better detection and code extraction quality, as well as has less computational complexity. Additionally, median blurring is applied to input image to suppress the noise.

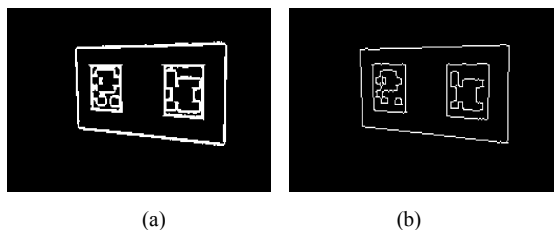


Fig. 4. Use of adaptive threshold (a) and canny edges detector (b) to prepare frame for finding contours

One of the challenges for marker recognition is existence of perspective distortions due to varying geometric conditions. In order to extract marker code, inverse perspective transformations is often applied to marker candidates. This involves calculation of corresponding transformation matrix, which is computational consuming.

In our version of the algorithm, we use improved method for code extraction that avoids the perspective transformation step. It only requires building of a grid equal to marker size directly from found contours of marker candidates, and testing pixels values on such grid (see Fig. 5).

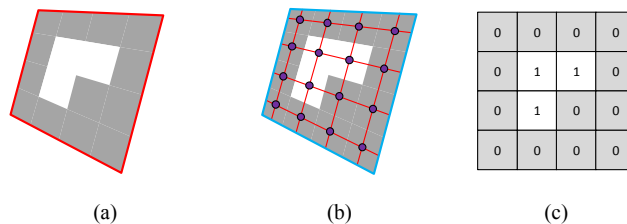


Fig. 5. Code extraction example. Found contour (a), grid building (b) and extracted code (c).

Another challenge in code extraction is presence of glare (Fig. 6a) or shadows on frame. In such cases wrong marker code extraction may appear due to incorrect thresholding parameter. For example, in Fig. 6 even dark regions of markers have a very high pixel intensity, and constant threshold leads to incorrect code extraction.

To increase the marker recognition robustness, we propose to estimate the minimum and maximum value in extracted marker image. Then threshold value is estimated as:

$$t = c \cdot (mx - mn) + mn, \quad (1)$$

where t is threshold value, mn and mx are minimum and maximum values within found contour, $c = 0.8$ is an empiric threshold coefficient.

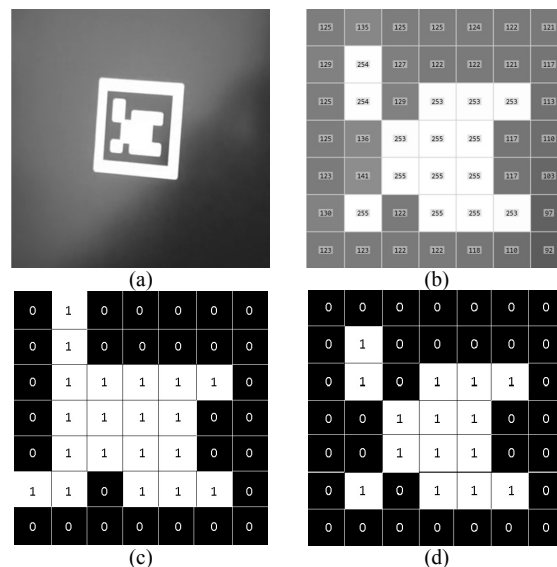


Fig. 6. Frame with marker in glare (a), extracted pixels values (b), wrong code extraction with fixed threshold (c) (value = 128), and correct code extraction with dynamically computed threshold (d) (value = 222.4)

The only requirement is that all true “white” pixels within marker region have higher intensity than any “black” one.

C. Dynamic scaling and tracking

Image pre-scaling has significant effect on frame processing time. For example, as it is shown in Fig. 7, pre-scaling input VGA frame with scale factor 0.3 reduces total

frame processing time more than 2 times with no loss in processing quality. It was found that it is possible to recognize ArUco marker even of only 30×30 pixels size for 7×7 bins. In this case, each marker bin will contain 4×4 pixels. Thus, for a particular marker we can find the scale factor make it of size 30×30 pixels and still successfully locate and recognize it:

$$scale = \frac{m}{s} \quad (2)$$

where $m = 30$ is a marker required size; s is marker current size.

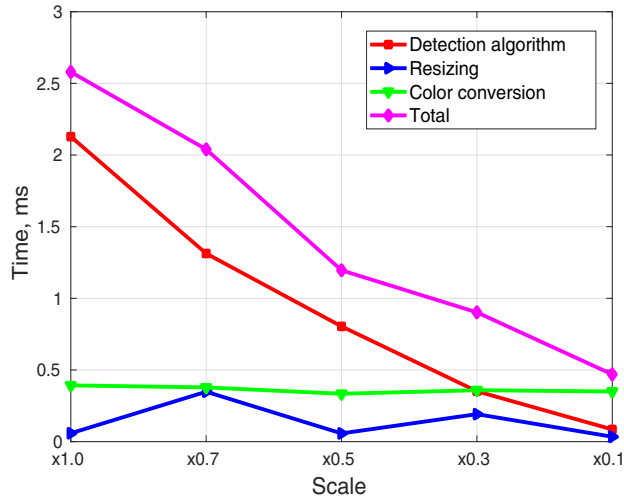


Fig. 7. Effect of pre-scaling on frame processing time.

Individual markers rescaling requires tracking of found markers and saving their last found size and position to process only small area of image or region of interest (ROI) with corresponding scale factor. Having predicted marker position and size we can build ROI around it. Because of marker position estimation error (both because of model discrepancy and marker relative velocity estimation error), found region can contain only part of the marker, or even totally miss it. In order to provide stable detection of marker, found region is extended by a precomputed value. We have found a specific empirical parameter allowing to find the required extension region

$$e = c \cdot \max(w, h), \quad (3)$$

where w and h are the current frame width and height, and extension constant $c = 0.075$. Then, ROI for analysis is equal to predicted marker's bounding box extended at each edge by value e :

$$\begin{aligned} x^* &= x - e, \quad y^* = y - e, \\ width^* &= width + 2 \cdot e, \quad height^* = height + 2 \cdot e. \end{aligned} \quad (4)$$

One should notice that additional analysis of ROI position and size should be performed to avoid violation of frame borders. Described solution of marker ROI scaling gives reasonable increase in frame processing performance (Fig. 8). One can see that tracking gives boost in performance in comparison with direct full frame analysis approach in the most cases. Only in rare cases tracking takes more time: when all markers are lost (because it requires full frame processing, but also additional resources for tracks re-initialization), and if integral ROIs' area exceeds source frame area. All of the improvements above provide increase

of FPS from 10 in initial setup to near 60 in final algorithm version (see Tab.1) with high recognition accuracy.

TABLE I. ACHIEVED FRAME RATE ON DIFFERENT STAGES OF ALGORITHM IMPROVEMENT

	Initial setup	Improved detection and code extraction	Pre-scaling	Tracking
FPS	10	15-20	35-40	~60

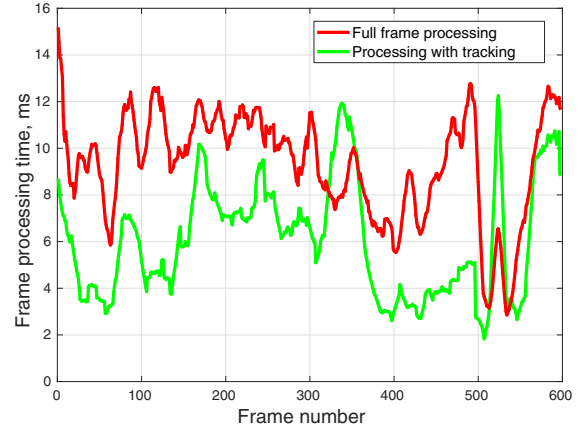


Fig. 8. Comparison of frame processing with (green) and without (red) marker tracking.

Thus, we have developed and integrated fiducial marker recognition solution on Raspberry.

III. AUTOMATIC TEXT RECOGNITION ON RASPBERRY

Another analyzed problem is related with text detection and recognition on Raspberry. For this purpose, we have made several outdoor videos with cars and tried to localize and recognize number plates. In order to perform scene text recognition, firstly we need to localize the number plate. To solve this problem, we have utilized the extremal regions (ER) detector [19]-[20]. The developed text detection pipeline is the following (Fig. 9).

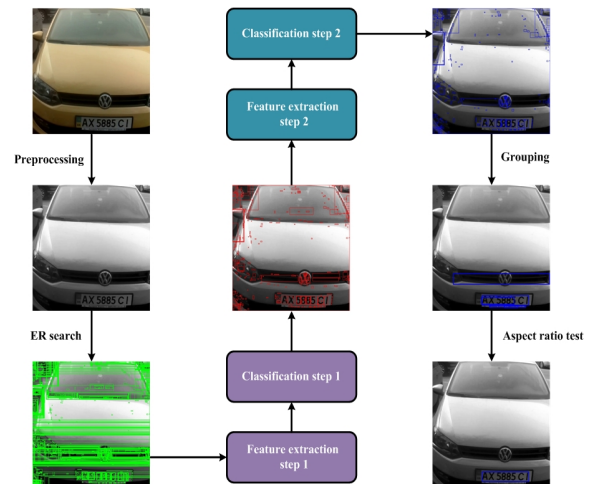


Fig. 9. Main steps of number plate detectoin algorithm

Firstly, contrast enhancement and adaptive thresholding are applied. After that a set of morphological operations is applied to filter out the noise (Fig. 10). After that letters segmentation is used. Finally, Tesseract engine is used for text recognition.



Fig. 11. Example of number plate recognition

Fig. 11 contains an example of number plate detection and recognition in Raspberry camera frames. In the near future we are planning to optimize the developed algorithm.

IV. CONCLUSIONS AND FUTURE WORK

In the paper, we considered two examples of embedded vision systems. Firstly, we comprehensively analyzed the fiducial marker recognition framework and developed several optimization steps. Experimental results indicate a significant improvement from both accuracy and speed points of view. In addition, we have shown an example of a prototype of number plate recognition system. Initial results show a good potential of production of low-cost and efficient module for intelligent transportation applications.

REFERENCES

- [1] R. Szeliski, *Computer vision: Algorithms and Applications*. London etc.: Springer, Sept, 2010.
- [2] D. Baggio, S. Emami, D. Escriva, K. Ievgen, J. Saragih and R. Shikrot, *Mastering OpenCV 3 - Second Edition*. Birmingham: Packt Publishing Ltd, Apr, 2017.
- [3] <https://www.raspberrypi.org>
- [4] A. Dziri, M. Duranton, and R. Chapuis, "Real-time multiple objects tracking on Raspberry-Pi-based smart embedded camera," *Journal of Electronic Imaging*, vol. 25(4), 2016
- [5] James Cooper et. al., "A Raspberry Pi 2-based Stereo Camera Depth Meter," *International Conference on Machine Vision Applications*, Nagoya, Japan, pp. 274-277, May 8-12, 2017.
- [6] Gang Jun Tu, Mikkel Kragh Hansen, Per Kryger, and Peter Ahrendt, "Automatic behaviour analysis system for honeybees using computer vision," *Computers and Electronics in Agriculture*, vol. 122, pp. 10-18, 2016.
- [7] R. Mo, and A. Shaout, "Portable Facial Recognition Jukebox Using Fisherfaces (Frj)," *International Journal of Advanced Computer Science and Applications*, vol. 7, no. 3, pp. 9-14, 2016.
- [8] K. Sri Sasikala, and Shakeel Ahmed, "Implementation of Number Plate Extraction for Security System using Raspberry Pi Processor," *International Journal of Engineering Research & Technology (IJERT)*, vol. 5, iss. 03, pp. 317-321, March-2016.
- [9] Gurjashan Singh Pannu, Mohammad Dawud Ansari, and Pritha Gupta, "Design and Implementation of Autonomous Car using Raspberry Pi," *International Journal of Computer Applications*, vol. 113, no. 9, pp. 22-29, March 2015.
- [10] Rizqi Andry Ardiansyah, "Design of An Electronic Narrator on Assistant Robot for Blind People," *MATEC Web of Conferences*, 42: 03013, 2016.
- [11] Rafael Munoz-Salinas, Manuel J. Marin-Jimenez, Enrique Yeguas-Bolivar, and R. Medina-Carnicer, "Mapping and localization from planar markers", *Pattern Recognition*, vol. 73, pp. 158-171, 2018.
- [12] K. Horak, and L. Zalud, "Image Processing on Raspberry Pi in Matlab," *Advances in intelligent systems and computing*, p. 25, 4 November 2015.
- [13] A. Babinec, L. Jurisica, P. Hubinsky, and F. Duchon, "Visual Localization of Mobile Robot Using Artificial Markers," *Procedia Engineering*, vol. 96, pp. 1-9, 2014.
- [14] Ievgen M. Gorovyi, and Dmytro S. Sharapov, "Advanced Image Tracking Approach for Augmented Reality Applications," *Signal Processing Symposium (SPSympo-2017)*, 12-14 September, Jachranka, Poland, pp.266-270, 2017.
- [15] Sherin M. Youssef, and Rana M. Salem, "Automated barcode recognition for smart identification and inspection automation," *Expert Syst. Appl.*, vol. 33, pp. 968-977, 2007.
- [16] C. Ozgur, C. Alias, and B. Noche, "Comparing sensor-based and camera-based approaches to recognizing the occupancy status of the load handling device of forklift trucks," *Logist. J. Proc.*, pp. 1-9, 2016.
- [17] C. Alias, C. Ozgur and B. Noche, "Monitoring production and logistics processes with the help of industrial image processing," *27th Annual POMS Conference 2016, Orlando (FL), USA, 2016*.
S. Garrido-Jurado, R. Muñoz-Salinas, F. J. Madrid-Cuevas, and M. J. Marín-Jiménez, "Automatic generation and detection of highly reliable fiducial markers under occlusion," *Pattern Recognition*, vol. 47, iss. 6, pp. 2280-2292, June 2014.
- [19] L. Neumann and J. Matas, "Real-Time Lexicon-Free Scene Text Localization and Recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 38, no. 9, pp. 1872-1885, 2016.
- [20] L. Neumann and J. Matas, "Text Localization in Real-World Images Using Efficiently Pruned Exhaustive Search," *ICDAR Proc. International Conference on Document Analysis and Recognition*, pp. 687-691, Sept, 2011.
- [21] R. Smith, "An Overview of the Tesseract OCR Engine," *ICDAR Proc. Ninth Int. Conference on Document Analysis and Recognition*, pp. 629-633, 2007.

Technology of Remote Recognition the Dart-Arrow on the Target

Roman Martsyshyn
ACS Department
Lviv Polytechnic National University
Lviv, Ukraine
mrs.nulp@gmail.com

Natalya Lysa
IST Department
Lviv Polytechnic National University
Lviv, Ukraine
lysa.nataly@gmail.com

Yulia Miyushkovych
ACS Department
Lviv Polytechnic National University
Lviv, Ukraine
jmiyushk@gmail.com

Lubomyr Sikora
ACS Department
Lviv Polytechnic National University
Lviv, Ukraine
lssikora@gmail.com

Rostyslav Tkachuk
CDCMEGP Department
Lviv State University of Life Safety
Lviv, Ukraine
Rlvtk@ukr.net

Abstract – In the work is proposed the technology of remote recognition of the image of the dart on the target. Proposed technology based on the IR sensor frame and realized as an external module.

Keywords – hit recognition, dart, target, hit detection, IR, sensor system.

I. INTRODUCTION

The problem of pattern recognition occurs in various spheres of life and needs research, taking into account the peculiarities of the nature of the recognizable image [1,3]. Often, there is a need to recognize the position of certain objects in relation to other objects. Especially often this problem arises when conducting various sporting events to determine the exact place of entry to the target or the playing field. Object recognition can be realized in many ways [5]. Some methods include mounting tracking devices in the object itself. In some cases, it is necessary to solve the task of tracking the position of an object without interfering with its design, as it may lead to deterioration of characteristics or decrease in the quality of the tracked object [4]. In the given article is considered the problem of recognition of the image of the dart on the target. The design features of the dart and the target have affected the proposed technology.

II. CHARACTERISTICS OF THE TARGET AND THE ARROWS

For a modern darts game, the targets are usually made of sisal (compressed fibers of agave). In Asia, common targets are made of horse hair. The idea of using sisal for the production of targets is owned by Nodor, and the first sisal targets appeared in 1932. The sections of the typical sisal target are separated from each other by a wire. The target is divided into sectors that are assigned numbers from 1 to 20 (Fig. 1) [2].

Sisal targets differ by the form of separation wire [2]:

- Normal (round) wire is characterized by a large percentage of dart (arrow) rebound (when arrow hit into a wire) and a low price. Used in the target of such manufacturers as: Winmau Pro SFB, Nodor Supabull II, Harrows Club.

- The triangular wire is characterized by a reduced percentage of wire rebound. When you hit the wire, the dart "go" along the border to the nearest sector. It is used in the target of such manufacturers as: Nodor Supawire, Harrows Apex Wire, Winmau Diamond.

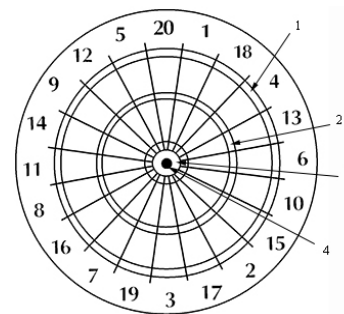


Fig. 1. Schematic view of the target with sectors (1 – "double ring", 2 – "triple ring", 3 – "Bull", 4 – "Bull eye").

- Thin split wire is used for professional purposes. Such a wire is characterized by a smaller number of dart rebounds and a high price. It is used in the target of such manufacturers as: Winmau Blade 5 Dual Core, Winmau Blade 5, Unicorn Eclipse Pro, Harrows Matrix, Nodor Supamatch II.

Standard target sizes:

- inner width of rings "double ring" and "triple ring" 8 mm.
- inner diameter of "bull's eye" 12.7 mm.
- inner diameter of the outer center ring "bull" 31.8 mm.
- the distance from the center of the target to the outer side of the "double ring" ring is 170.0 mm.
- distance from the center of the target to the outside of the wire of the ring "triple ring" 107.0 mm.
- the total diameter of the target is 451.0 mm \pm 10.0 mm.
- thickness of wire is 1.5 mm.

Dart – a special arrow for playing darts. The main parts of the dart (Fig. 2) are the tip, barrel, shank and plumage [1].

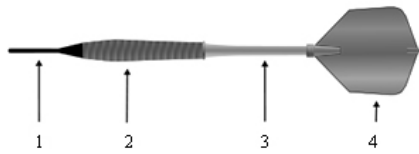


Fig. 2. Components of the dart (schematic view): 1 – tip, 2 – barrel, 3 – shank, 4 – plumage.

The dart's tip can be made in the form of a sharpened metal needle, which is designed to be stuck in the target. Also, the tip may be as a plastic needle (suitable for playing in electronic darts). The dart tip is firmly fixed in the barrel. On the other hand, the barrel includes a shank with plumage. The material and shape of the barrel determine the flight path of the dart, it is made of various metals: brass (massive cheap dart models), silver-nickel alloy (models of the middle price category), tungsten (expensive professional models).

The length of the dart should not exceed 30.5 cm. The weight of the dart should not exceed 50 grams. The most popular weight of darts is 19-25 gr.

The design dimensions of the target and the dart allow to consider several variants of placement of sensors for realization of the image recognition system (to determine dart position in the target).

III. AUTOMATIC REGISTRATION OF THE DART HIT IN TARGET - A REVIEW OF POSSIBLE SYSTEMS

A. "Electronic Darts" System

Classic darts are the most common and preferred: durable, especially if made by leading manufacturers. The playing field consists of compressed sisal fibers that are glued to the support board. Darts have a standard pointed tip that pierces the hole in the material.

In electronic darts, in contrast to the traditional darts, the construction is made of plastic lining and has a number of holes through which are able to pass special soft dart tips (Fig. 3). An electronic board is located under the playing field. Each time the tip of the arrow hits the board, the account is digitally registered on the board.

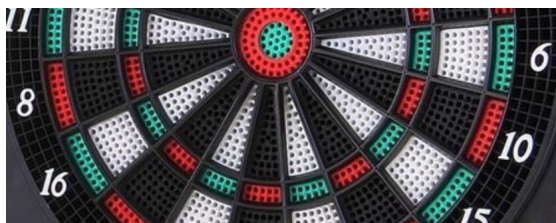


Fig. 3. Example of an electronic darts target with multiple holes (enlarged image)

Electronic darts are suitable for novice players, since the holes in the target are large, which helps to easily hit the target. Electronic darts are entertaining and not widely used for professional competitions. Using darts with a plastic tip reduces the choice of possible options for darts [2]

B. Installation of the registration system in the basis of the sisal target

It is theoretically possible to install sensors for automatic registration of the dart hit in the base of the dart target. These sensors can record the hit of the dart in the target based on pressure or other signs. The disadvantage of this approach is the short-lived gaming field of the target, which is due to the material of its manufacture.

The material from which the target is made (sisal or horse hair) gets damaged from the hit of darts under intense exploitation. This necessitates the periodic replacement of the target board.

With the intense use of the game field (targets), there is a need for frequent replacement of the target. When mounting a hit registration system at the target, the cost of making the game field will be increased. It is economically unprofitable to make a new registration system for hit and fit it into a target at each change of target.

C. Installation of the registration system in the dart arrow

Another approach to registering a dart in the target is to mount the sensor to the dart. Dart – strictly individual inventory.

Given the diversity of the range of dart arrows, there are several drawbacks in installing a fixing system in the dart. These disadvantages are critical for the use of such a system of fixation in professional sports.

Such disadvantages are:

- Increase in the mass of the dart arrow due to the additional mass of the built-in sensor;
- Possible deterioration of dart aerodynamic properties due to deflection of the center of gravity;
- Restricting the ability of players to select darts. Some professional players prefer dart arrows of a certain weight or a certain brand. Installation of the registration system in the dart prevents the use of "favorite sports equipment" and makes use of only the "sensors" darts.

IV. EXTERNAL MODULE FOR DETECTING THE HIT OF A DART IN THE TARGET BASED ON THE INFRARED SENSOR FRAME

The disadvantages of implementing a system for detecting hits in the body of the target and in the body of the dart make it impossible to implement the previously considered variants for the recognition of hits. To use sisal targets and various darts it is proposed to use an external module to fix the hit.

One possible solution to the problem may be the installation of a sensor frame around the target. The shape of the sensors frame is rectangular (square) with a side of at least $451.0 \text{ mm} \pm 10.0 \text{ mm}$ (preferably larger because of the theoretical possibility of hit outside the target).

The operating principle of the proposed sensor system for detecting hits in the target is based on the optical (infrared) touch screen (infrared grid).

A. Infrared Touch Screen Technology

Infrared touch screens are based on light-beam interruption technology. Instead of an overlay on the surface, a frame surrounds the display. The frame has light sources, or light emitting diodes (LED's) on one side and light detectors on the opposite side (Fig. 4), creating an optical grid across the screen. When an object touches the screen, the invisible light beam is interrupted, causing a drop in the signal received by photosensors [6,7]. Thus the contact coordinates are determined.

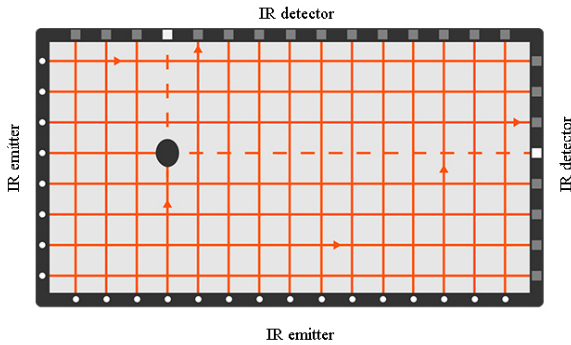


Fig. 4. Infrared touch screen technology (schematic view)

Infrared touch screens, performed in the form of an empty frame (without glass) with sensors installed in it. This constructive feature (empty frame without glass) allows you to use infrared touch screens to register the hit of a dart in the target.

B. IR Module for Detecting the Hit of a Dart

The system for detecting the dart hit the target involves the use of two layers of infrared sensors. The corresponding sensors frames are arranged parallel to each other at a fixed distance from the target (Fig. 5).

With an "ideal" perpendicular hit of a dart in the target, the definition of the position of the dart can be realized on the basis of only one infrared frame.

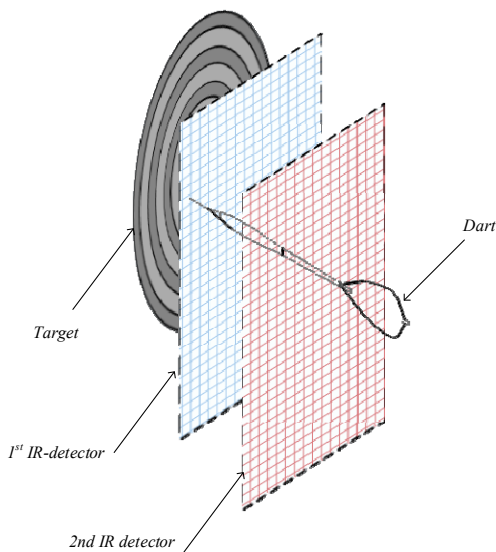


Fig. 5. The system of fixing the target hit on the basis of the IR sensor frame

Under such conditions (perpendicular hit), the coordinate of the dart on the sensor grid is projected onto the target plane.

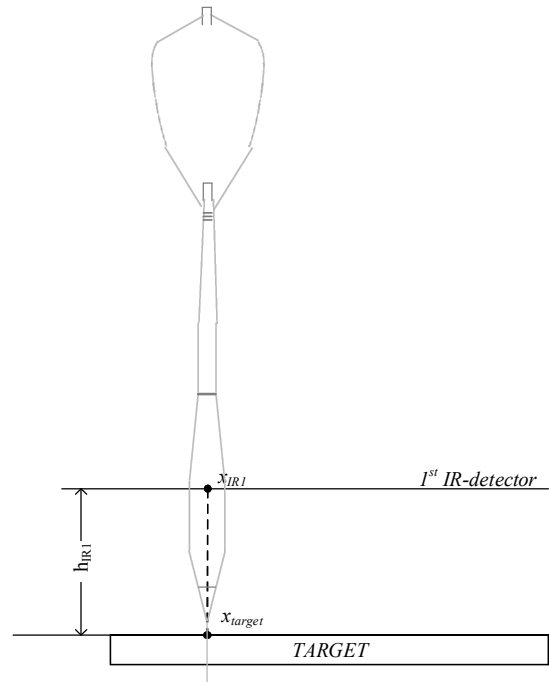


Fig. 6. "Ideal" perpendicular hit of a dart in the target

In this case (perpendicular hit) the x -coord = x_{IR} -coord, and the y -coord = y_{IR} -coord.

The need to use two infrared sensor frames is due to the fact that the arrow position relative to the target can (and usually is) not strictly perpendicular. Usually the arrow is tilted towards the target plane (Fig. 7).

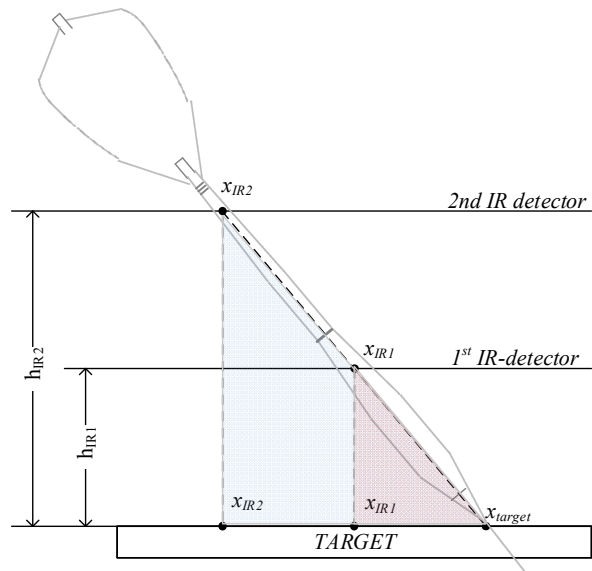


Fig. 7. The arrow is tilted towards the target plane (the most popular type of hit)

In case, described on Fig.7, we need to use some equations to determine x_{target} coordinate and the y_{target} coordinate.

The use of two layers of parallel infrared frames allows us to determine the position of the dart in the target in case of non-perpendicular hit.

From similar triangles (formed between a dart and a perpendicular line to the point of its passing through the sensors), we can determine the point of attack on the target (Fig.7).

Schematic to find the coordinate x_{target} (the x-coordinate on target) it is necessary (according to similar triangles):

$$\frac{h_{IR2}}{(x_{target} - x_{IR2})} = \frac{h_{IR1}}{(x_{target} - x_{IR1})} \quad (1)$$

where

$$x_{target} = \frac{h_{IR2}x_{IR1} - h_{IR1}x_{IR2}}{h_{IR2} - h_{IR1}} \quad (2)$$

By the same principle we can find the coordinate y_{target} (4).

$$\frac{h_{IR2}}{(y_{target} - y_{IR2})} = \frac{h_{IR1}}{(y_{target} - y_{IR1})} \quad (3)$$

where

$$y_{target} = \frac{h_{IR2}y_{IR1} - h_{IR1}y_{IR2}}{h_{IR2} - h_{IR1}} \quad (4)$$

According to the initial settings of the sensory system, the target position of the dart can be determined on the received coordinates.

V. CONCLUSION

In the work is proposed the technology of remote recognition of the image of the dart on the target. The advantages of the proposed technology are:

- Possibility to use professional models of the target with the possibility of replacing the target as needed.
- Infrared rays do not interfere with the players (they are not visible) and fix the hit.
- The possibility of fixing many hits of the dart - manufacturers offer multi-touch screens (frames) up to 32 points.
- The ability to use any dart (according to the players' tastes).
- Durability and maintainability of the fixation system.

REFERENCES

- [1] Ryan J. Tibshirani, Andrew Price and Jonathan Taylor, "A statistician plays darts," *Journal of the Royal Statistical Society Series A*, vol. 174, iss. 1, pp. 213-226, 2011
- [2] Darts [Online]. Available: <https://en.wikipedia.org/wiki/Darts>
- [3] L. Sikora, N. Lysa, R. Martsyshyn and Y. Miyushkovych, "Models of combining measuring and information systems for evaluation condition parameters of energy-active systems," *IEEE First International Conference on Data Stream Mining & Processing (DSMP)*, Lviv, pp. 290-294, 2016. doi: 10.1109/DSMP.2016.7583561
- [4] O. Riznyk, I. Yurchak and O. Povshuk, "Synthesis of optimal recovery systems in distributed computing using ideal ring bundles," *XII International Conference on Perspective Technologies and Methods in MEMS Design (MEMSTECH)*, Lviv, pp. 220-222, 2016. doi:10.1109/MEMSTECH.2016.7507545
- [5] M. Nazarkevych, B. Yavourivskiy and I. Klyuynyk, "Editing raster images and digital rating with software," *The Experience of Designing and Application of CAD Systems in Microelectronics*, Lviv, pp. 439-441, 2015. doi: 10.1109/CADSM.2015.7230897
- [6] G. Blindmann, "Multitouch technologies" [Online]. Available: <http://multi-touch-solution.com/knowledge-base-en/>.
- [7] Touch technology. Infrared [Online]. Available: <http://www.tci.de/en/service/download/touch-technology/infrared/>

Elements of RSA Algorithm and Extra Noising in a Binary Linear-Quadratic Transformations During Encryption and Decryption of Images

Anatoliy Kovalchuk
Lviv Polytechnic National University
Department of Information Technology Publishing
Lviv, Ukraine
akm0519@gmail.com

Nataliia Lotoshynska
Lviv Polytechnic National University
Department of Information Technology Publishing
Lviv, Ukraine
nsvlot@gmail.com

Abstract — An algorithm is proposed for encrypting-decrypting images using RSA algorithm elements, as the most cryptographically resistant to unauthorized decryption related to images with strictly clear edges. It is proposed to use RSA algorithm elements as coefficients of some linear-quadratic affine transformation. The proposed algorithm has a higher cryptographic stability compared to RSA algorithm.

Keywords — encryption, decryption, image, outline, cryptographic stability.

I. INTRODUCTION

An important image feature is the image edges availability. The edge separation task requires using of operations on adjacent elements sensitive to changes and obliterating regions of constant brightness levels, that is, edges are those areas where changes occur, becoming bright, while other parts of the image remain dark [2].

There are certain issues with image encryption, namely, the edges are partially preserved on sharply fluctuating images [3, 4].

The mathematically perfect edge is the rupture of brightness levels spatial function within image plane. Therefore, the edge determination means a search for the sharpest changes, that is, the maxima of the gradient vector module [2]. This is one of the reasons why edges remain within the image while encrypting in RSA system, since encryption here is based on the elevation up to the degree by modulus of a certain natural number [1]. In this case, brightness gives an even greater gap on the edge and edge adjacent pixels of raising to a degree.

As is known [5, 6], the theoretical stability is determined on the condition that there are no time limits for unauthorized decryption, and therefore it is an answer to the question that the cryptosystem cannot be split in principle. They can be constructed using a random equally probable encryption key, which key length is not less than the length of the open text. Completely stable systems are extremely expensive in implementation. Therefore, the systems are used in practice, which in principle can be split, but at an unacceptable time.

Let us assume that the image is assigned a matrix of color

$$\mathbf{C} = \begin{pmatrix} c_{1,1} & \dots & c_{1,m} \\ \dots & \dots & \dots \\ c_{n,1} & \dots & c_{n,m} \end{pmatrix}$$

Consider the next affine linear quadratic transformation, where A, B, C, D coefficients are arbitrary real numbers

$$\begin{cases} Ax + By = u \\ Cx^2 + Dy^2 = v \end{cases} \quad (1)$$

II. ENCRYPTING AND DECRYPTING BY ONE LINE OF IMAGE MATRIX

Let P and Q be a pair of arbitrary prime numbers.

Construct numbers:

$$N = PQ, \quad \varphi(N) = (P-1)(Q-1), \quad (2)$$

$$e_1 d_1 \equiv 1 \pmod{\varphi(N)}, \quad (3)$$

$$e_2 d_2 \equiv 1 \pmod{\varphi(N)}, \quad (4)$$

$$e_3 d_3 \equiv 1 \pmod{\varphi(N)}. \quad (5)$$

Encryption occurs using elements of the same line according to the following scheme:

Two consecutive values of color intensity are selected from image matrix C line (each value is selected once) and the following three values are calculated:

$$I = P^{e_1} \pmod{N}, \quad J = Q^{d_2} \pmod{N}, \quad (6)$$

$$K = (P+Q)^{e_3} \pmod{N},$$

where $e_1, e_2, e_3, d_1, d_2, d_3$ is the number, which are obtained from the relations (3) - (5) – respectively.

In (1), the coefficients are chosen $A = I, B = C = J,$

$$D = K \text{ i } x = c_{i,j}, y = c_{i,j+1}, 1 \leq i \leq n, 1 \leq j \leq m.$$

Encoded are the values $u' = u + f(i)$ and $v' = v + g(i)$, where $f(i)$ and $g(i)$ are some of the noises (u, v derived from (1)) are recorded as two consecutive values in the line of encrypted image, each value per one line.

Decryption is carried out according to the following formulas (after solving system (1) relative x, y)

$$y = \frac{\beta \pm \sqrt{\beta^2 - 4\alpha\gamma}}{2\alpha}, \quad x = \frac{u - By}{A},$$

$$\alpha = CB^2 + A^2D, \quad \beta = 2CBu, \quad \gamma = Cu^2 - A^2v$$

where $u = u' - f(i)$, $v = v' - g(i)$.

Results for $I = -1$, $C = B$, $D = K$, $P = 23$, $Q = 13$, $f(i) = Pi^2$, $g(i) = Qi^2$ are shown in Fig. 1 - Fig.3.



Fig.1 Initial Image

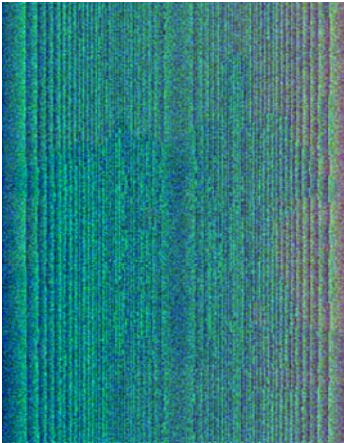


Fig.2 Encrypted image



Fig. 3 Decrypted image

III. ENCRYPTING AND DECRYPTING BY TWO LINES OF IMAGE MATRIX

In each two lines of the image matrix C the corresponding color intensity values are selected from each line of x and y . The lines are selected sequentially. Each line is selected only once. Encryption is performed similar to using one line of the image matrix by the formulas (1) – (6) with other functions of noise masking. Decryption is performed using the same formulas as in case of using one line:

$$y = \frac{\beta \pm \sqrt{\beta^2 - 4\alpha\gamma}}{2\alpha}, \quad x = \frac{u - By}{A},$$

$$\alpha = CB^2 + A^2D, \quad \beta = 2CBu, \quad \gamma = Cu^2 - A^2v.$$

Results for $I = -1$, $C = B = J$, $D = K$, $P = 23$, $Q = 13$, $f(i) = i^3$, $g(i) = i^3$ are shown in Fig. 4 - Fig.9.



Fig. 4 Initial Image

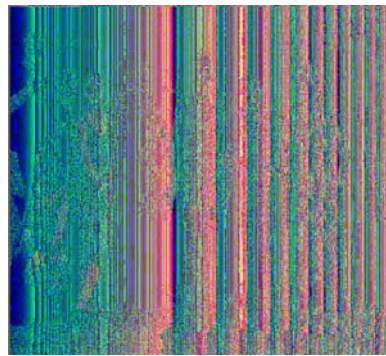


Fig. 5 Encrypted image



Fig.6. Decrypted image



Fig.7. Initial Image

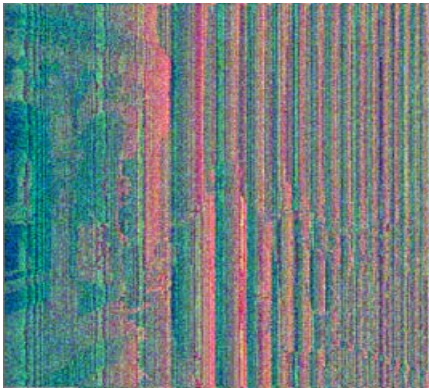


Fig. 8 Encrypted image

CONCLUSION

Comparison of Fig. 2 and Fig. 5, Fig. 8 shows that the encryption by one line of the image matrix differs from encryption by two lines of this matrix. There no edges in the encrypted images. All encrypted images are visually different. The specified algorithm can be used for the graphic images transmission.

- The proposed encryption modifications are intended for encryption of images in grayscale and are based on using the ideas of basic RSA algorithm.
- The proposed algorithms can be used for any type of image, but the greatest benefits are achieved when using the images that allow for clear edge definition.
- Both modifications can be used with no reservations for color images. However, regardless of the image type, the size of the encrypted image increases proportionally to the size of the input image.

- Resistance to unauthorized decryption in the proposed modifications is provided by RSA algorithm with additional stability, which is determined by binary transformations.
- Modified encryption methods are constructed so that at the low key values it is also possible to achieve qualitative encryption, but provided the correct selection of encryption parameters. This allows achieving the high speed of the algorithm.



Fig.9. Decrypted image

REFERENCES

- [1] Bryus Shnayer, *Prykladna kryptohrafiya*. M.: Tryumf, 2003.
- [2] B. Jähne. *Digitale image processing* (6th ed.), Springer-Verlag Berlin Heidelberg, 2005.
- [3] A. Kovalchuk, D. Peleshko, M. Navytka and T. Sviridova, "Using of affine transformations for the encryption and decryption of two images," 11th International Conference The Experience of Designing and Application of CAD Systems in Microelectronics (CADSM), Polyana-Svalyava, Ukraine, pp. 348-349, 2011.
- [4] Y. Rashkevych, A. Kovalchuk, D. Peleshko and M. Kupchak, "Stream modification of RSA algorithm for image coding with precise contours extraction," 10th International Conference - The Experience of Designing and Application of CAD Systems in Microelectronics, Lviv-Polyana, Ukraine, pp. 469-473, 2009.
- [5] M. Nazarkevych, R. Oliarnyk, O. Troyan and H. Nazarkevych, "Data protection based on encryption using Ateb-functions," XIth International Scientific and Technical Conference Computer Sciences and Information Technologies (CSIT), Lviv, Ukraine, pp. 30-32, 2016. doi: 10.1109/STC-CSIT.2016.7589861
- [6] M. Nazarkevych, R. Oliarnyk, H. Nazarkevych, O. Kramarenko and I. Onyshchenko, "The method of encryption based on Ateb-functions," IEEE First International Conference on Data Stream Mining & Processing (DSMP), Lviv, Ukraine, pp. 129-133, 2016. doi:10.1109/DSMP.2016.

Representative Based Clustering of Long Multivariate Sequences with Different Lengths

Sergii Mashtalir
Informatics dept.
Kharkiv National University
of Radio Electronics
Kharkiv, Ukraine
sergii.mashtalir@nure.ua

Volodymyr Mashtalir
Informatics dept.
Kharkiv National University
of Radio Electronics
Kharkiv, Ukraine
volodymyr.mashtalir@nure.ua

Mykhailo Stolbovyi
Informatics dept.
Kharkiv National University
of Radio Electronics
Kharkiv, Ukraine

Abstract—Video streams as unstructured or poorly structured data issue a challenge to create a unified framework capable to depict and convey high-level stories. Up-to-date indexing and search techniques to manage video data are able to operate the voluminous amounts of contained in video information in order to detect spatial and temporal events. Nevertheless, bridging semantic gap between the low-level frame or video features and high-level semantic concepts necessitates extremely high-speed procedures of temporal unlabeled data. Automatic video annotation in visual forms appears one of the promising approaches representing most pertinent and crucially important information. This goal is achieved by (among others) clustering large collections of video data.

Keywords—video stream, clustering, metric

I. INTRODUCTION

To detect spatial and temporal events in video streams as a whole, to be able to have easily understandable visual abstracts of video collections, it is essential to have the means for creation of summaries in forms of video skims (frame sequences composed with excerpts of the original video) and key frames (the most semantically typical frame or several ones from the consecutive image sequence) sets. Summarizing video content is substantial for at least such applications [1,2] as video browsing archiving content based retrieval, access to teleconferences, video mail and video news, etc. It should also be noted that advanced tools for film trailers making, generating of TV programs previews and other similar treatments to quick overview of the content can not be simple reductions of the original. Video processing requires to keep, at least, the restricted inquiries content aspects and to reduce the semantic gap between the features extracted from images (or/and their time sequences) [3,4].

There arise remarkable achievabilities in video processing approaches based on specific feature integration via CNN (Convolutional Neural Networks) to extend, in some measure, capabilities of task-dependent and scene-dependent explanations of visual spatial and temporal events to index, store, linear and nonlinear edit, retrieve and present video records and excerpts in convenient mode [5].

Nevertheless, promising way of solving video streams analysis problems remains in temporal (pre-, on line, post) processing of multivariate time series. Such series are induced by frame sequences in various ways. First of all, it is necessary to allocate time series in terms of local and global features concerning color, texture, shape extracted from each frame. In the second, it is it should be emphasized time

series discovered in matrix form which are agreed with one or more subframes covering field of view or surrounding some content associated localities in the image. Local features may be related with edge, corner, ridge, blob detection, the histogram of oriented gradients (HOG), the gradient location and orientation histogram (GLOH), the scale-invariant feature transform (SIFT), the speeded up robust features (SURF) etc. [1,6].

It is absolutely recognizable that perceptual process is closely related with salient regions of the visual scene. To consider all the visual impressions contained in videos it is required to take into most complete account useful for human attention concept with the object of perceptual video complexity reduction and support event decision on the base of the relevant parts of sensory input selections. The ability to extract the relevant slices of an image is of large interest, especially if each frame from video stream is unavoidable processed in real time. So, prevalent belief, that visual attention models are useful starting-point for arbitrary type of content based video parsing, predefines objects of an automatic frame by frame analyze. In field of view such objects can be represented as some significant feature points and their neighborhoods [7].

The points that satisfy the properties of repeatability distinctiveness, stability, uniqueness, interpretability are usually called saliency points [6,8]. The saliency map can be elucidated as spatially arranged point collection that provides semantic stability of a visual scene. Moreover, temporal proximity assumption ensures that the salient points of the current frame are localized in windows surrounding the locations of each previously found salient point. Thereby, any saliency model based on spatiotemporal salient points should predict what attracts the (human) attention and sequence of saliency maps or window sets, covering a subset of salient points, produce a multivariate time series what gives serious reason to get brief valid video summary via temporal clustering.

Clustering of time series data induced by video streams, first and foremost, is aimed to an exploration of video tagging. Shots and key frames, as a rule, are adapted as units for excerpts labeling when segmenting video and searching for sets of similar temporal fragments, desirably, with nearly equivalent content [4,9]. If any known clustering algorithms can be applied when using key frames (at least in feature spaces) [10, 11], then the clustering of time series is associated with a number of difficulties, the main one of which lies in different lengths of the sequences to be processed [12]. Thus, techniques advancement to get a quick

video overview is remained profound challenge.

The remainder of the paper is structured as follows. The second section is devoted to available capabilities destined for different lengths streams matching and additional refinement of the series under consideration. The third section proposes a clustering technique providing similar temporal segment of video grouping. Further, the results of video data clustering are discussed and the tasks to be solved in the future are defined.

II. PRELIMINARY REMARKS AND STATEMENT OF THE PROBLEM

To estimate the similarity of two multidimensional time series induced by video streams, one can use a modification of the popular DTW (Dynamic Time Warping) method [11, 13], which was, by and large, widely used to estimate the distance between arbitrary sequences of different lengths. The essence of it is as follows. Introduce two multidimensional sequences $X = \{x(1), \dots, x(k), \dots, x(M)\}$ and $Y = \{y(1), \dots, y(l), \dots, y(N)\}$, $N \neq M$ where $x(k)$ and $y(l)$ can be represented either vectors $x(k) = (x_1(k), \dots, x_n(k))$, $y(l) = (y_1(l), \dots, y_n(l))$ in any feature space $\Omega \subset \mathbb{R}^n$ or matrices (subframe set for each frame from video stream) $x(k) = \{x_{i_1 i_2}(k)\}$, $y(l) = \{y_{i_1 i_2}(l)\}$, $x(k), y(l) \in \mathbb{R}^{h \times v}$ mostly in the image space. Next, $(N \times M)$ matrix of distances in the chosen local (element-wise) metric with elements $d(x(k), y(l))$, $k = \overline{1, N}$, $l = \overline{1, M}$ between all elements of the sequences to be matched is introduced into consideration. On the basis of this matrix a warping path is constructed as distance sequence $W = \{w_1, w_2, \dots, w_q, \dots, w_L\}$, $w_q = d(x(k), y(l))_q$, $\max\{N, M\} \leq L \leq M + N - 1$, $q = \overline{1, L}$ which, in fact, determines the similarity between X and Y on the basis of the accumulative distance

$$D(k, l) = d(x(k), y(l)) + \min\{D(k, l-1), D(k-1, l), D(k-1, l-1)\}. \quad (1)$$

It should be emphasized that the resulting warping path, generally speaking, is a proximity measure, but not a metric.

As the distance between local elements of X and Y, the Euclidean metric is usually used

$$d(x(k), y(l)) = \|x(k) - y(l)\|_2 \quad (2)$$

which, when processing subframes of images from video, takes the form of the Frobenius norm metric

$$d(x(k), y(l)) = (\text{Sp}(x(k) - y(l))(x(k) - y(l))^T)^{\frac{1}{2}}. \quad (3)$$

In situations where the processed data are noisy and have outliers, it is appropriate to use the Manhattan metric, which has some robust properties. In this case, (2) corresponds to expression

$$d(x(k), y(l)) = \sum_{i=1}^n |x_i(k) - y_i(l)|, \quad (4)$$

and for (3) we have

$$d(x(k), y(l)) = \sum_{i_1=1}^n \sum_{i_2=1}^v |x_{i_1}(k) - y_{i_2}(l)|. \quad (5)$$

Now turn to the clustering problem. Introduce a set of multidimensional time series $X_1, X_2, \dots, X_q, \dots, X_Q$ that must be grouped into clusters. It is assumed that each of the clusters contains a different number of observations, $N_1, \dots, N_q, \dots, N_Q$ respectively. Direct clustering of the original sequences seems to be ineffective, since calculation (1) is based on dynamic programming on long series produced by video streams and such clustering is associated with high computational complexity. In addition, the processed sequences are, as a rule, nonstationary, i.e. different segments can belong to different classes.

In such situations, it is possible to take advantage of the window-type approach, when each series X_q , $q = 1, 2, \dots, Q$ is divided into P window sections, resulting in a set of new series $WX_{q1}, X_{q2}, \dots, WX_{qp}, \dots, WX_{qP}$, $p = 1, 2, \dots, P$, which are further considered as independent sequences to be clustered. It is not difficult to understand that if the windows WX_{qp} of one sequence X_q fall into different clusters, which indicates that the original sequence is nonstationary, then the resulting segments can be classified as signals independent of each other.

Thus, QP signals are introduced, each of which contains samples, and the ultimate goal is a partition of these series in the self-learning mode in homogeneous in the sense of (1) – (5) classes.

III. CLUSTERING BASED ON REPRESENTATIVES

Among the known clustering algorithms, the most widely explored ones are based on prototype-centroids due to the simplicity of computational models and the interpretability of the results obtained [10,11]. However, these methods are of little use for clustering multidimensional time series produced by video streams, since the processed patterns in the known approaches have the same dimension, but subject to video processing the center of attention is different lengths of the series: WX_{qp} has N_{pq} elements.

In such situations, approaches connected with sample exemplary usage may be more preferable when instead of the computed prototype-centroid one of the vectors (matrices), available in the processed sample $WX_{11}, \dots, X_{1p}, \dots, WX_{21}, \dots, WX_{qp}, \dots, WX_{QP}$ with QP temporal segments, is selected. Consider such a clustering procedure.

The algorithm functioning begins with the selecting of the initial representatives $WX_{qp}^1(0), WX_{qp}^2(0), \dots, WX_{qp}^m(0)$ where m is the given number of clusters. Here, it is important to emphasize that the choice of m would not be productive without deep preliminary semantic analysis of the video parsing goals. In the capacity of $WX_{qp}^1(0)$, the sample component furthest from all others is chosen, i.e. $\forall q, p \in \{1, 2, \dots, Q\}$, $\forall r, s \in \{1, 2, \dots, P\}$

$$DTW(WX_{qp}^1(0), WX_{qp}(0)) > DTW(WX_{rs}, WX_{qp}).$$

Further, $WX_{qp}^1(0)$ is temporarily excluded from the sample and the second representative is selected, as before, the most distant from the elements left in the sample, i.e.

$$DTW(WX_{qp}^1(0), WX_{qp}(0)) > DTW(WX_{qp}^2(0), WX_{qp}(0)) > DTW(WX_{rs}, WX_{qp}).$$

This selection procedure (with at each step the exception of the already selected representatives, being the starting centers of clusters) is repeated m times till to the formation of initial patterns such that

$$DTW(WX_{qp}^1(0), WX_{qp}(0)) > \dots > DTW(WX_{qp}^m(0), WX_{qp}(0)) > DTW(WX_{rs}, WX_{qp}).$$

This procedure for selecting initial patterns is effective if the original data do not contain outliers. Otherwise, the initial representatives can be randomly chosen, just as it happens in greedy clustering algorithms, e.g. in the classical k -means. With respect to the video streams processing, such outliers can be, e.g. temporal segments of inserted advertising products, and clustering with such comparable attractors is semantically meaningless. In other words, the preprocessing of sequence $WX_{11}, \dots, X_{1p}, \dots, WX_{21}, \dots, WX_{qp}, \dots, WX_{QP}$ passes into meaningful procedures.

In the second stage, the remaining $QP - m$ patterns are allocated over clusters Cl_j , $j = \overline{1, m}$ according to the relation: $\forall j \neq l \in \{1, 2, \dots, m\} WX_{qp} \in Cl_j$ if

$$DTW(WX_{qp}, WX_{qp}^j(0)) < DTW(WX_{qp}, WX_{qp}^l(0)).$$

Thus, all available patterns have been collected in the neighbourhoods of each of initial representatives $WX_{qp}^j(0)$.

At the third stage in each of the groups formed, a new representative is determined, which selects as the observation with the minimum total distance to all points of the initial cluster. In other words, for the improved representative for all admissible q, p, r, s, j relations

$$\sum DTW(WX_{qp} \in Cl_j, WX_{qp}^j(1)) < \sum DTW(WX_{qp} \in Cl_j, WX_{rs} \in Cl_j).$$

have to be hold.

After discovery m refined representatives $WX_{qp}^j(1)$, $j = \overline{1, m}$ the return to the second stage takes place, where the patterns are attributed to the newly formed patterns $WX_{qp}^j(1)$ according to the rule:

$$WX_{qp} \in Cl_j \text{ if } DTW(WX_{qp}, WX_{qp}^j(1)) < DTW(WX_{qp}, WX_{qp}^l(1)) \forall j \neq l \in \{1, 2, \dots, m\}.$$

This process is repeated until all representatives cease to change, i.e. for all j from 1 to m the stabilization condition $WX_{qp}^j(\alpha+1) = WX_{qp}^j(\alpha)$ is fulfilled, where $\alpha = 0, 1, 2, \dots$ is the iteration number.

The original series X_q are obtainable from WX_{qp} at the end of the procedure. If for any q all elements belong to the same cluster, it means that the series X_q is stationary and does not change properties in the time interval from 1 to N_q . Otherwise, we have that in such segments there have been changes in the properties of the time series, and a more detailed analysis of the corresponding segments of the series is required.

Turning to the data under examination viz to video streams, one significant for applications detail must be noted. Plausible content-identification is required when shots are split up separated by different tricks e.g. gradual effects (fade, dissolve, wipe) and inheritances of lens distortion, magnify blur and sharpen as well as temporal smoothing, shadow/highlight effects. Due to the difficulties and challenges of such insignificant information for abstracting video (but not for segmentation of time series), the need for a more acceptable content based filtering has arisen. There exists sufficient ground to delete video transitions at clustering stage and serious attention should be paid to the detection shots equiprobably belonging to two consecutive segments, which are obtained by editing the video and in fact are obstacles to the content analysis.

IV. RESULTS AND DISCUSSIONS

Associated with video clustering experiments were conducted based on 40 episodes "Destroyed in seconds" (Discovery Channel) with time length about 22 minutes each, with a resolution of 688×422 and a frame rate of 25 per second. The peculiarity of the video data used is the sufficiently rich availability of trailers that were manually excluded from further processing as well as the transitions between consecutive pairs of shots that seems quite reasonable for maintaining the adequacy of delivering the core information. Previously, corresponding to the episode time series, produced by simplest shape features of each frame segmentation, was segmented on the base of spatio-temporal approach [14]. In a number of cases, this temporal segmentation was repeated for already found shots (if their length exceeded 1500 frames).

Fig. 1 illustrates a typical shot, the aggregate of which generates multidimensional time series to be clustered. Fig. 2 shows the final representative-shot, which is used as an element in summarizing sequence.

Because of the great length and rich content of video data clustering can be repeated for the found representative-shots until the required duration of result (given or found empirically) will be obtained. At this point one important detail must be emphasized, substantial difficulties remain in the construction of semantic structure.

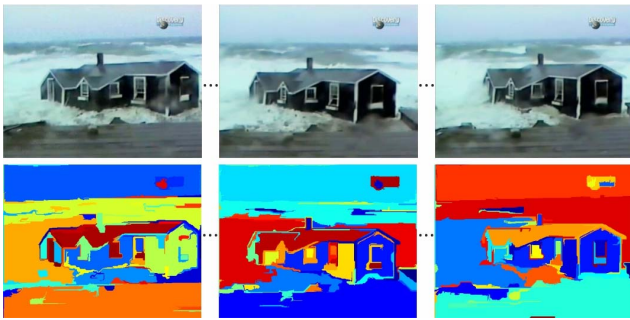


Fig. 1. Example of a shot (with usage shape features).

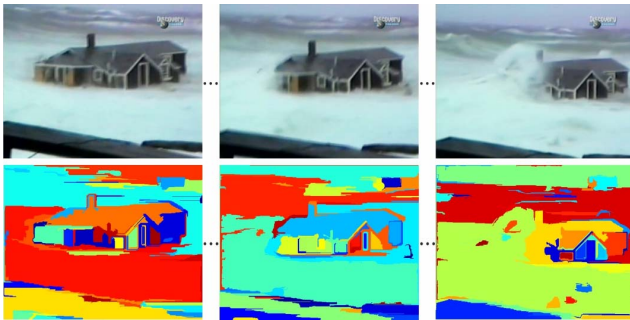


Fig. 2. Example of a key shot as clustering representative.

Similar results, but with some semantic deterioration and a noticeable reduction in computational complexity, were obtained by temporal segmentation of matrix pixel series and clustering of subframe set sequences, examples of which are shown in Fig. 3. To further elucidate clustering results, it is necessary to consider disjoint matrix sequences in time separately and together.

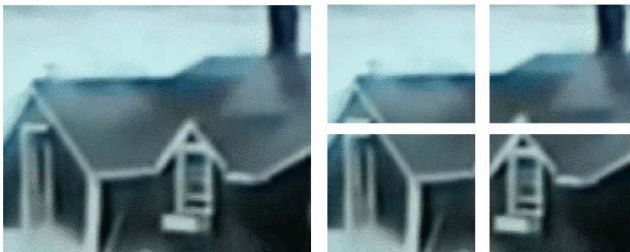


Fig. 3. Subframe examples for video segmentation and clustering.

Seen as a whole, the results obtained highlight the potential of the clustering approach to video summarizing. A few words of comment are necessary here about DTW. DTW opens up a whole range of new opportunities to match time series with different lengths and under certain circumstances fully valid comparisons take place. However, if there exists a considerable difference in the sequences lengths, there can be a peculiar averaging when a point (frame) from a sequence is associated with a significant number of points (frames) from

another sequence. It should also be noted that the proposed approach operates with a known number of clusters, which can not be a priori determined under large video streams processing and there are no clear ways of reasonable choice.

Summing up, it can be argued that the main findings of the study reveal a clustering procedure intended to analysis of long multivariate sequences (including matrix type series) with different lengths. Further investigation will stimulate an understanding of rational detection and application of content based saliency maps with possible backward analysis of partly other regions of visual attention.

REFERENCES

- [1] C. Liu, Recent Advances in Intelligent Image Search and Video Retrieval. Intelligent Systems Reference Library, vol. 121, Cham: Springer, 2017.
- [2] G. Csurka, Domain Adaptation in Computer Vision Applications. Advances in Computer Vision and Pattern Recognition, Cham: Springer, 2017.
- [3] B. T. Truong, and S. Venkatesh, "Video abstraction: A systematic review and classification," *ACM Trans. on Multimedia Computing, Communications, and Applications*, vol. 3, iss. 1, pp. 1–37, 2007.
- [4] S. Mashtalir, and O. Mihnova, "Key frame extraction from video: framework and advances," *Int. J. of Computer Vision and Image Processing*, vol. 4, iss. 2, pp. 68–79, 2014.
- [5] T. Wiatowski, and H. Bölskei, "A mathematical theory of Deep Convolutional Neural Networks for feature extraction." *IEEE Trans. on Information Theory*, vol. 64, iss. 3, pp. 1845–1866, 2018.
- [6] F. Shih, *Image Processing and Pattern Recognition: Fundamentals and Techniques*, Hoboken: John Wiley & Sons, Inc., 2010.
- [7] L. Elazary, and L. Itti, "Interesting objects are visually salient," *J. of Vision*, vol. 8, iss.3, pp. 1–15, 2008.
- [8] D. Liu, G. Hua, and T. Chen "A hierarchical visual model for video object summarization," *IEEE Trans. on PAMI* vol. 32, iss. 12, pp. 2178–2190, 2010.
- [9] Ye. Bodyanskiy, D. Kinoshenko, S. Mashtalir, and O. Mikhnova, "On-line video segmentation using methods of fault detection in multidimensional time sequences," *Int. J. of Electronic Commerce Studies*, vol. 3, iss. 1, pp. 1–20, 2012.
- [10] C.C. Aggarwal, *Data Mining*. Cham: Springer, 2015.
- [11] G. Gan, C. Ma, and J. Wu, *Data Clustering: Theory, Algorithms, and Applications*, ASA-SIAM Series on Statistics and Applied Probability. Philadelphia : SIAM, VA, 2007.
- [12] Z. Hu, S. V. Mashtalir, O. K. Tyshchenko, and M. I. Stolbovyi "Video shots' matching via various length of multidimensional time sequences," *Int. J. of Intelligent Systems and Applications (IJISA)*, vol. 9, iss. 11, pp.10–162, 017.
- [13] X. Wang, A. Mueen, H. Ding, G. Trajcevski, P. Scheuermann, and E. Keogh, "Experimental comparison of representation methods and distance measures for time series data," *Data Mining and Knowledge Discovery*, vol. 26, iss. 2, pp. 275–309, 2012.
- [14] S. Mashtalir, and V. Mashtalir, "Sequential temporal video segmentation via spatial image partitions," *IEEE First Int. Conf. on Data Stream Mining and Processing (DSMP'2016)*, Lviv, Ukraine, pp. 239–242, 2016.

Sequence Matching for Content-Based Video Retrieval

Sergii Mashtalir
Informatics Department
Kharkiv National University of Radio
Electronics
Kharkiv, Ukraine
sergii.mashtalir@nure.ua

Olena Mikhnova
Cibernetics Department
Kharkiv Petro Vasylenko National
Technical University of Agriculture,
Kharkiv, Ukraine
elena_mikhnova@ukr.net

Mykhailo Stolbovyi
Informatics Department
Kharkiv National University of Radio
Electronics
Kharkiv, Ukraine
d_inf@nure.ua

Abstract— In this paper the authors propose a novel technique for comparing video frame sequence presented in an arbitrary metric space. By reviewing existing best practices in spatio-temporal video segmentation and frame matching, the authors suggest mathematical grounding for efficient video content analysis. Variants of relationships are observed between the frame sequences under comparison (perfect match, inclusion, equality of cardinality of sets). Examples of application as well as estimation metrics are also provided.

Keywords— Video Content Matching, Spatio-Temporal Segmentation, Set Theory, Metric Space

I. INTRODUCTION

Great diversity of artificial intelligence problems emerged during the last two decades. Researchers around the globe are trying to streamline people's activities by introducing contemporary methods and techniques that aid machines in decreasing human mental workload. Multimedia processing is among leading areas of research and development in numerous companies. Motorola Multimedia Research Lab, IBM Research, FX Palo Alto Laboratory, Google, just to name a few.

In this paper we consider frame sequence matching which may turn out quite complicated for a machine because of blends and dissolves that make color and texture changes almost impossible to track under some visual conditions [1-3]. The concept of video segmentation and frame matching is also typical for any frame extraction procedure. In video processing, the closer the frames are to each other in terms of some metric, the harder is to pick up a boundary between them. The nature of segments and segmentation process itself is crucial for successful matching. The simplest way of segmenting video content is dividing it into fragments of equal length. Despite it may reduce time needed for processing for more than a half part, of course it is not a perfect idea of segmentation as a scene may appear in consecutive segments or several scenes may be contained in one segment. A strong post-processing is needed after such a temporal segmentation, which is very hard to ensure [4, 5].

Inter-frame difference (for cuts) and skipping frame difference (for dissolves) are considered to be the main sources for spatio-temporal segmentation. It can be measured by dissimilarity of pixels, frame blocks, or the whole frames. Among color feature comparison techniques, histogram difference gains most popularity and simplicity. Most of the current methods can cope with both kinds of inter-frame transitions (cuts and dissolves). When cuts occur, a method should detect changes in two consecutive frames, while for

dissolves a method should analyze a number of consecutive frames to detect a new scene. Dissolves are harder to detect with traditionally used color-based methods only (or skipping frame difference should be used), but a good method should distinguish both kinds of transitions at the same video, as no one knows editor's plans of video organization [4, 6].

More than fifty spatio-temporal segmentation and frame matching techniques are briefly observed in [7-15]. Such a large number approaches to the video segmentation and the shot detection indicate, on the one hand, the interest of scientists in these methods development, and on the other, the need to develop new ones, because there are no universal approaches suitable for analyzing arbitrary video data. The main problem they face lies in a variety of video content genres, without mentioning object and camera motion, flashes and other changes in lighting conditions. Most of the available methods take into account frame difference, without paying great attention to content which changes in time. Fig. 1 below details how frame matching techniques are distributed according to their popularity. The most widely used techniques that remain fundamental parts of the most successful approaches turn out to be color histograms and machine learning. Other techniques such as detecting camera flashes or working only in the compressed domain are not yet of widespread applicability [4, 7].

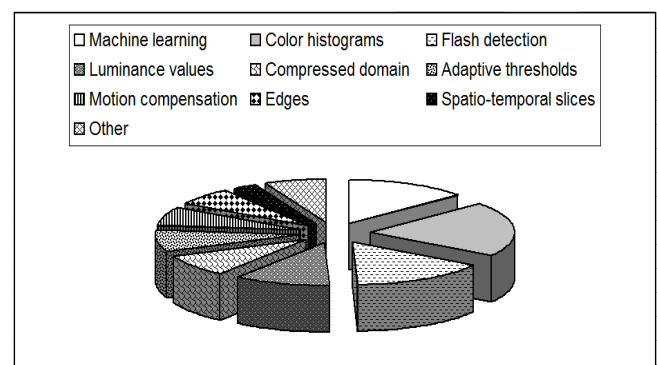


Fig. 1. Pie chart of the most popular frame matching techniques

It is important to figure out how video frames should be compared to each other to determine exact and fuzzy matches. The next section outlines the novel unique model proposed for such purposes. It has been already mentioned that a good spatio-temporal segmentation is delimited by fades and wipes along with lighting conditions, camera angle change, etc. With this in mind, such a model should be

constructed that solves the aforementioned problems from mathematical and applied points. The model should also comply with existing standards along with contemporary level of video feature presentation and processing. After introducing the model, the latter section of the paper testifies its application significance and provides measures for estimation.

II. VIDEO SEGMENT MATCHING MODEL

Consider an arbitrary metric space Ω with a specified metric $\rho(x, y)$ where $x, y \in \Omega$. In addition, assume a set F which elements are finite sequences of elements from Ω . In fact, the set F represents temporal segment in video Ω , which is a tuple of frames x_1, x_2, \dots, x_n . To be more precisely, this means $\bar{x} = (x_1, x_2, \dots, x_n) \in F$ if $x_1, x_2, \dots, x_n \in \Omega$, n is an arbitrary non-negative integer that is greater than zero, and the order of elements x_1, x_2, \dots, x_n is essential. In other words, they cannot be rearranged without changing the element from the set F . In this case, such sequences will be called tuples. Thus, the set F is constructed from element tuples of the metric space Ω . Then, assume the following pattern on the set F . Introduce the notion of distance matrix for the tuple pair.

Definition 1. Suppose matrix $A(\bar{x}, \bar{y})$ is the distance matrix for the pair of elements $\bar{x}, \bar{y} \in F$, which is constructed in the following way:

Compare $card(\bar{x}) = n$ and $card(\bar{y}) = m$. Suppose the first tuple is less or equal to the second one $n \leq m$, then the first row of distance matrix $A(\bar{x}, \bar{y})$ looks as follows: $\rho(x_1, y_1) \rho(x_2, y_2) \dots \rho(x_n, y_n)$. The second row looks like: $\rho(x_1, y_2) \rho(x_2, y_3) \dots \rho(x_n, y_{n+1})$, etc. Then, s -th row is $\rho(x_1, y_s) \rho(x_2, y_{s+1}) \dots \rho(x_n, y_{s+n-1})$. Consequently, the number of rows in $A(\bar{x}, \bar{y})$ is equal to $s = m - n + 1$, and the matrix itself looks as follows:

$$A(\bar{x}, \bar{y}) = \begin{pmatrix} \rho(x_1, y_1) & \dots & \rho(x_n, y_n) \\ \rho(x_1, y_2) & \dots & \rho(x_n, y_{n+1}) \\ \dots & \dots & \dots \\ \rho(x_1, y_s) & \dots & \rho(x_n, y_{s+n-1}) \end{pmatrix}. \quad (1)$$

The size of the above matrix is $s \times n$, taking into consideration that $n \leq m$, $s = m - n + 1$, $n = card(\bar{x})$, $m = card(\bar{y})$. Assume the following properties of the matrix $A(\bar{x}, \bar{y})$.

Property 1. If the distance matrix contains a zero row, then it means that the smaller tuple (in terms of cardinality) is fully included somewhere in the bigger one. In this case $\bar{x} \subset \bar{y}$.

Property 2. If the total number of elements in the first and the second tuple are equal to each other $card(\bar{x}) = card(\bar{y})$, then the distance matrix $A(\bar{x}, \bar{y})$ is constructed from a single row:

$$A(\bar{x}, \bar{y}) = (\rho(x_1, y_1), \dots, \rho(x_n, y_n)). \quad (2)$$

Property 3. When the two tuples are fully equal to each other $\bar{x} = \bar{y}$, then the distance matrix $A(\bar{x}, \bar{y})$ is constructed from a single zero row, and it looks like this:

$$A(\bar{x}, \bar{y}) = (0, \dots, 0). \quad (3)$$

Property 4. $A(\bar{x}, \bar{y}) = A(\bar{y}, \bar{x})$ for $\forall \bar{x}, \bar{y} \in F$.

All of the above properties follow from the definition of the distance matrix and from the fact that $\rho(x, y)$ is initially a metric. Now, consider a series of functionals for the set $F \times F$, i.e. for its Cartesian square. Suppose $\bar{x}, \bar{y} \in F$ and $card(\bar{x}) = n \leq m = card(\bar{y})$, then the following functionals are correspondent with them:

$$\begin{aligned} g_1(\bar{x}, \bar{y}) &= \sum_{i=1}^n \rho(x_i, y_i), \\ g_2(\bar{x}, \bar{y}) &= \sum_{i=1}^n \rho(x_i, y_{i+1}), \\ &\dots \\ g_s(\bar{x}, \bar{y}) &= \sum_{i=1}^n \rho(x_i, y_{i+s-1}) \end{aligned} \quad (4)$$

where $s = m - n + 1$. The following theorem is held.

Theorem 1. Each functional specified by the equations (4) is a metric on the set F .

The above theorem is easily proved by ensuring reflexivity, symmetry and triangle inequality.

Reflexivity. If the two tuples are fully equal to each other $\bar{x} = \bar{y}$, then from the Property 3 it follows that $A(\bar{x}, \bar{y}) = (0, \dots, 0)$, i.e. there exists $g_1(\bar{x}, \bar{x}) = 0$.

Symmetry. Symmetry apparently follows from the Property 4.

Triangle inequality. Triangle inequality can be explained using the example $g_1(\bar{x}, \bar{y})$. Suppose there are three tuples $\bar{x}, \bar{y}, \bar{z}$. The following Fig. 2 illustrates these in a schematic manner.

Then, it is clear that:

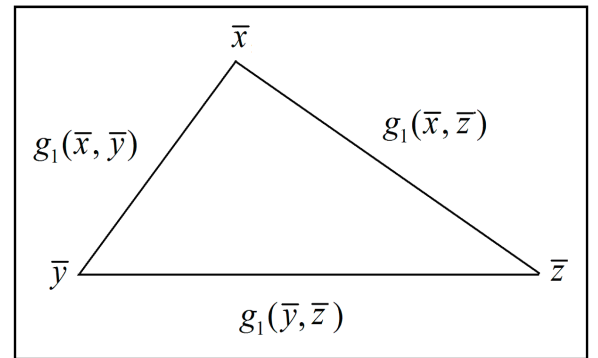


Fig. 2. Triangle inequality for the three tuples

$$\begin{aligned}
g_1(\bar{x}, \bar{y}) &= \sum_{i=1}^n \rho(x_i, y_i), \\
g_1(\bar{x}, \bar{z}) &= \sum_{i=1}^n \rho(x_i, z_i), \\
g_1(\bar{y}, \bar{z}) &= \sum_{i=1}^n \rho(y_i, z_i)
\end{aligned} \quad (5)$$

where $card(\bar{x}) = n$, $card(\bar{y}) = m$, $card(\bar{z}) = k$ and $n \leq m \leq k$. For the sake of certainty, consider $g_1(\bar{x}, \bar{y}) + g_1(\bar{x}, \bar{z})$, then the following equation is held from (5):

$$g_1(\bar{x}, \bar{y}) + g_1(\bar{x}, \bar{z}) = \sum_{i=1}^n [\rho(x_i, y_i) + \rho(x_i, z_i)]. \quad (6)$$

As $\rho(x, y)$ is a metric, then the following is held for any i :

$$g_1(\bar{x}, \bar{y}) + g_1(\bar{x}, \bar{z}) = \sum_{i=1}^n [\rho(x_i, y_i) + \rho(x_i, z_i)]. \quad (7)$$

By taking equation (6) into account, the following can be obtained:

$$g_1(\bar{x}, \bar{y}) + g_1(\bar{x}, \bar{z}) \geq g_1(\bar{y}, \bar{z}). \quad (8)$$

The other two couples of summands needed in the triangle inequality may be proved the same way, i.e. $g_1(\bar{x}, \bar{y})$ is a metric. Now, we shall explain why this theorem is held for all the other functionals in (4). With this, we understand it in case of their existence for particular tuple cardinalities. The following Fig. 3 shows this relation in a form of a schema.

The number of functionals in (4) corresponds to the number of times the smallest tuple can be included into the medium-sized tuple. For those number of functionals it is essential to consider the triangle inequality. At the end, we may conclude that $g_1(\bar{x}, \bar{y})$ always exists. The theorem is proved.

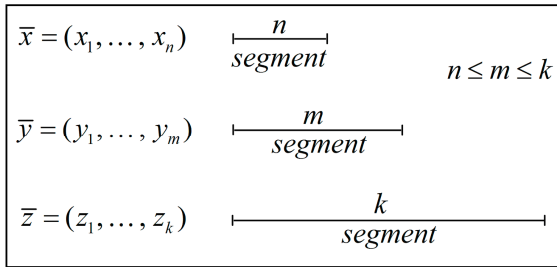


Fig. 3. Schema of possible relations between the three tuples

Consider the functional $g(\bar{x}, \bar{y})$ as the sum of all the elements in the distance matrix. It is easily seen that this functional is symmetric and reflexive as well as $g_1(\bar{x}, \bar{y})$. The triangle inequality should be observed separately. Suppose $card(\bar{x}) = card(\bar{y}) = card(\bar{z}) = n$, then the

distance matrices $A(\bar{x}, \bar{y})$, $A(\bar{x}, \bar{z})$, $A(\bar{y}, \bar{z})$ will look as (2), and the functionals $g(\bar{x}, \bar{y})$, $g(\bar{x}, \bar{z})$, $g(\bar{y}, \bar{z})$ will look as (4). By analogy to the theorem 1, the triangle inequality is grounded for $g(\bar{x}, \bar{y})$. By transferring to a more general case, consider $card(\bar{x}) = n$, $card(\bar{y}) = m$, $card(\bar{z}) = k$. Assume the following relations are held:

$$\begin{cases} n \leq m \leq k, \\ m - n = s_1, \\ k - m = s_2. \end{cases} \quad (9)$$

It is clear that $k - n = s_1 + s_2$, and the distance matrices will look as follows:

$$\begin{aligned}
A(\bar{x}, \bar{y}) &= \begin{pmatrix} \rho(x_1, y_1) & \dots & \rho(x_n, y_n) \\ \vdots & \ddots & \vdots \\ \rho(x_1, y_{s_1}) & \dots & \rho(x_n, y_{s_1+n-1}) \end{pmatrix}, \\
A(\bar{x}, \bar{z}) &= \begin{pmatrix} \rho(x_1, z_1) & \dots & \rho(x_n, z_n) \\ \vdots & \ddots & \vdots \\ \rho(x_1, z_{s_1+s_2}) & \dots & \rho(x_n, z_{s_1+s_2+n-1}) \end{pmatrix}, \\
A(\bar{y}, \bar{z}) &= \begin{pmatrix} \rho(y_1, z_1) & \dots & \rho(y_m, z_m) \\ \vdots & \ddots & \vdots \\ \rho(y_1, z_{s_2}) & \dots & \rho(y_m, z_{s_2+m-1}) \end{pmatrix}.
\end{aligned} \quad (10)$$

The following equations can be obtained from the above:

$$\begin{aligned}
g(\bar{x}, \bar{y}) &= [\rho(x_1, y_1) + \dots + \rho(x_n, y_n)] + \dots \\
&+ [\rho(x_1, y_{s_1}) + \dots + \rho(x_n, y_{s_1+n-1})], \\
g(\bar{x}, \bar{z}) &= [\rho(x_1, z_1) + \dots + \rho(x_n, z_n)] + \dots \\
&+ [\rho(x_1, z_{s_1+s_2}) + \dots + \rho(x_n, z_{s_1+s_2+n-1})], \\
g(\bar{y}, \bar{z}) &= [\rho(y_1, z_1) + \dots + \rho(y_m, z_m)] + \dots \\
&+ [\rho(y_1, z_{s_2}) + \dots + \rho(y_m, z_{s_2+m-1})].
\end{aligned} \quad (11)$$

The equations in (11) enable checking triangle inequality directly. By rearranging the summands and assuming that $\rho(x, y)$ is a metric, this check testifies fulfillment of the triangle inequality in a general case. The following section provides information on the model implementation on video sequences and estimation of the results. To enhance this paradigm in future, it may be interesting to divide the first (smaller) set into subsets and perform search of these smaller subsets in the second (bigger) set. The practical application of it seems quite trivial as not always the whole video scene is repeated, but a small fragment of it.

III. EXPERIMENTAL RESULT

Assume Ω is a video sequence of frames. Let $\rho(x, y)$ be a metric or a distance between the two elements from this

set Ω . Here, x and y are minimum possible video elements, i.e. frames in case of video. Suppose $\bar{x} = (x_1, x_2, \dots, x_n) \in F$, $\bar{y} = (y_1, y_2, \dots, y_m) \in F$ where F is a set of all the scenes in a video. Thus, \bar{x} and \bar{y} are the two video segments (fragments or scenes) for comparison, and x_1, x_2, \dots, x_n and y_1, y_2, \dots, y_m are the frames in these segments, which number equals to n and m respectively. The above matrix (1) indicates how two scenes match each other.

For example, as a first sequence, a fragment of a documentary was taken, illustrating the rescue helicopter crash. In Fig. 4 a) the complete video sequence is shown in the form of 501 frames, which can be divided into a set of segments by the video spatio-temporal segmentation approaches proposed in [4,16]. As a result, you can get the following partition: the first segment, 1..66 frames are illustrating the lone flight of the helicopter; 67...151 frames are helicopter flight against the shore; 152...260 frames are the landing process and rescuers work; 261...332 frames are departure of the rescue helicopter; 333...451 frames are the process of crash; 451...501 frames are segment showing the result of the rescue helicopter crash into the water.

Next, for the experiment, the first and last segments from the original data were taken, with a length of 66 and 49 frames, respectively. Examples of frames from this sequence are shown in Fig. 4 b) and c).

To these segments, the same approach to segmentation was applied similarly and the results corresponding to these segments were obtained. The graphs in Fig. 5 illustrate result of video segmentation for all 3 video. Despite the fact that the segmentation approach was applied to different sequences, the values obtained for segments 4b) and c) correspond to values

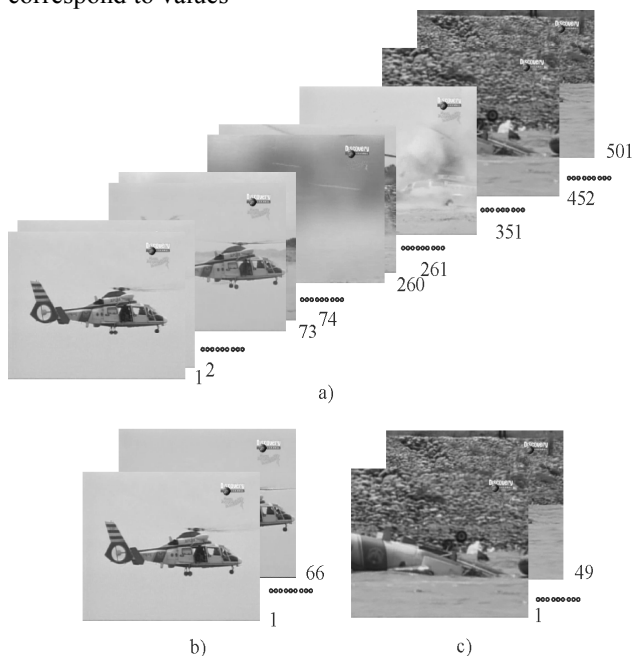


Fig. 4. Example of initial video data and segments to compare

in the intervals from 1 to 66 frames and from 452 to 501 frames of the original video data. It is quite logical that when we using the proposed metric for comparison with the initial sequence (Fig. 4a)) one of the analyzed segments (Fig. 4b) or

Fig. 4c)), we obtained at the appropriate places of the distance matrix (1) a zero sequences values 66 and 49 frames, respectively.

Thus, we can compare video sequences and find the same or similar if we establish a certain threshold value for possible differences in the distance matrix (1). In other words, if we compare the results of different video sequences segmentation with the proposed approach, we obtain a certain sequence in the distance matrix whose values do not exceed the established threshold, then we can say about the similarity of the compared data, and in the case of obtaining a zero sequences about the conjunction of the compared data corresponding parts.

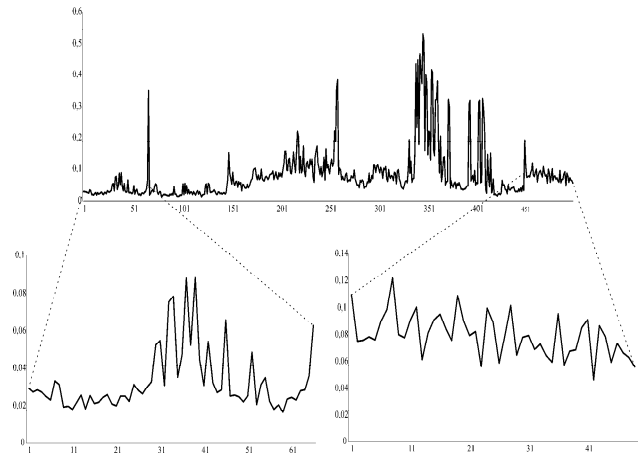


Fig. 5. Result of spatio-temporal video segmentation

IV. CONCLUSION

The novel proposed model may be effectively implemented for video scene comparison. The effectiveness of it can be testified by traditional precision-recall metrics in terms of finding perfect matches for the subsets under analysis. For video content presentation and scene comparison, temporal segmentation plays the key role because extraction of correct scene fragment duplicate is basically what we are trying to reach using the model. Precision-recall metrics reveals correspondence of multidimensional data processing results to human expectations from mental analysis of such data. Although this estimation is performed by human experts, which may be subjective in a sense, the combination of metric parameters more precisely indicates opportunities of the model. The true positives show how many relevant frame sequences are extracted. The false positives are considered being mismatches of extracted segments. The false negatives are the omitted segments that should actually be extracted. The true negatives are not being used as they are the inverse from the above [18, 19]. The only drawback of the estimation is that it does not take into account fuzzy matches and half-satisfaction of the experts, which may be the topic of further research.

REFERENCES

- [1] D. Schonfeld, et. al., Video search and mining. Studies in Computational Intelligence. Springer, Berlin, 2010.
- [2] R. Szeliski, Computer vision. Algorithms and applications. Springer, London, 2011.

- [3] L. Chen, and F. W. M. Stentiford "Video sequence matching based on temporal ordinal measurement," *Pattern Recognition Letters.*, vol. 29, pp. 1824-1831, 2008.
- [4] S. Mashtalir, and O. Mikhnova, "Key frame extraction from video: framework and advances," *J. Computer Vision and Image Processing.* vol. 4(2), pp. 67-78, 2014. (<https://www.igi-global.com/article/key-frame-extraction-from-video/115840>)
- [5] H. Lu, and Y.-P. Tan, "An effective post-refinement method for shot boundary detection," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 15(11), pp. 1407–1421, November, 2005.
- [6] W. Heng, and K. Ngan, "Shot boundary refinement for long transition in digital video sequence", *IEEE Transactions on Multimedia*, vol. 4(4), pp. 434-445, December, 2002.
- [7] A. F. Smeaton, P. Over, and A. R. Doherty, "Video shot boundary detection: Seven years of TRECVID activity," *J. Computer Vision and Image Understanding.* vol. 114(4), pp. 411-418, 2010.
- [8] Zhang Y.-J. (ed.), *Advances in image and video segmentation.* Hershey- London-Melbourne-Singapore: IRM Press, 2006.
- [9] S. Porter, M. Mirmehdi, and B. Thomas, "Temporal video segmentation and classification of edit effects", *Image and Vision Computing.*, vol. 21, pp. 1097-1106, December 2003.
- [10] S. Piramanayagam, E. Saber, N. D. Cahill, and D. Messinger, "Shot boundary detection and label propagation for spatio-temporal video segmentation" *Proc. SPIE 9405, Image Processing: Machine Vision Applications VIII, 94050D 7 p.*, February 2015.
- [11] S. Thakare, "Intelligent processing and analysis of image for shot boundary detection," *International Journal of Emerging Technology and Advanced Engineering.*, vol. 2, no. 2, pp. 208-212, Mar.-Apr. 2012.
- [12] R. Vázquez-Martín, and A. Bandera, "Spatio-temporal feature-based keyframe detection from video shots using spectral clustering," *Pattern Recognition Letters*, vol. 34, no. 7, pp. 770-779, 2013.
- [13] G. I. Rathod, and D.A. Nikam, "An algorithm for shot boundary detection and key frame extraction using histogram difference," *Int. J. Emerging Technology and Advanced Engineering*, vol. 3(8), pp. 155-163, August, 2013.
- [14] J. Nesvadba, F. Ernst, J. Perhac, J. Benois-Pineau, and L. Primaux, "Comparison of shot boundary detectors", *Int. Conf. on Multimedia and Expo*, IEEE Press, Amsterdam, pp. 6-8, 2005.
- [15] H. Jiang, G. Zhang, H. Wang and H. Bao, "Spatio-temporal video segmentation of static scenes and its applications" *IEEE Transactions on Multimedia.*, vol. 17, no. 1, pp. 3-15, January, 2015.
- [16] Y. Bodyanskiy, D. Kinoshenko, S. Mashtalir, and O. Mikhnova, "On-line video segmentation using methods of fault detection in multidimensional time sequences", *Int. J. of Electronic Commerce Studies*, vol. 3(1), pp. 1-20, 2012.
- [17] O. Mikhnova, and N. Vlasenko, "Key frame partition matching for video summarization," *Int. J. of Information Models and Analyses*, vol. 2(2), pp. 145-152, 2013.
- [18] C. D. Manning, P. Raghavan, and H. Schütze, *Introduction to Information Retrieval.* Cambridge University Press, Cambridge, 2008.
- [19] S. V. Mashtalir, and O. D. Mikhnova, "Stabilization of key frame descriptions with higher order Voronoi diagram", *J. Bionics of intelligence.* vol. 1, pp. 68-72, 2013.

Image Segmentation Metric-Based Adaptive Method

Oleh Berezsky

Department of computer engineering
TNEU

Ternopil, Ukraine
ob@tneu.edu.ua

Kateryna Berezska

Department of computer engineering
TNEU

Ternopil, Ukraine
km.berezska@tneu.edu.ua

Oleh Pitsun

Department of computer engineering
TNEU

Ternopil, Ukraine
o.pitsun@tneu.edu.ua

Nadiya Savka

Department of computer engineering
TNEU

Ternopil, Ukraine
nadya_savka@ukr.net

Natalia Batryn

Department of computer engineering
TNEU

Ternopil, Ukraine
nbatryn@gmail.com

Taras Dolynyuk

Department of computer engineering
TNEU

Ternopil, Ukraine
trsdln@gmail.com

Abstract—In this work, the authors analyzed the existing algorithms of image pre-processing and segmentation and determined their advantages and disadvantages. There was developed a metric-based method to choose automatically the segmentation algorithms and their parameters for cytological and histological image processing. This approach allows increasing the speed of segmentation algorithms and their parameters choice for different classes of images.

Keywords— *segmentation, adaptive method, histological image, cytological, metric.*

I. INTRODUCTION

In the existing automated microscopy systems (AMS), such as imageJ, AxioVison, Amira, BioImageXD, image segmentation plays a key role. On the basis of segmentation, the investigated micro-objects are distinguished and their quantitative characteristics are calculated for further classification. Histological and cytological images are characterized by a high level of processing complexity because this type of images is impacted by significant impulse noise [1]. In the work [2], the researchers compared different ways of pre-processing images based on their ability to remove noise and segment the image. In [3], an improved computational approach based on a gradient image filter was proposed. The developed approach allowed effectively detecting contours of objects and reducing inappropriate information created from the background and texture. Spatial filter algorithms were analyzed in the works [4,5]. In [7-10], the analysis of different approaches to the classification of segmentation algorithms was conducted. In [11], an overview of the basic methods of segmentation of objects in a digital image was carried out, as well as the analysis of the effectiveness of the application of these methods for the irregular object segmentation. Quantitative evaluation of segmentation quality allows objectively determining the best method and its input parameters. The advantage of the quality evaluation is absence of a subjective (human) factor. Compared to the FRAG algorithms, the use of metrics to find the distance between objects allows a more accurate estimation of their similarity [13-14]. Using a database to store the results of image testing and rules for selecting the methods of segmentation and its parameters can increase the universality of the developed system [15]. With client-server technology, it is possible to test a large number

of images, thus increasing the accuracy of the selection of parameters.

The purpose of the research study is to develop a metric-based adaptive method of image segmentation and conduct the computer experiments on the choice of segmentation algorithms and their parameters for cytological and histological images.

II. SEGMENTATION ALGORITHMS TESTING

The method of automatic selection of segmentation algorithms consists of two main parts: training of the system on the training sample and work on the testing image set that was not included in the training sample. To receive the best results, the existing algorithms and their combinations were selected. The process of image segmentation is time-consuming, therefore, as the result of manual testing, the limits of the algorithms parameters were selected. For example, for the threshold segmentation, the lower threshold is in the range from 35 to 175. The application of the value of the lower threshold outside the range significantly reduces the processing quality.

The method of choosing an algorithm and segmentation parameters is the following:

1. Determining the input parameters of the image (brightness level, average values of red, green and blue channels);

2. Image segmentation. At this stage, the following methods are used: threshold segmentation, watershed method, and k-mean method. For threshold segmentation, a set of values for the lower threshold (35 - 175) is used with step 5. The k-mean method uses a set of different values of the flags for testing (Figure 1);

3. Segmentation evaluation. Each image is compared with the standard segmentation performed by an expert. To evaluate the similarity between images, the Gromov-Hausdorff metric, the Gromov-Fréchet metric, and the FRAG parameter are used. Additionally, a subjective individual's evaluation is used;

4. The best result is saved in the database for further formulation of the rules.

III. SEGMENTATION QUALITY EVALUATION

To evaluate the segmentation quality, a metric-based method of finding the distance between images is used. To find distances between images, the Hausdorff and Fréchet metrics are used. To find the smallest distances between images, the metrics of Gromov-Hausdorff and Gromov-Fréchet are used. Let us give the basic definition of metrics.

Fréchet metric. For two curves $f : [a, b] \rightarrow X$ and $g : [a', b'] \rightarrow X$ the Fréchet distance between them is equal:

$$d_F = \inf_{\alpha, \beta} \sup_{t \in [0, 1]} d(f(\alpha(t)), g(\beta(t))),$$

where $d(x, y)$ is an Euclidean distance between points X and Y , α and β are arbitrary continuous non-intermittent functions with an interval $[0, 1]$ on intervals $[a, b]$ and $[a', b']$ respectively. Function value $\alpha(0) = 0$ and $\alpha(1) = 1$, and similar to function β [16-18].

Hausdorff metric. For a metric space (X, d) , the Hausdorff metric d_H is called metric on the set \mathfrak{S} of all compact subsets X , which is defined as:

$$d_H^X(A, B) := \max \left\{ \max_{x \in A} \min_{y \in B} d(x, y), \max_{y \in B} \min_{x \in A} d(x, y) \right\}$$

Gromov-Hausdorff metric. The distance between two compact sets A and B is equal to:

$$d_{GH}(A, B) := \inf_{X, f, g} d_H^X(f(A), g(B)),$$

where $f : A \rightarrow X$, $g : B \rightarrow X$ – isometric embeddings to some metric space (X, d) . Gromov-Fréchet metric. Similarly, to find the smallest distance between two curves, we use the Gromov-Fréchet metric:

$$d_{GF}(S, Q) := \inf_{X, S, Q} d_F^X(f(S), g(Q)),$$

where $f : S \rightarrow X$, $g : Q \rightarrow X$ – isometric embeddings to metric space (X, d) .

IV. AUTOMATIC IMAGE SEGMENTATION MODULE

The graphical representation of automatic segmentation module is shown in Fig. 1. In the learning phase, the first step is to select the input image parameters.

The algorithm of automatic segmentation module is the following:

1. Image upload;
2. Selection of input image parameters (brightness level, average values of red, green and blue channels);

3. Search for the segmentation method and its parameters in the database;
4. Segmentation by the selected method;
5. Saving the result.

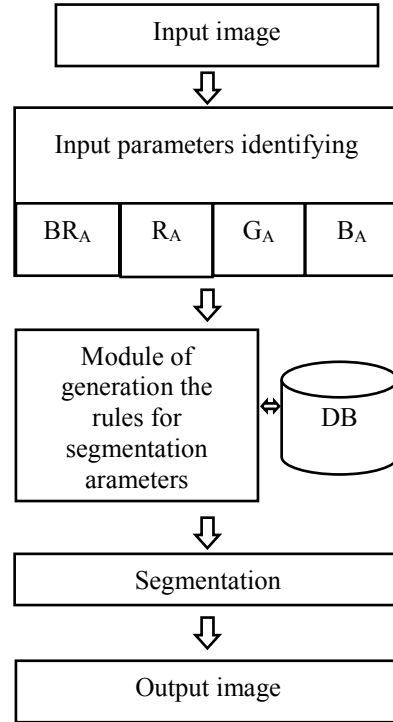


Fig. 1. Automatic image segmentation

V. DATABASE STRUCTURE

The database for storing the learning outcomes of the system consists of two tables. The "Input Parameters" table is designed to store information about the input parameters of the image. The "Algorithm Parameters" table is designed to store learning outcomes. The table consists of fields where information about the best algorithm and its parameters is stored, depending on the input parameters of the image. The structure of the database tables is shown in Fig. 2.

VI. RESULTS

The results of the automatic selection of histological and cytological image segmentation parameters are presented in Table 1. In the table, the input data is the initial image, the mask processed by the expert, the result of the automatic threshold segmentation by the ImageJ software, and the result of the developed module work.

Based on the above mentioned results, it is possible to conclude that the developed system of automatic selection of segmentation parameters has better results compared with the methods of automatic segmentation of the Image J software. For each group of input parameters, on the basis of expert evaluations, the linguistic estimates of the distribution of segmentation algorithms and their parameters were selected, depending on the input parameters. Table 2 contains the linguistic estimates of the values of image segmentation, where VL refers to very low quality, L means low quality, M is Medium, H is high quality of segmentation.

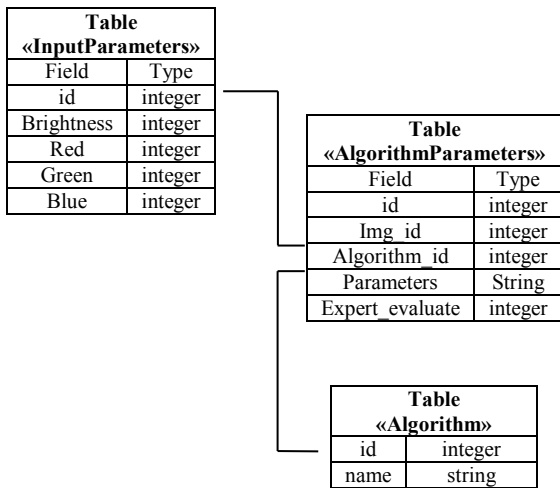


Fig. 2. Structure of DB Tables

Lower threshold value	65	85	95	105	120	140	160
Method							
THRESHOLD thresh_binary	VL	VR	L	M	M	L	L
THRESHOLD thresh_otsu	L	L	L	H	M	M	L
THRESHOLD thresh_binary + thresh_otsu	M	M	M	H	H	M	L
THRESHOLD adaptive_thresh _gaussian_c	L	M	M	M	M	M	L
WATERSHED thresh_binary + thresh_otsu	M	M	H	H	H	M	M
K-MEANS thresh_binary + thresh_otsu	L	M	M	H	M	M	L

TABLE I. SEGMENTATION RESULTS

Input image	Expert processing	Automatic segmentation (ImageJ)	Developed module

The fuzzy knowledge base model is developed in the Fuzzy Logic Toolbox editor. The term Brightness has the following meanings: Very Low, Low, Medium, High, Very High, and Brightness. The membership function is Gaussian.

The term "averageBlueValue" characterizes the level of the blue RGB image channel in the range from 0 to 255 and consists of 6 values. For this variable, a generalized bell-like membership function was chosen.

An example of implementing rules using Fuzzy Logic Toolbox is shown in Fig. 3.

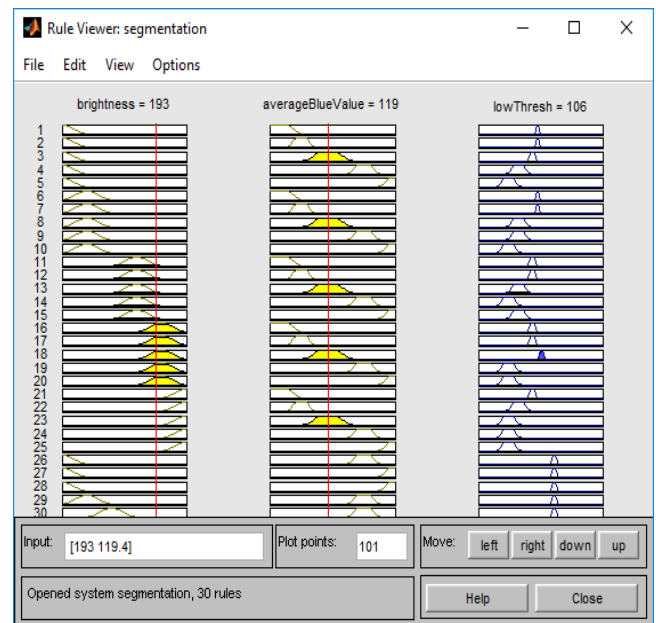


Fig. 3. Example of rules application by means of Fuzzy Logic Toolbox

The example is the following:

IF Brightness = 32 **AND** BlueValue = 201 **THEN** Algorithm = Watershed + Threshold, lowThreshold = 150

IF Brightness = 119 **AND** BlueValue = 229 **THEN** Algorithm = Watershed + Threshold, lowThreshold = 95

IF Brightness = 194 **AND** BlueValue = 114 **THEN** Algorithm = Watershed + Threshold, lowThreshold = 100

IF Brightness = 230 **AND** BlueValue = 142 **THEN** Algorithm = Watershed + Threshold, lowThreshold = 100

IF Brightness = 41 **AND** BlueValue = 139 **THEN** Algorithm = Watershed + Threshold, lowThreshold = 90

IF Brightness = 120 **AND** BlueValue = 138 **THEN** Algorithm = Threshold, lowThreshold = 85, type = THRESH_BINARY

IF Brightness = 65 **AND** BlueValue = 110 **THEN** Algorithm = Watershed + Threshold, lowThreshold = 95

TABLE II. LINGUISTIC ESTIMATES OF IMAGE SEGMENTATION METHODS

IF Brightness = 3 AND BlueValue = 144 **THEN** Algorithm = Watershed + Threshold, lowThreshold = 105

VII. CONCLUSION

1. The basic algorithms of image segmentation have been analyzed, which allowed identifying a set of algorithms for cytological and histological image segmentation.

2. The method of automatic selection of segmentation algorithms and their parameters is developed on the basis of the used Gromov-Hausdorff and Gromov-Fréchet metrics.

3. Based on computer experiments, a base of rules for combinations of segmentation algorithms and their input parameters was formed, which makes it possible to perform segmentation in automatic mode for a certain class of images.

REFERENCES

- [1] O. Berezsky, and O. Pitsun, "Automated Processing of Cytological and Histological Images" Perspective Technologies and Methods in MEMS Design (MEMSTECH'2016): XII th International Conference, Lviv-Polyana, Ukraine, pp. 51-53, 20-24 April 2016. DOI:10.1109/MEMSTECH.2016.7507518
- [2] S. Adatrao, and M. Mittal, "An analysis of different image preprocessing techniques for determining the centroids of circular marks using hough transform," 2nd International Conference on Frontiers of Signal Processing (ICFSP), pp. 110-115, 15-17 Oct. 2016. DOI: 10.1109/ICFSP.2016.7802966
- [3] P. Kaur, and A. Gupta, "Contour Detection of Gradient Images Using Morphological Operator and Transform Domain Filtering," IEEE International Conference on Computational Intelligence & Communication Technology, pp. 107-111, 2015.
- [4] V. N. Tyapkin, I. N. Kartsan, D. D. Dmitriev and A. E. Goncharov, "Spatial filtering algorithms in adaptive multi-beam hybrid reflector antennas," International Siberian Conference on Control and Communications (SIBCON), Omsk, Russia, 2015
- [5] C.-V. Gustavo, T. Devis, L. Bruzzone and J. Benediktsson "Advances in Hyperspectral Image Classification: Earth Monitoring with Statistical Learning Methods," IEEE Signal Processing Magazine, vol. 31(1), pp. 45 – 54, 2014. DOI: 10.1109/MSP.2013.2279179.
- [6] Y. Zhu, and C. Huang "An Improved Median Filtering Algorithm for Image Noise Reduction", Physics Procedia, vol. 25, pp. 609 – 616, 2012,
- [7] S. Masood, M. Sharif, A. Masood, M. Yasmin and M. Raza, "A Survey on Medical Image Segmentation," Current Medical Imaging Reviews, vol. 11, pp. 3-14, 2015.
- [8] S. Y. Yeo, X. Xie, I. Sazanov and P. Nithiarasu "Segmentation of biomedical images using active contour model with robust image feature and shape prior," International Journal for Numerical Methods in Biomedical Engineering, vol. 30(2), pp. 232-248, 2014.
- [9] S. Divya, and K. B. Jayanthi "Analysis of contour evolution methods for segmentation of medical images," International Conference on Innovations in Information, Embedded and Communication Systems (ICIIECS), pp. 1-4, 2015.
- [10] P. Campadelli, E. Casiraghi, and S. Pratisoli, "A segmentation framework for abdominal organs from CT scans," Artificial Intelligence in Medicine, vol 50(1), pp. 3-11, 2010. ISSN 0933-3657
- [11] A. Norouzi, "Medical Image Segmentation Methods, Algorithms, and Applications," IETE Technical Review, vol. 31, pp. 199-213, 2014.
- [12] M. I. Schlesinger, E. V. Vodolazskiy, and V. M. Yakovenko "Fréchet Similarity of Closed Polygonal Curves," International Journal of Computational Geometry, vol. 26, pp. 53-66. 2016, DOI: 10.1142/S0218195916500035
- [13] H.-K. Ahn, C. Knauer, and M. Scherfenberg, "Computing the discrete Fréchet distance with imprecise impute," International Journal of Computational Geometry, vol. 22, pp. 27-44, 2016. DOI: 10.1142/S0218195912600023
- [14] J. Gudmundsson, and M. Smid "Fast algorithms for approximate Fréchet matching queries in geometric trees," Computational Geometry, vol. 48, pp. 479-494, 2015. DOI:10.1016/j.comgeo.2015.02.003
- [15] O. Berezsky, L. Dubchak, and O. Pitsun, "Access distribution in automated microscopy system", 14th International Conference The Experience of Designing and Application of CAD Systems in Microelectronics (CADSM), Lviv, Ukraine, pp. 241 – 243, 21-25 Feb. 2017.
- [16] O. M. Berezsky, G. M. Melnyk, Y. M. Batko and O.Y. Pitsun, "Regions Matching Algorithms Analysis to Quantify the Image Segmentation Results," Sensors & Transducers, vol. 208(1), pp. 44-49, 2017.
- [17] O. M. Berezsky, "Fréchet metric for trees," IEEE 1st International Conference on Data Stream Mining and Processing, DSMP, Lviv, Ukraine, pp. 213-217, 23-27 Aug. 2016,
- [18] O. Berezsky, M. Zarichnyi, and O. Pitsun, "Development of a metric and the methods for quantitative estimation of the segmentation of biomedical images," Eastern European Journal of Enterprise Technologies, vol. 6 (90), pp.4-11, 2017.

Interactive Computer Simulators in Rescuer Training and Research of their Optimal Use Indicator

Igor Malets

*Department of Project Management, Information
Technologies and Telecommunications
Lviv State University of Life Safety
Lviv, Ukraine
igor.malets@gmail.com*

Oleksandr Prydatko

*Department of Project Management, Information
Technologies and Telecommunications
Lviv State University of Life Safety
Lviv, Ukraine
o_prydatko@ukr.net*

Vasyl Popovych

*Department of Environmental Safety
Lviv State University of Life Safety
Lviv, Ukraine
popovych2007@ukr.net*

Andriy Dominik

*Department of vehicle operation and fire-rescue equipment
Lviv State University of Life Safety
Lviv, Ukraine
dominik.andrij@ukr.net*

Abstract — The scientific work is devoted to the description of the results of long-term work on the development and research of the effectiveness of interactive computer simulators training rescuers at the Lviv State University of Life Safety. Focuses on the study of the indicator of optimal use of developed simulators provided that the required indicator of the effectiveness of training is achieved. Flash technologies are used to develop interactive simulators, and for the processing of the results of the experimental part, the methods of mathematical statistics. The obtained results provided the grounds for substantiating the optimal quantitative and qualitative indicator of the use of the developed technology.

Keywords — *computer simulator, optimal use, statistical indicator*

I. INTRODUCTION

The rapid pace of development of technologies and methods of training leads to the irreversible process of adapting the educational space to the use of information technology in any field of knowledge. The integration of interactive techniques based on the comprehensive use of computer technologies into the process of preparing a "modern lifeguard" capable of working in conditions of global informatization of society is not an exception [1, 2].

It is obvious that the quality of the formed competence, acquired skills and skills in the process of preparation determines the professional level of the future rescuer and his competitiveness in the labor market. That is why the development and research of the efficiency of a modern innovative approach to the process of formation of professional competence is an actual scientific and applied problem. The main techniques and concepts used in world practice to improve the quality of the training process for rescuers are aimed at developing tactical skills and improving the decision-making process for successful emergency response [3, 4, 5, 6].

The works devoted to the principles of artificial intelligence and machine learning in the educational process should be noted. Significant contribution can be found in papers [7, 8, 9]. Development of educational process using automated computer systems is reflected in scientific works [10, 11]. Significant contribution to the management of the educational environment lead by the example of the Federal Republic of Germany is reflected in the work [12]. Some issues of sustainable development of higher education with the full application of information and communication technologies are presented in the works [13, 14, 15, 16].

From domestic and world experience in the development of interactive simulators for the training of rescue specialists found that existing technologies allow you to master the skills "what to do", but without the skills to "how to do". In this regard, the research team was tasked with the development of interactive simulators that will enable them to master the skills of work with technical means of salvation.

II. DEVELOPMENT OF INTERACTIVE COMPUTER SIMULATORS

In search of a new and more effective form of teaching material, a number of interactive learning tools for cadets and students of educational institutions working in the field of human life safety are actively being developed. As experience shows [17, 18], there are computer technologies, the use of which requires only a personal computer with the corresponding software, as an alternative to traditional technology means that can be used exclusively in the landfill. Of course, a complete replacement of traditional technology is not possible, because in practice the expert will need to work with real equipment. However, it is possible to substantially limit the amount of their use at the expense of the developed means of alternative technology. Consequently, the general principle of innovation technology, which is proposed to

improve the process of practical training of rescuers, consists in the combined (reciprocal) application of innovative computer tools and real equipment. Of course, such an innovative approach will stimulate the reduction of the cost of training, but it will not be innovative until it is of a quality benefit to the existing system. So, let's consider the following component of the proposed training technology, which is related to the development and use of computer simulators.

Fash-technology was used to implement the idea of creating interactive computer simulators. The software package is the perfect medium for creating the most diverse multimedia products.

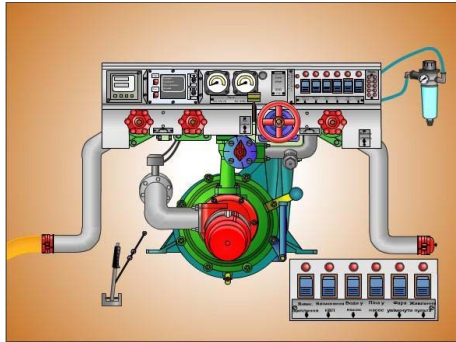


Fig. 1. A general view of an interactive computer simulator

With the help of the developed interactive computer simulators you can learn to do any exercises on work with fire pumps. The simulators allow practically to practice the exercise without significant physical activity and with the corresponding saving of resources. The only requirement is the presence of a computer with the appropriate software. Exercise training on the simulator is carried out in stages, according to a clear algorithm.

When an error is made, the simulator worksheet tells the user about the wrong effect by text and voice comment and allows you to fix it yourself. The user, if desired, can work out the exercise as many times as required.

To prevent mistakes in practice, which may cause a failure of a particular unit or node, after the end of an exercise user is able to see the typical mistakes that occur when working with real equipment.

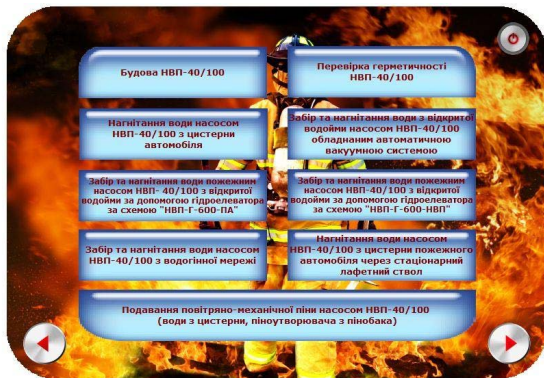


Fig. 2. Window menu of software designed facility (in Ukr.)

Exercises using interactive simulators allow you to create the foundation in the form of prevailing knowledge

and skills for the future formation of professional skills. The main advantage of such simulators is that they allow us to process the key, we can say, technological skills to solve more complex, combined tasks in the traditional form in practice. This ensures active mental and manipulative activity of the cadet and student.

For ease of use of the developed simulators, we have formed a program menu, which reduces the choice of the necessary exercise to perform within the same window.

III. RESEARCH COMPUTER SIMULATORS INDEX OF OPTIMAL USE

Obviously, the actual question is the study of the effectiveness of the application of the developed means of innovative learning technology, which is partly reflected in the works [1, 5]. However, in this paper, we focus our efforts on determining the optimal number of cyclic applications of developed simulators in the process of practical training. In our opinion, the optimality will be determined at the limit of the number of practical training and quality of the received competencies (skills). That is why the experimental part will be built on the basis of determining the minimum-permissible number of cyclic workouts on the simulator with the achievement of the maximum value of the quality of the acquired practical skills (working out a regular exercise without any error).

Moreover, the experiment was conducted as part of the training of future rescuers by monitoring the results of practical exercises. As experimental units three types of exercises on simulators for work with fire pumps have been chosen. The main indicator, which was determined during the experimental part, is the average number of cyclic attempts to perform the exercise until a no mistakes are made. Users, before exercising, do not get acquainted with the hands-on training provided for the class. The training was limited to the theoretical part. The results of experimental studies are presented as a consolidated sample in Table 1.

TABLE 1. SUMMARY RESULTS OF EXPERIMENTS

Simulator	Q-ty Participants	numb. of attempts	Mistakes		Execution time	
			n_{min}	n_{max}	t_{min}	t_{max}
simulator1	56	8	0	19	2,0	13,4
simulator2	53	6	0	16	1,5	15,1
simulator3	51	7	0	17	2,1	14,5

As the first exercise (simulator 1), one of the most common tasks for working with a fire pump was chosen - the injection of water by the pump NPP-40/100 from the tank of the fire truck. Second exercise (simulator 2) - injection of water from an open reservoir by a pump of NPP-40/100 equipped with an automatic vacuum system. And, accordingly, the third exercise (simulator 3) - the capture and injection of water by the pump NPP-40/100 from the water supply network.

The results of experimental studies are presented as an interval statistical distribution and are presented in the form of histograms of frequencies.

The results of experimental studies are the source data for predicting the optimal rate of use of interactive simulators. However, before switching to forecasting, it is necessary to determine which distribution law corresponds to the results obtained.

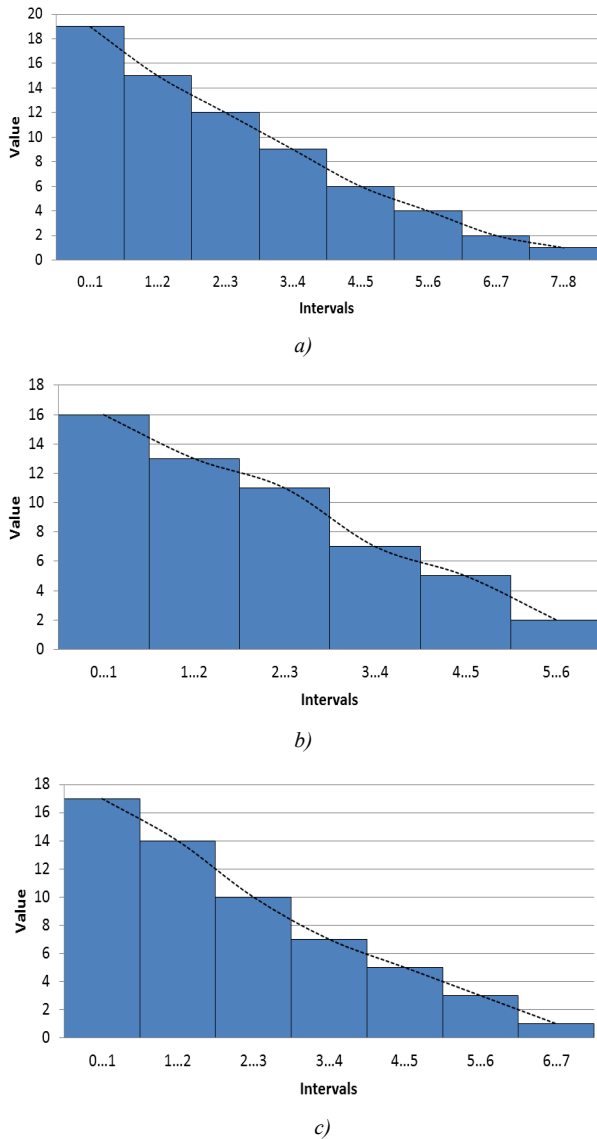


Fig. 3. Histograms of the frequencies of statistical distribution by observation results: a - simulator 1; b - simulator 2; c - the simulator 3

According to the form of the histogram spectra (Fig. 3), we can assume that the X sign has an exponential law of probability distribution. However, our assertions are only hypothetical, and the correctness of this hypothesis needs to be checked. To test the null hypothesis of the exponential law of the distribution of the sign of the general population, mathematical statistics use the Pearson consistency criterion. Therefore, in order to calculate this criterion, the numerical characteristics of the interval statistical distribution of the samples were further determined.

Upon learning variation interval of three rows of statistical distributions obtained results are listed in Table 2.

TABLE 2. NUMERICAL CHARACTERISTICS OF THE STUDIED INTERVAL STATISTICAL DISTRIBUTIONS

simulator	$n = \sum_{i=1}^k n_i$	$\bar{x}_B = \frac{\sum_{i=1}^k x_i^* n_i}{n}$	$\lambda = \frac{1}{\bar{x}_B}$
simulator 1	56	2,91	0,3436
simulator 2	53	2,132	0,469
simulator 3	51	2,313	0,4322

As is well known, the Pearson consistency criterion has a distribution χ^2 with $k = q - m - 1$ degrees of freedom and

determined by dependence $\chi^2 = \sum_{i=1}^q \frac{(n_i - np_i)^2}{np_i}$, where n_i – the empirical sampling frequencies, and np_i – the theoretical sampling frequencies respectively. Next, in order to verify the validity of the hypothesis of the exponential distribution law, we have carried out the definition of theoretical frequencies in the investigated cases and on the common graphical grid in comparison with the results of empirical frequencies (Fig. 4).

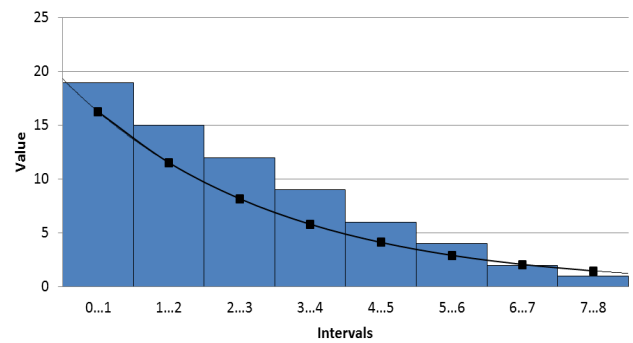


Fig. 4. Comparison of theoretical and empirical frequencies of the statistical distribution of the results of observation by simulator 1

From Figure 4 we can conclude that the sign of the general population is hypothetically consistent with the exponential distribution law, since the difference between empirical and theoretical frequencies is insignificant. However, due to the fact that this statement is hypothetical, it needs to be checked using the criterion of coherence

$$\chi_{\text{crit}}^2 = \sum_{i=1}^8 \frac{(n_i - np_i)^2}{np_i} = 6,15.$$

We determine the significance level $\alpha = 0.05$ and the number of degrees of freedom $k = 6$ critical point $\chi_{\text{kp}}^2 (\alpha = 0,05; k = 8 - 1 - 1) = 12,6$.

Consequently, we can conclude that there are no grounds for the rejection of the null hypothesis of the exponential law of the distribution of the results of the observation of exercises with the help of simulator # 1, because $\chi_{\text{crit}}^2 \in [0; 12,6]$.

Next, in the same way as the given case, we carry out the verification of the null hypothesis for the

correspondence of the exponential law for the cases studied with simulators # 2 and 3.

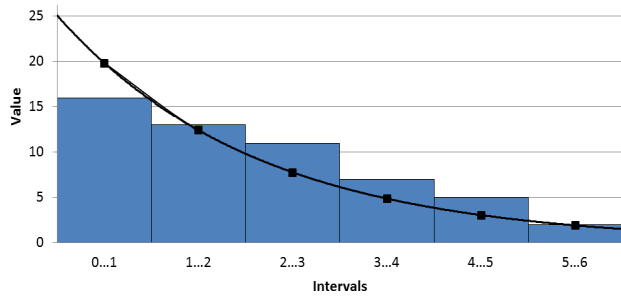


Fig. 5. Comparison of theoretical and empirical frequencies of the statistical distribution of the results of observation by the simulator 2

The sign of the general population is consistent with the exponential distribution law, since the observational value $\chi_{cn}^2 = \sum_{i=1}^6 \frac{(n_i - np_i)^2}{np_i} = 4,19$. And the results of determining the critical boundary are $\chi_{kp}^2 (\alpha = 0,05; k = 6 - 1 - 1) = 9,5$.

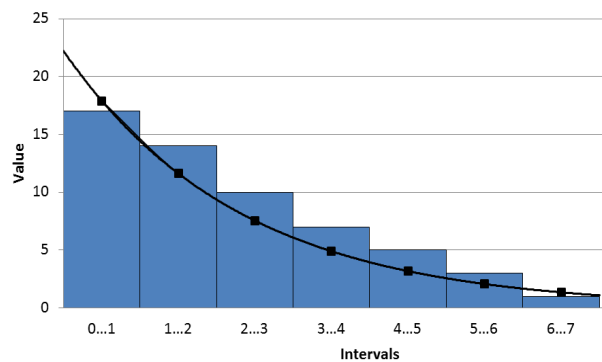


Fig. 6. Comparison of theoretical and empirical frequencies of the statistical distribution of the results of observation by the simulator 3

For simulator number 3 the value of the observation criterion is $\chi_{cn}^2 = \sum_{i=1}^6 \frac{(n_i - np_i)^2}{np_i} = 3,52$. And the value of a critical criterion - $\chi_{kp}^2 (\alpha = 0,05; k = 7 - 1 - 1) = 11,1$. Determining the significance level $\alpha = 0.05$ and the number of degrees of freedom $k = 5$, the critical point χ_{kp}^2 we can conclude that there are no grounds for the rejection of the null hypothesis about the exponential law of distribution of the results of observation of the simulator 3, because $\chi_{cn}^2 \in [0; 11,1]$.

As can be seen from the performed inspections, the results of monitoring the innovative approach to the practical training of rescuers are consistent with the exponential distribution of probability laws, and therefore, the hypotheses put forward are confirmed.

Thus, guided by the results of the conducted research, we carry out forecasting of the probable amount of cyclic application of the developed simulators to achieve an absolute qualitative indicator (implementation of the exercise without errors) in the students who will be trained in the proposed technology.

And a standard measure of the probability of error assuming the exercise in the next cycle of the use of the simulator, we take the density (density) distribution of the random variable:

$$y = \lambda e^{-\lambda x}, \quad (1)$$

where λ is the inverse value of the mathematical expectation; e is the Euler number; x is the mean value of the studied interval of attempts.

The results of the prediction for the three cases studied are summarized in Table 3.

TABLE 3. THE PROBABILITY OF ERROR ASSUMPTION WHEN WORKING OUT THE EXERCISE AT THE N-TH STAGE

Exercise number	Probability of errors at the stage:							
	1	2	3	4	5	6	7	8
simul.1	0,29	0,24	0,16	0,12	0,07	0,05	0,04	0,03
simul.2	0,37	0,25	0,18	0,1	0,06	0,04		
simul.3	0,35	0,24	0,17	0,11	0,06	0,04	0,03	

The purpose of the calculations carried out is to determine the required number of cycles of practical exercises until the absolute qualitative indicator is reached (the development of a defined exercise by all participants without error).

Of course, the absolute value is relative, so we reserve the right to set the upper limit value, which will satisfy the qualitative indicator, at 95%. Taking into account this and based on the results of the forecasting (Table 3), we can conclude that the optimal amount of practical training using the developed simulators, in order to achieve the appropriate level of training quality, is 6 cycles. According to the results of the conducted researches, it was also established that the average time to perform a certain number of attempts and exercises, until the achievement of the established quality indicator, will be approximately 60 minutes. In view of the length of the academic hour, it can be argued that one training session (2 academic hours) will be sufficient for successful mastering of the three basic practical exercises on fire pumps. In contrast, it should be noted that the achievement of this qualitative indicator is not always possible under the traditional approach to mastering such practical skills even for 24 academic hours. In this regard, we can argue that the use of a combined approach to the formation of practical competences for future rescuers (interactive simulators and traditional fire-fighting equipment) will provide an opportunity to improve the quality of training with simultaneous savings in resources.

IV. CONCLUSION

According to the results of the work we can formulate the following conclusions:

1. Using Flash-technologies, a complex of interactive computer simulators for the practical training of future rescuers has been developed, which provides an opportunity to build innovative approaches to the process of forming professional competencies.

2. According to the results of experimental researches using mathematical methods, the indicator of optimal use of developed interactive simulators was determined, which gave grounds for preparing proposals for changes in the organization of practical training of rescuers.

REFERENCES

- [1] O. Prydatko, and I. Pasnak, "Investigation of the processes of the information technologies integration into the training of specialists at mine rescue departments," *Scientific Bulletin of National mining university, Lviv*, no.1 (157), pp. 108-113, 2017.
- [2] Y. Rak, T. Rak, I. Malets, and A. Renkas, "Interactive methods of preparation of specialists for fire and safety service," *Development trends in rescue techniques and technical equipment, School of Aspirants of the State Fire Service in Krakow, Poland*, pp. 128-130, 2006.
- [3] Yu. Rak, O. Zachko, "Improvement of the process of making design decisions for fire suppression by means of computer simulator," *Lviv State University of Life Safety, Fire safety*, no. 19, pp. 124-130, 2011.
- [4] B. Shtain, V. Loik, and V. Dubasiuk, "3-D training apparatus as project of preparation firefighter-rescuer," *Bulletin Lviv State University of Life Safety*, no. 7, pp. 147-154, 2013.
- [5] A. Renkas, O. Prydatko, D. Mozol, and T. Gangur, "Innovative technologies quality management projects in training future rescuers," *Bulletin of Lviv State University of Life Safety*, no. 11, pp. 80-88 2015.
- [6] Yu. Rak, O. Zachko, and T. Rak, "Formal logical models of planning the computer trainer from working off the tactical skills of head of fire liquidation," *Bulletin of the National University "Lviv Polytechnic"* no. 688, pp. 197-203, 2010.
- [7] H. Rana, and Lal M. Rajiv, "Role of artificial intelligence based technologies in e-learning," *International Journal of Latest Trends in Engineering, Science and Technology*, vol. 1, no. 24-26, 2014
- [8] G. Samigulina, *Technology immune network modeling for intellectual control systems and forecasting of the complex objects*. Monograph. USA: Science Book Publishing House, 172, 2015
- [9] G. A. Samigulina, and A. S. Shayakhmetova, "Smart-system of distance learning of visually impaired people based on approaches of artificial intelligence," *J. Open Engineering*, vol. 6, pp. 359-366, 2016
- [10] G. V. Rybina, "Development and use of teaching integrated expert systems in the study process," *Scientific Methodological Conference "Development of the preparation of IT-specialists in the sphere of applied informatics on the basis of innovation technologies and e-Learning"*, MESI, Moscow, pp. 219-226, 2007.
- [11] A. G. Khmelev, "Neuronet technologies in the systems of automated students' knowledge control," *XVI All-Ukrainian Scientific Methodological Conference "Problems of Economical Cybernetics"*, Odessa, pp. 120-125, 14th - 16th September 2011.
- [12] V. Gumbold, "On internal and external organization of higher educational establishments in Berlin," *Reserve stock: Debates on politics and culture*, no.2(22), 2002, pp. 3-7. ISBN 5-86793-037-8.
- [13] D. Tilbury, "Another world is desirable: A global rebooting of higher education for sustainable development," Sterling, S., Maxey, L. and Luna, H. (Eds.) *The Sustainable University: Progress and prospects*. Routledge studies in sustainable development, vol. 2. Earthscan/Routledge, London, 2013, pp. 71-86.
- [14] Wals, A. E.J. (2007), "Social Learning towards a Sustainable World. Principles, Perspectives, and Praxis" Wageningen, The Netherlands: Wageningen Publishers, pp. 17-32.
- [15] A. E. J. Wals, "Sustainability in higher education in the context of the UN DESD: a review of learning and institutionalization processes," *Journal of Cleaner Production* 62, pp. 8-15, 2014.
- [16] K. Brundiars, A. Wiek, C. L. Redman, "Real-world learning opportunities in sustainability: From classroom into the real world," *International Journal for Sustainability in Higher Education*, vol. 11(4), pp. 309-324, 2010.
- [17] R. Gurevych, M. Kademiia, and M. Koziar, *Information-communicational technologies in the professional education of future specialists*, Monograph. SPOLOM, Lviv, Ukraine, 2012
- [18] M. Koziar, "Interactive teaching methods university," *Problems and prospects of forming a national humanitarian and technical elite, National Technical University "Kharkiv Polytechnic Institute"*, vol. 42(46), pp. 285-292, 2015.

Analysis of Metal Defects by Clustering the Sample and Distributed Cumulative Histogram

Roman Melnyk
Software Department
Lviv Polytechnic National University
Lviv, Ukraine
ramelnyk@polynet.lviv.ua

Yurii Kalychak
Software Department
Lviv Polytechnic National University
Lviv, Ukraine
yurii.i.kalychak@lpnu.ua

Abstract—In this paper the clustering algorithm was used to classify the regions of the metal sample with defects to determine their coordinates. The informative distributed cumulative histogram is proposed. To measure sizes and intensity of defects the IDCH image is transformed and clustered.

Keywords—image intensity, surface, defects, clustering, pixel, segmentation, inversion, distributed cumulative histogram.

I. INTRODUCTION

A big number of defect determination methods differ between themselves by features and extraction algorithms. The paper [1] considers the probability of detecting size and magnitude of defects in addition to the probability of error alarms and proposes an adaptive generalized likelihood ratio (AGLR) technique. The algorithm in [2] calculates the difference between the original signal and a smooth one in the amplitude spectrum, and the defect map is then obtained by transforming the difference to spatial domain.

The approach for defect detection in [3] consists of two phases: global estimation and local refinement. First, by applying a spectral-based approach in a global manner roughly estimates defects. Second locally refines the estimated region based on the distributions of pixel intensities. The paper [4] presents an automatic system based on Hough Transform, Principal Component Analysis and Artificial Neural Networks to classify three defects with well defined geometric shapes: welding, clamp and identification hole. The paper [5] describes the algorithm that extracts local statistical features from grey-level texture images decomposed with wavelet frames.

Many papers present the image segmentation techniques using clustering [8-12]. For example, the algorithm in [8] uses k-means algorithm to split the original image into regions based on Euclidean color distance to produce an over-segmentation result. In [9] cluster analysis (TCA) method for automatic defect detection is based on three-dimensional image segmentation. Fuzzy, C-Means, K-Means clustering methods [10-13] are the most wide-spread approaches for image segmentation, pattern recognition, finding the optimal segmentation threshold and classification.

The majority of the above-mentioned approaches are quite complicated and time-consuming. In this paper, the clustering algorithm for the calculation of image intensity distribution is developed.

II. DETERMINATION OF DEFECT COORDINATES BY INTENSITY CLUSTERING

To illustrate a work of the clustering algorithm we consider the image of a metal sample with two holes (Fig. 1a) [5]. In order to obtain the lowest nodes of the tree (leaves), the input image is divided by the set of horizontal and vertical lines (Fig. 1b). For each rectangle, the relative value of the full intensity is calculated. The relation is taken to the pixel intensity from full image (all pixels intensity).

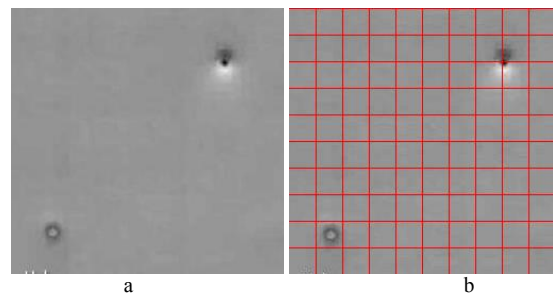


Fig. 1. Metal sample and its coverage by a grid

After the rolling up process has been performed, one more characteristic for every rectangle – its number of a cluster to which it belongs – is obtained. Fig. 2 demonstrates the clustering process and Fig. 3 shows the 6x4 part of clustered matrix containing one hole. Input data were the metal image, covered by the 10x10 grid and a number of clusters as seven. In the image each cluster is marked by a corresponding grayscale color. The clusters with higher intensity are lighter. The image of the metal sample has dimensions 250x250. Thus, each rectangle has dimension 25x25.

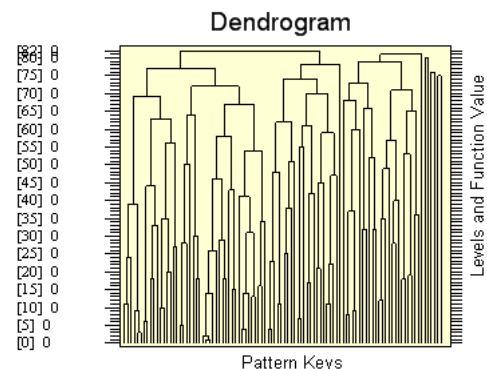


Fig. 2. Dendrogram of clustering of original image

cumulative histogram (DCH). We distinguish two types of DCH: view from OX and view from OY.

At first we calculate two distributed histograms as two sets of N (M) ordinary histograms (for every column and row of the image pixel matrix):

$$V_i(c) = \{V_{ij}(c)\}, j = 0, 255, i = 1, N \quad (1)$$

$$V_j(r) = \{V_{ji}(r)\}, i = 0, 255, j = 1, M \quad (2)$$

The distributed histogram shows frequency of pixels intensity values in columns $V_i(c)$ and in rows $V_j(r)$. In the image histograms, the OX axis shows the gray level intensities in N columns (M rows) and the OY axis shows the frequency of these intensities.

Then we calculate two distributed cumulative histograms as sets of frequency sums:

$$V_j(cc) = \left\{ \sum_{l=0}^i V_{jl}(c), i = 0, 255 \right\}, j = 1, N \quad (3)$$

$$V_j(cr) = \left\{ \sum_{l=0}^i V_{li}(r), i = 0, 255 \right\}, j = 1, M \quad (4)$$

where $V_i(c)$, $V_j(cc)$ – histogram and cumulative histogram in columns, $V_{ij}(c)$ – an intensity frequency in column, N , M – numbers of columns and rows.

A schematic example of distributed histogram is given in Fig. 7.

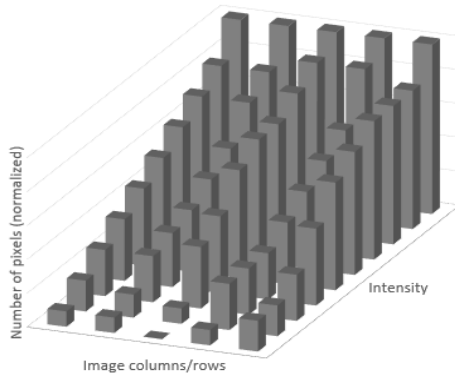


Fig. 7. Distributed cumulative histogram of abstract image

For further processing of DCH we present it by a flat 2D image on the plane OX, OI – a top view on the three-dimensional distributed histogram in Fig. 7 along the OI axis. In the new image, each value of the pixel intensity corresponds to the pixel frequency in the columns or rows given by DCH in Fig. 7:

$$\begin{aligned} I_i(c) &= 255 \times V_{i-1}(cc) / N, \\ i &= 0, 255, V_{-1}(cc) = 0 \end{aligned} \quad (5)$$

$$\begin{aligned} I_j(r) &= 255 \times V_{j-1}(cr) / M, \\ j &= 0, 255, V_{-1}(cr) = 0 \end{aligned} \quad (6)$$

For the image in Fig. 2a the DCH is presented in Fig. 8.

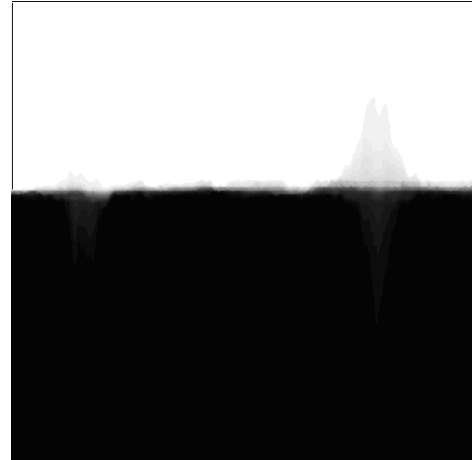


Fig. 8. Distributed cumulative histograms (view from OX plane)

In Fig. 8 it is difficult to distinguish a small number of pixels responsible for defects. To make them to be more visible we fill the closed regions of white and black colors by grey color using the flood-fill algorithm. So, we get informative part of the distributed cumulative histogram (IDCH) in Fig. 9. On it the grey color marks an absence of information. All the other colors stay unchanged.

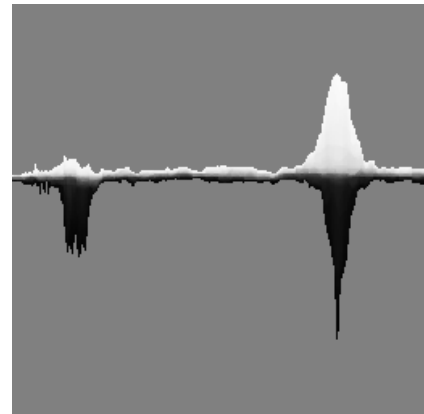


Fig. 9. Informative distributed cumulative histograms (view from OX plane)

The image of the IDCH by OX plane gives us possibility to determine the next features of defects: coordinates and sizes by axis OX, size by axis OY, integral intensity as the definite integral of the enveloping function in different intervals etc.

To exclude reciprocal influence of dark and light defects we divide IDCH into two parts: upper and lower according to color distribution – white and black. The graph of cumulative histogram helps us to find the point of division.

Then with two parts of the IDCH we do transformations similar to those earlier performed with the metal image: on the upper part we change grey color by black and the lower part we invert and change grey clot by black. As it was shown earlier it is necessary to remote intensity of informative and uninformative pixels.

In result in Fig. 10 we get two images illustrating sizes and intensity of dark and light defects on the metal surface.

In the previous chapter the clustering algorithm was used to detect and to rough measurement of defects on the metal surface. Now we apply it for more precise measurement of their intensity.

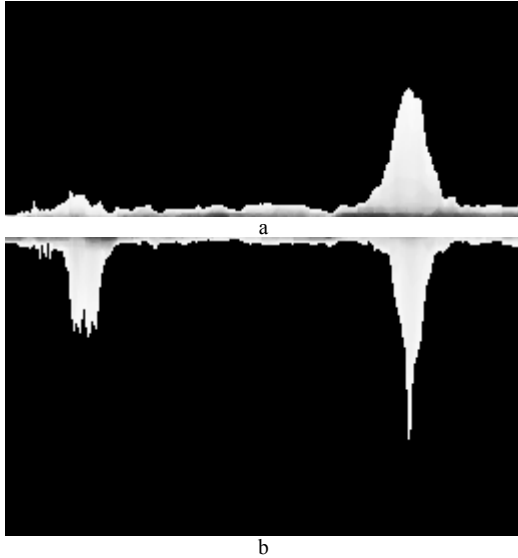


Fig. 10. Two transformed parts of informative distributed cumulative histogram: a – upper, b – lower

The algorithm is being applied to two parts of the IDCH from Fig. 10. Black color of zero intensity does not affect the intensity value of the rectangles of the image. Thus, on the clustered matrix we measure defects by a number of indexed rectangles and intensity value of every rectangles. In sum this data gives us response for the question: to reject or to accept the metal sample.

We estimate intensity of defects by indexes K_i of rectangles which form the closed area of some defect:

$$D(x) = \sum_{i=1}^f (K_f - K_i) \quad (7)$$

where K_f is the biggest index of cluster, x – coordinate of defect. Or by a sum of the rectangle intensity features:

$$S(x) = \sum_{i=1}^f I_i \quad (8)$$

Let us calculate a mean value of a pixel intensity in the j -th columns of the image pixels matrix:

$$I(j) = 1/H \sum_{i=1}^H b_{ij} \quad (9)$$

where b_{ij} is a pixel intensity in j -column ($1 \leq j \leq W$), W and H represent the number of columns and the number of rows respectively.

IV. EXPERIMENTS

In Fig. 11 we see two clustered matrices: for white and black defects. Elements with index 7 represent background.

In Fig. 11a the white defect is marked by the elements with next indices: 6, 5, 4, 4, 4, 3, 2 from the eighth column and 6, 6, 5, 4 from the ninth column. So, a value of defect is $19+7=26$. In Fig. 11b the black defects are marked by the elements with next indices: 3, 2, 1 from the second column and 5, 5, 5, 5, 5, 4, 3, 2; 5, 5, 5 from the eighth and ninth columns. So, a value of two black defects is $15+30=46$.

VI.0,002869	VI.0	VI.0	VI.0	VI.0	VI.0	VI.0	VI.0	VI.0	VI.0,006752
cIIa.7	cIIa.7	cIIa.7	cIIa.7	cIIa.7	cIIa.7	cIIa.7	cIIa.7	cIIa.7	cIIa.3
VI.0	VI.0	VI.0	VI.0	VI.0	VI.0	VI.0	VI.0	VI.0	VI.0,006752
cIIa.7	cIIa.7	cIIa.7	cIIa.7	cIIa.7	cIIa.7	cIIa.7	cIIa.7	cIIa.7	cIIa.3
VI.0	VI.0	VI.0	VI.0	VI.0	VI.0	VI.0	VI.0,01644	VI.0,003509	VI.0,006752
cIIa.7	cIIa.7	cIIa.7	cIIa.7	cIIa.7	cIIa.7	cIIa.7	cIIa.7	cIIa.6	cIIa.7
VI.0	VI.0	VI.0	VI.0	VI.0	VI.0	VI.0	VI.0,02531	VI.0,006451	VI.0,006752
cIIa.7	cIIa.7	cIIa.7	cIIa.7	cIIa.7	cIIa.7	cIIa.7	cIIa.7	cIIa.5	cIIa.7
VI.0	VI.0	VI.0	VI.0	VI.0	VI.0	VI.0	VI.0,03251	VI.0,01055	VI.0,006752
cIIa.7	cIIa.7	cIIa.7	cIIa.7	cIIa.7	cIIa.7	cIIa.7	cIIa.7	cIIa.4	cIIa.7
VI.0	VI.0	VI.0	VI.0	VI.0	VI.0	VI.0	VI.0,03226	VI.0,01688	VI.0,006752
cIIa.7	cIIa.7	cIIa.7	cIIa.7	cIIa.7	cIIa.7	cIIa.7	cIIa.7	cIIa.4	cIIa.7
VI.0,004294	VI.0	VI.0	VI.0	VI.0	VI.0	VI.0	VI.0,03698	VI.0,01863	VI.0,006752
cIIa.7	cIIa.7	cIIa.7	cIIa.7	cIIa.7	cIIa.7	cIIa.7	cIIa.7	cIIa.6	cIIa.7
VI.0	VI.0	VI.0	VI.0	VI.0	VI.0	VI.0	VI.0,04645	VI.0,027	VI.0,006752
cIIa.7	cIIa.7	cIIa.7	cIIa.7	cIIa.7	cIIa.7	cIIa.7	cIIa.7	cIIa.5	cIIa.7
VI.0,005021	VI.0,007594	VI.0	VI.0	VI.0	VI.0	VI.0	VI.0,05397	VI.0,03575	VI.0,006752
cIIa.7	cIIa.7	cIIa.7	cIIa.7	cIIa.7	cIIa.7	cIIa.7	cIIa.7	cIIa.4	cIIa.3
VI.0,04594	VI.0,07685	VI.0,04382	VI.0,04455	VI.0,04526	VI.0,04634	VI.0,04402	VI.0,08208	VI.0,07593	VI.0,0532
cIIa.3	cIIa.1	cIIa.3	cIIa.3	cIIa.3	cIIa.3	cIIa.3	cIIa.1	cIIa.1	cIIa.2

a									
VI.0,05482	VI.0,06957	VI.0,03564	VI.0,02946	VI.0,03113	VI.0,03119	VI.0,03452	VI.0,06661	VI.0,05513	VI.0,04009
cIIa.2	cIIa.1	cIIa.4	cIIa.4	cIIa.4	cIIa.4	cIIa.4	cIIa.1	cIIa.2	cIIa.3
VI.0,02317	VI.0,06331	VI.0,0004634	VI.0	VI.0	VI.0	VI.0	VI.0,05051	VI.0,03422	VI.0,005964
cIIa.5	cIIa.1	cIIa.7	cIIa.7	cIIa.7	cIIa.7	cIIa.7	cIIa.2	cIIa.3	cIIa.7
VI.0	VI.0,05788	VI.0	VI.0	VI.0	VI.0	VI.0	VI.0,04161	VI.0,01755	VI.0,005964
cIIa.7	cIIa.2	cIIa.7	cIIa.7	cIIa.7	cIIa.7	cIIa.7	cIIa.3	cIIa.5	cIIa.9
VI.0	VI.0,04439	VI.0	VI.0	VI.0	VI.0	VI.0	VI.0,03904	VI.0,01095	VI.0,005964
cIIa.7	cIIa.3	cIIa.7	cIIa.7	cIIa.7	cIIa.7	cIIa.7	cIIa.4	cIIa.6	cIIa.7
VI.0	VI.0,003841	VI.0	VI.0	VI.0	VI.0	VI.0	VI.0,02838	VI.0,00962	VI.0,005964
cIIa.7	cIIa.7	cIIa.7	cIIa.7	cIIa.7	cIIa.7	cIIa.7	cIIa.5	cIIa.7	cIIa.2
VI.0	VI.0	VI.0	VI.0	VI.0	VI.0	VI.0	VI.0,02098	VI.0,001217	VI.0,005964
cIIa.7	cIIa.7	cIIa.7	cIIa.7	cIIa.7	cIIa.7	cIIa.7	cIIa.5	cIIa.7	cIIa.7
VI.0	VI.0	VI.0	VI.0	VI.0	VI.0	VI.0	VI.0,02027	VI.0	VI.0,005964
cIIa.7	cIIa.7	cIIa.7	cIIa.7	cIIa.7	cIIa.7	cIIa.7	cIIa.5	cIIa.7	cIIa.7
VI.0	VI.0	VI.0	VI.0	VI.0	VI.0	VI.0	VI.0,01987	VI.0	VI.0,005964
cIIa.7	cIIa.7	cIIa.7	cIIa.7	cIIa.7	cIIa.7	cIIa.7	cIIa.5	cIIa.7	cIIa.7
VI.0	VI.0	VI.0	VI.0	VI.0	VI.0	VI.0	VI.0,01591	VI.0	VI.0,005964
cIIa.7	cIIa.7	cIIa.7	cIIa.7	cIIa.7	cIIa.7	cIIa.7	cIIa.5	cIIa.7	cIIa.7
VI.0	VI.0	VI.0	VI.0	VI.0	VI.0	VI.0	VI.0,01742	VI.0	VI.0,006816
cIIa.7	cIIa.7	cIIa.7	cIIa.7	cIIa.7	cIIa.7	cIIa.7	cIIa.5	cIIa.7	cIIa.7

b									
VI.0,05482	VI.0,06957	VI.0,03564	VI.0,02946	VI.0,03113	VI.0,03119	VI.0,03452	VI.0,06661	VI.0,05513	VI.0,04009
cIIa.2	cIIa.1	cIIa.4	cIIa.4	cIIa.4	cIIa.4	cIIa.4	cIIa.1	cIIa.2	cIIa.3
VI.0,02317	VI.0,06331	VI.0,0004634	VI.0	VI.0	VI.0	VI.0	VI.0,05051	VI.0,03422	VI.0,005964
cIIa.5	cIIa.1	cIIa.7	cIIa.7	cIIa.7	cIIa.7	cIIa.7	cIIa.2	cIIa.3	cIIa.7
VI.0	VI.0,05788	VI.0	VI.0	VI.0	VI.0	VI.0	VI.0,04161	VI.0,01755	VI.0,005964
cIIa.7	cIIa.2	cIIa.7	cIIa.7	cIIa.7	cIIa.7	cIIa.7	cIIa.3	cIIa.5	cIIa.9
VI.0	VI.0,04439	VI.0	VI.0	VI.0	VI.0	VI.0	VI.0,03904	VI.0,01095	VI.0,005964
cIIa.7	cIIa.3	cIIa.7	cIIa.7	cIIa.7	cIIa.7	cIIa.7	cIIa.4	cIIa.6	cIIa.7
VI.0	VI.0,003841	VI.0	VI.0	VI.0	VI.0	VI.0	VI.0,02838	VI.0,00962	VI.0,005964
cIIa.7	cIIa.7	cIIa.7	cIIa.7	cIIa.7	cIIa.7	cIIa.7	cIIa.5	cIIa.7	cIIa.2
VI.0	VI.0	VI.0	VI.0	VI.0	VI.0	VI.0	VI.0,02098	VI.0,001217	VI.0,005964
cIIa.7	cIIa.7	cIIa.7	cIIa.7	cIIa.7	cIIa.7	cIIa.7	cIIa.5	cIIa.7	cIIa.7
VI.0	VI.0	VI.0	VI.0	VI.0	VI.0	VI.0	VI.0,02027	VI.0	VI.0,005964
cIIa.7	cIIa.7	cIIa.7	cIIa.7	cIIa.7	cIIa.7	cIIa.7	cIIa.5	cIIa.7	cIIa.7
VI.0	VI.0	VI.0	VI.0	VI.0	VI.0	VI.0	VI.0,01987	VI.0	VI.0,005964
cIIa.7	cIIa.7	cIIa.7	cIIa.7	cIIa.7	cIIa.7	cIIa.7	cIIa.5	cIIa.7	cIIa.7
VI.0	VI.0	VI.0	VI.0	VI.0	VI.0	VI.0	VI.0,01591	VI.0	VI.0,005964
cIIa.7	cIIa.7	cIIa.7	cIIa.7	cIIa.7	cIIa.7	cIIa.7	cIIa.5	cIIa.7	cIIa.7
VI.0	VI.0	VI.0	VI.0	VI.0	VI.0	VI.0	VI.0,01742	VI.0	VI.0,006816
cIIa.7	cIIa.7	cIIa.7	cIIa.7	cIIa.7	cIIa.7	cIIa.7	cIIa.5	cIIa.7	cIIa.7

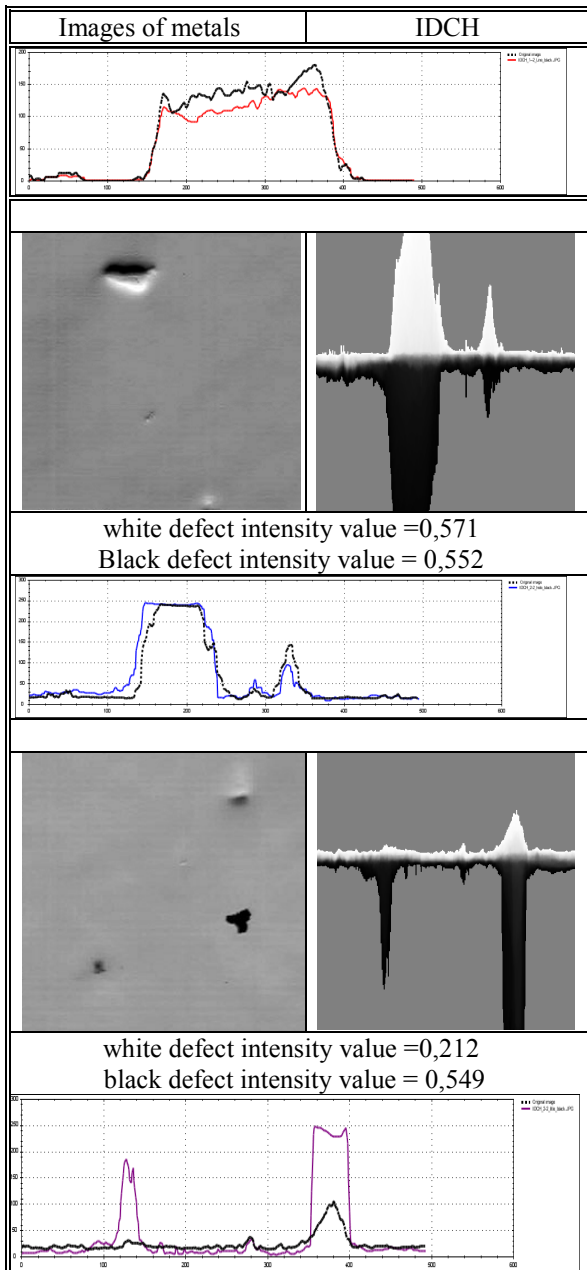
Fig. 11. Two clustered matrixes of transformed informative distributed cumulative histograms: a – upper, b – lower

Calculating intensity we have in the first case $S(x) = 0,307$ and in the second case $S(x) = 0,166 + 0,270 = 0,436$. Defect measurement by intensity values is more accurate for the reason that every rectangle contains only information connected with defects.

Some experiments were held with metal (scratches and holes), paper (creases) [6, 7]. They are given in Table IV with their IDCH images. Also in the Table IV calculated intensity values of white and black defects for every metal sample are given. To confirm the accuracy of calculation we use the formula (9) to get the graphs of mean intensity in the columns of transformed IDCH images. We can see that dark and light small anomalies in the metal samples appear in histograms and could be compared with etalon samples.

TABLE IV. INFORMATIVE DISTRIBUTED CUMULATIVE HISTOGRAMS OF DEFECTED MATERIAL SAMPLES

Images of metals	IDCH
white defect intensity value = 0,710 black defect intensity value = 0,688	



V. CONCLUSION

The intensity features and coordinates of defects on the metal surface were obtained by clustering algorithm applied

to the metal image. The informative distributed histogram on a base of distributed histogram is proposed for more precise determination of intensity of metal defects. The IDCH image transformation and clustering algorithm were used for these purposes. The developed software allows to analyze the images of different materials.

REFERENCES

- [1] L.J. Wells, M. S. Shafae and J.A. Camelio, "Automated Surface Defect Detection Using High-Density Data," *J. Manuf. Sci. Eng.*, 138(7), Mar, 2016.
- [2] I. Ahn and Ch. Kim, "Finding Defects in Regular-Texture Images," *16th Korea-Japan Joint Workshop on Frontiers of Computer Vision*, Hiroshima, Japan, pp. 478-480, Feb. 2010.
- [3] J. Choi and Ch. Kim, "Unsupervised Detection of Surface Defects: A Two-Step Approach," *IEEE International Conference of Image Processing (ICIP)*, Orlando, USA, pp. 1037-1040, Sep. 2012.
- [4] L.A.O. Martins, F.L.C. Padua, and P.E.M. Almeida, "Automatic Detection of Surface Defects on Rolled Steel Using Computer Vision and Artificial Neural Networks IECON," *36th Annual Conference on IEEE Industrial Electronics Society*, pp. 1081-1086, 2010.
- [5] S. Jahanbina, A.C. Bovika, E. Perezb, and D. Nair, "Automatic Inspection of Textured Surfaces by Support Vector Machines," [Electronic resource] Link: <https://live.ece.utexas.edu/publications>
- [6] Wintriss defects gallery [Electronic resource] Link: <http://www.weco.com/surface-inspection>
- [7] Defect detection of various films [Electronic resource] Link: <http://cilabs.kaist.ac.kr/research/image-analysis/defect-detection>
- [8] V.H. Pham, and B.R. Lee, "An image segmentation approach for fruit defect detection using k-means clustering and graph-based algorithm," *Vietnam Journal of Computer Science*, vol. 2, iss. 1, pp 25-33, February 2015,
- [9] K. Zheng, Y.-S. Chang, K.-H. Wang, and Y. Yao, "Thermographic clustering analysis for defect detection in CFRP structures," *Polymer Testing*, vol. 49, pp. 73-81, February 2016.
- [10] R. Xu, and D. Wunsch, "Survey of clustering algorithms", *IEEE Transactions on Neural Networks*, vol. 16, iss. 3, pp. 645 - 678, May 2005.
- [11] S. Naz, H. Majeed, and H. Irshad, "Image segmentation using fuzzy clustering: A survey," *6th International Conference on Emerging Technologies (ICET)*, pp. 181 - 186, 18-19 Oct. 2010.
- [12] S. Thilagamani and N. Shanthi, "A Survey on Image Segmentation Through Clustering," *International Journal of Research and Reviews in Information Sciences*, vol. 1, no. 1, pp. 14-17, March 2011.
- [13] Y. Yang, D. Xu, F. Nie, S. Yan, and Y. Zhuang, "Clustering Using Local Discriminant Models and Global Integration," *IEEE Transactions on Image Processing*, vol. 19, iss. 10, pp. 2761 - 2773, Oct. 2010

Image Contrast Enhancement Using a Modified Histogram Equalization

Sergei Yelmanov
Special Design Office of Television Systems
Lviv, Ukraine
sergei.yelmanov@gmail.com

Yuriy Romanyshyn^{1,2}
¹Department of Electronics and Computer Technologies
¹Lviv Polytechnic National University
²University of Warmia and Mazury
¹Lviv, Ukraine, ²Olsztyn, Poland
yuriy.romanyshyn1@gmail.com

Abstract— The histogram equalization is the basic technology in image processing, which is widely used to increase the integral contrast of the image and is characterized by high efficiency. However, known techniques of histogram equalization have a number of disadvantages, the major one of which is a decrease in the contrast of small objects in the image. To address this disadvantage, a new technique of modified histogram equalization is proposed based on the assessment of the two-dimensional probability distribution of brightness in the image. A new method of increasing the contrast of the image based on the modified equalization of histogram is also proposed. The proposed technique of modified histogram equalization can be recommended to increase the image contrast in automatic mode in imaging and image processing.

Keywords—image, contrast, histogram, modified equalization.

I. INTRODUCTION

The widespread use of modern digital technologies in image processing and image processing requires the solution of the task of effective improving of image quality in automatic mode [1, 2].

The histogram equalization is the basic techniques to enhance the image quality [3]. The histogram equalization technique and its modifications are widely used for images processing to increase integral contrast of image and are characterized by high efficiency [4]. However, known techniques of histogram equalization have a number of disadvantages. Their main disadvantage is a decrease in the contrast of small-sized objects and their details in the image.

To address this disadvantage, we propose a new technique of image enhancement based on the modified equalization of histogram.

In this work solves the task of increasing the efficiency of processing of multi-element images. The problem of increasing the image contrast based on histogram transformations is considered (Section II). A new technique of modified equalization of image histogram is proposed based on the assessment of the two-dimensional probability distribution of brightness (Section III). A new method of increasing the contrast of the image based on the modified equalization of histogram is also proposed.

Research of known and proposed methods of images processing is carried out by measuring the contrast for the results of processing test images (sections IV and V).

II. IMAGE CONTRAST ENHANCEMENT

Various approaches to enhance the contrast of image are known at present [1]. Techniques of processing in spatial domain using nonlinear statistical no-inertial transformations of brightness are of the greatest interest to the image processing in automatic mode and in real time mode [1, 3].

The most widely used method of image pre-processing is the linear stretching of the dynamic range of brightness to the range of its possible values [1]:

$$y_i = y_{\min} + \frac{y_{\max} - y_{\min}}{x_{\max} - x_{\min}} \cdot (x_i - x_{\min}), \quad (1)$$

where x_{\min} , x_{\max} , y_{\min} , y_{\max} - minimum and maximum values of brightness for the source and processed images:

$$x_i, x_{\min}, x_{\max}, y_i, y_{\min}, y_{\max} \in [0, 1]. \quad (2)$$

Generally, it is assumed that $y_{\min} = 0$, $y_{\max} = 1$.

In [5], a method of nonlinear stretching of the dynamic range of images using the sigmoid function was proposed:

$$y_i = y_{\min} + \frac{y_{\max} - y_{\min}}{x_{\max} - x_{\min}} \cdot \frac{1}{1 + e^{-\alpha(x_i - \beta)}}, \quad (3)$$

where α , β - parameters of the sigmoid function.

Another widely used approach to contrast enhancement is power law transformation, also known as gamma correction [1]:

$$y_i = y_{\min} + (y_{\max} - y_{\min}) \cdot \left(\frac{x_i - x_{\min}}{x_{\max} - x_{\min}} \right)^\gamma, \quad (4)$$

where γ - exponent, parameter.

The histogram equalization is the most well-known standard procedure in image processing [1, 4]:

$$y_i = y_{\min} + (y_{\max} - y_{\min}) \cdot \int_0^{x_i} p(x) dx, \quad (5)$$

where $p(x)$ - distribution of brightness.

However, the discussed methods (1), (3)-(5) of image enhancement and their modifications have several significant disadvantages.

The main disadvantage of the known methods of image enhancement by histogram equalization is a decrease in the contrast of small-sized objects.

To address these disadvantages, we propose a new method of modified equalization of histogram based on the assessment of the two-dimensional probability distribution of brightness.

III. PROPOSED METHOD

We propose a new approach to the modified equalization of histogram on the basis of the assessment of the two-dimensional probability distribution of brightness.

A new method of image processing based on the modified equalization of histogram is also proposed.

Image enhancement is carried out by the modified histogram equalization based on the assessment of the two-dimensional probability distribution of brightness in image.

Proposed method is an adaptive transformation in the form:

$$y_i = \alpha \cdot \int_0^1 \int_0^1 \omega(x_k, x_n) \cdot \varphi_{x_i}(x_k, x_n) dx_k dx_n, \quad (6)$$

where $\omega(x_k, x_n)$ – two-dimensional density distribution of the brightness; $\varphi_{x_i}(x_k, x_n)$ – weight function; α – normalizing coefficient, multiplier.

The weight function $\varphi_{x_i}(x_k, x_n)$ (6) characterizes the measure of proximity between two values of brightness x_i and x_k (or x_n) on a preset interval (x_k, x_n) of brightness and is defined as:

$$\varphi_{x_i}(x_k, x_n) = \begin{cases} 0, & \text{if } x_k > x_i \wedge x_n > x_i \\ \frac{\rho(x_k, x_i)}{\rho(x_k, x_n)}, & \text{if } x_k \leq x_i \wedge x_n > x_i \\ \frac{\rho(x_n, x_i)}{\rho(x_n, x_k)}, & \text{if } x_k > x_i \wedge x_n \leq x_i \\ 1, & \text{if } x_k \leq x_i \wedge x_n \leq x_i \end{cases}, \quad (7)$$

where $\rho(x_k, x_i)$ – measure of proximity for two values x_k and x_i of brightness.

Normalizing coefficient α (6) is equal to:

$$\alpha = \left[\int_0^1 \int_0^1 \omega(x_k, x_n) dx_k dx_n \right]^{-1}. \quad (8)$$

Expressions (6) - (8) define the proposed method of modified histogram equalization in general terms.

To demonstrate the proposed technique of modified histogram equalization let's assume that the assessment of two-dimensional distribution for the brightness has the form:

$$\omega(x_k, x_n) = p(x_k)^\gamma \cdot p(x_n)^\gamma, \quad (9)$$

where $p(x_k)$ – distribution of brightness of source image, γ – parameter.

Assume also that the measure $\rho(x_k, x_i)$ of proximity for two brightness values is defined as:

$$\rho(x_i, x_k) = |x_i - x_k|. \quad (10)$$

Expressions (6)-(8) and (9)-(10) define the proposed method to increase the image contrast using modified equalization of image histogram.

Research of known and proposed methods of image enhancement by histogram transformation are carried out in Sections IV and V.

IV. RESEARCH

Research was carried out by measuring the contrast of test images and the results of their processing using contrast metrics and expert estimates of perceived contrast.

The processing of the source images was carried out using six different methods:

- 1) linear stretching (1) [1], $\alpha = 0.01$ (1%);
- 2) nonlinear stretching using sigmoid funktion (3) [5];
- 3) gamma correction (4) [1], $y_{mean} = 0.5$;
- 4) histogram equalization (5) [1, 4];
- 5) BBHE [6];
- 6) modified histogram equalization (6)-(10).

The four source images are shown in Fig. 1, Fig. 3, Fig. 5 and Fig. 7. The results of processing for the source images with using earlier considered methods are shown in Fig. 2, Fig. 4, Fig. 6 and Fig. 8.

To measure the contrast of images, the no-reference metrics of contrast were used:

- 1) complete integral weighted contrast [7]:

$$C_{gen}^{wei_1} = \int_0^1 \int_0^1 \frac{|x_i x_j - \bar{x}|}{x_i x_j + \bar{x}} p(x_i) p(x_j) dx_i dx_j, \quad (11)$$

where \bar{x} – average value of brightness for the initial image.

- 2) generalized weighted contrast [8]:

$$C_{gen}^{wei_2} = \int_0^1 \int_0^1 \frac{|x_i - x_j|}{x_i + x_j} p(x_i) p(x_j) dx_i dx_j, \quad (12)$$

- 3) generalized integral linear contrast [8]:

$$C_{gen}^{lin} = \int_0^1 \int_0^1 \frac{|x_i - x_j|}{x_{max} - x_{min}} p(x_i) p(x_j) dx_i dx_j, \quad (13)$$

- 4) incomplete integral weighted contrast [8]:

$$C_{inc}^{wei} = \int_0^1 (x - \bar{x}) / (x + \bar{x}) p(x) dx, \quad (14)$$

- 5) incomplete integral linear contrast [9]:

$$C_{inc}^{lin} = \int_0^1 \left| \frac{x - \bar{x}}{x_{mpv}} + \frac{1}{2} - \left| \frac{x - \bar{x}}{x_{mpv}} - \frac{1}{2} \right| \right| p(x) dx, \quad (15)$$

where x_{mpv} - maximum possible value of brightness.

The results of measuring the integral contrast for test images and for results of their processing are shown in Fig. 9, Fig. 10, Fig. 11 and Fig. 12. Analysis of results of the research is carried out in Section V.

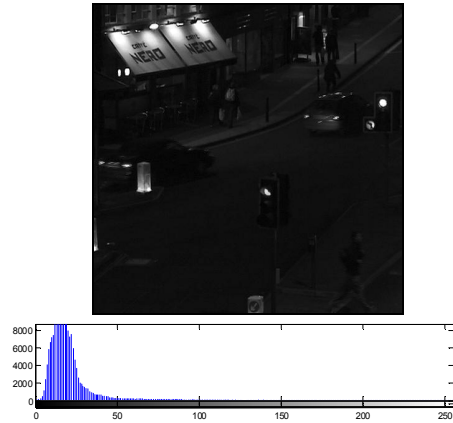


Fig. 1. The source image A.

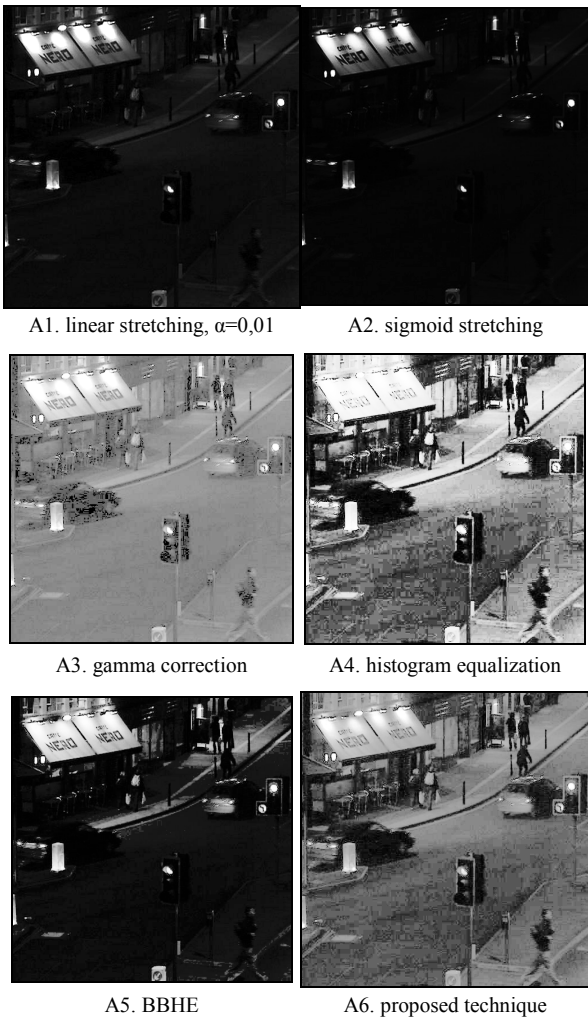


Fig. 2. Processed images.

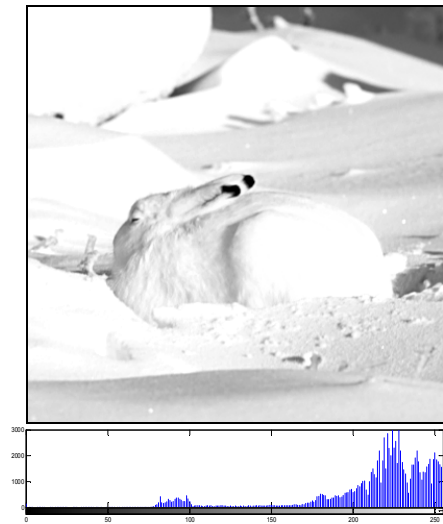


Fig. 3. The source image B.

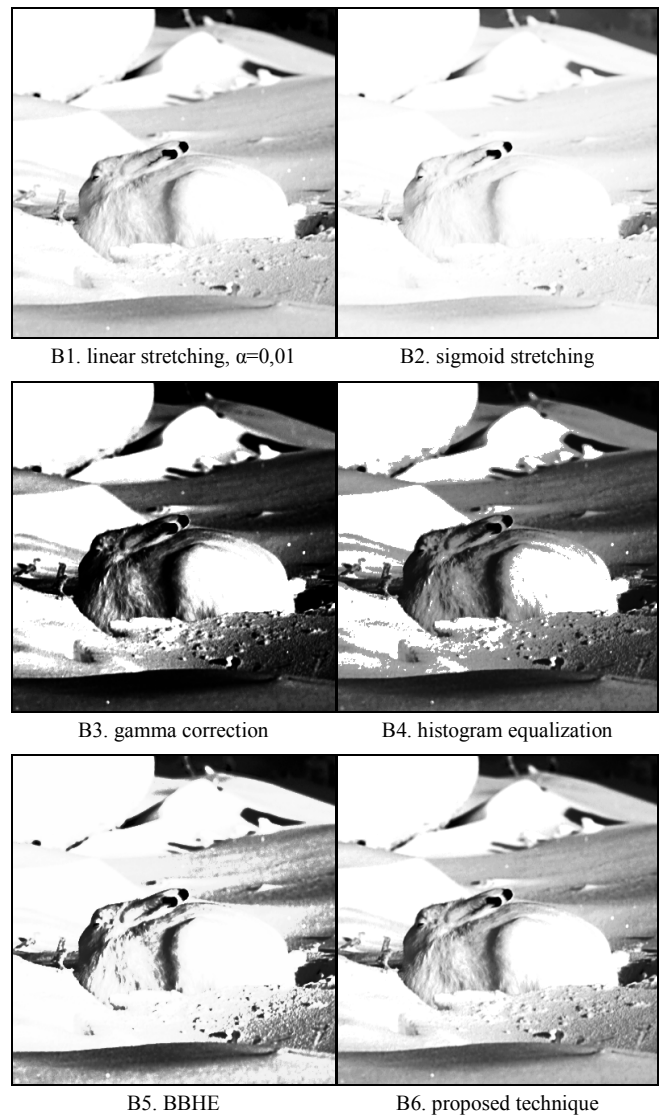


Fig. 4. Processed images.

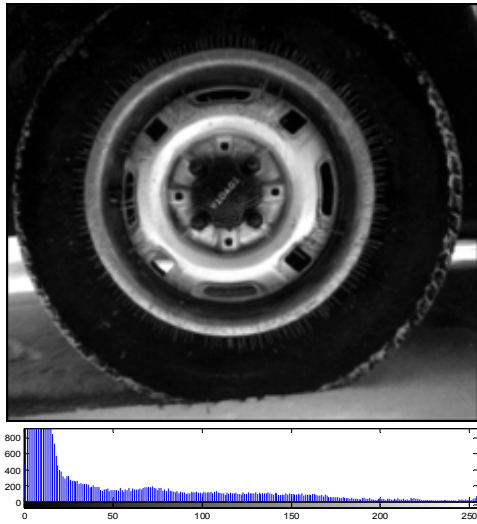


Fig. 5. The source image D.

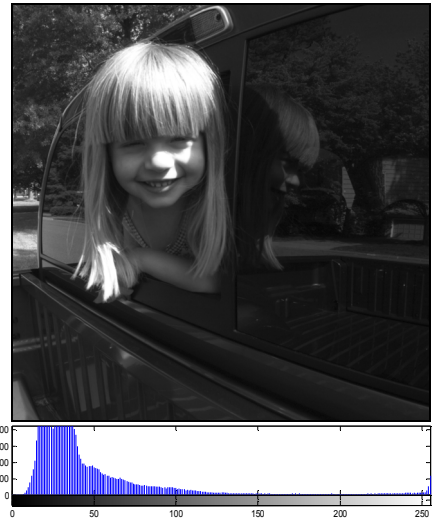
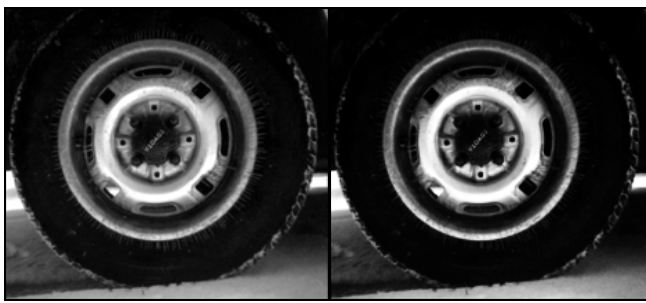


Fig. 7. The source image E.



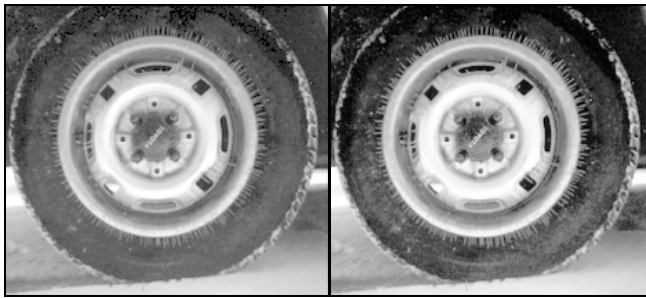
D1. linear stretching, $\alpha=0,01$

D2. sigmoid stretching



E1. linear stretching, $\alpha=0,01$

E2. sigmoid stretching



D3. gamma correction

D4. histogram equalization



E3. gamma correction

E4. histogram equalization



D5. BBHE

D6. proposed technique



E5. BBHE

E6. proposed technique

Fig. 6. Processed images.

Fig. 8. Processed images.

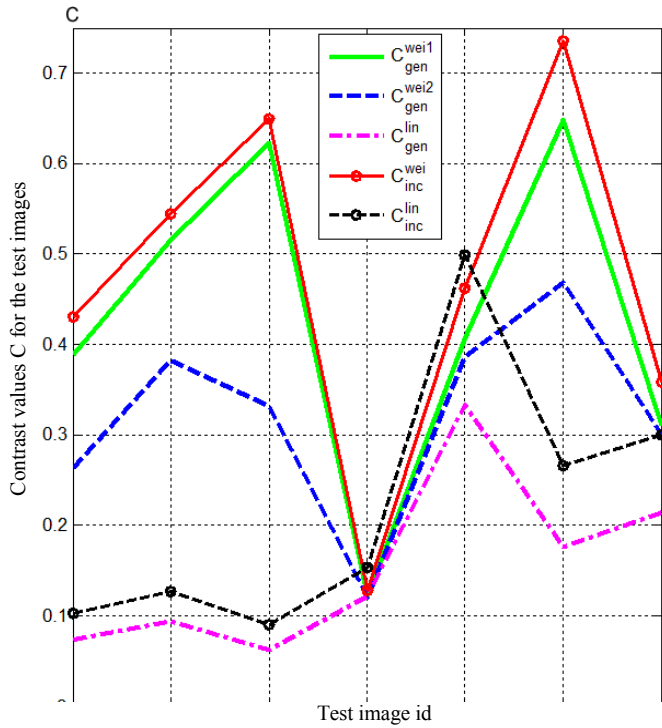


Fig. 9. Contrast for processed images

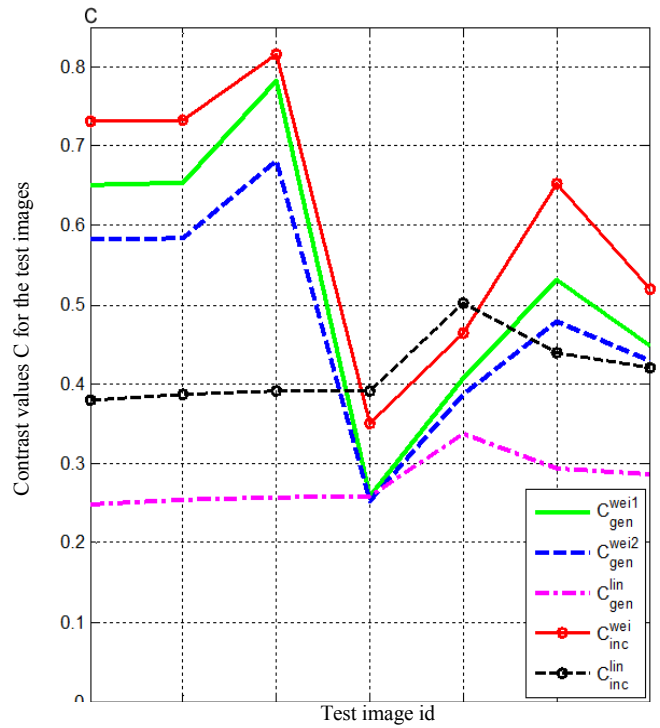


Fig. 11. Contrast for processed images

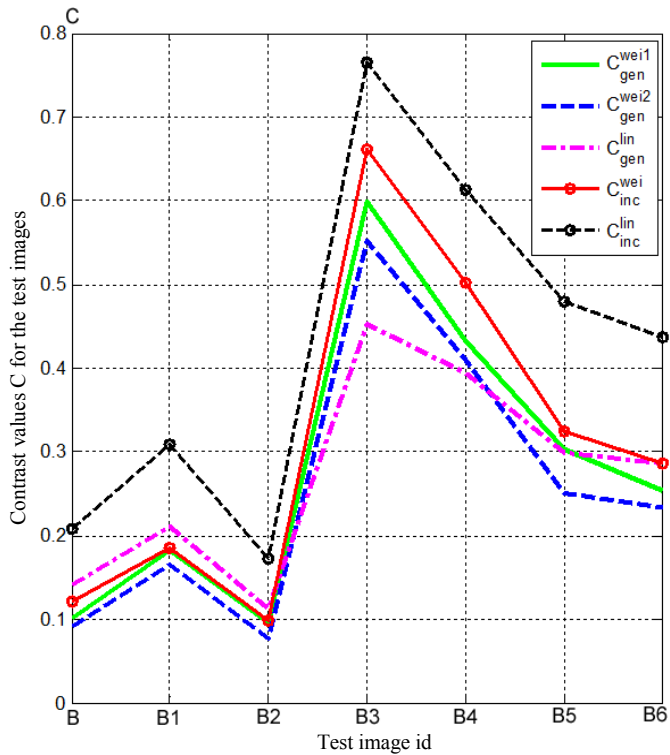


Fig. 10. Contrast for processed images

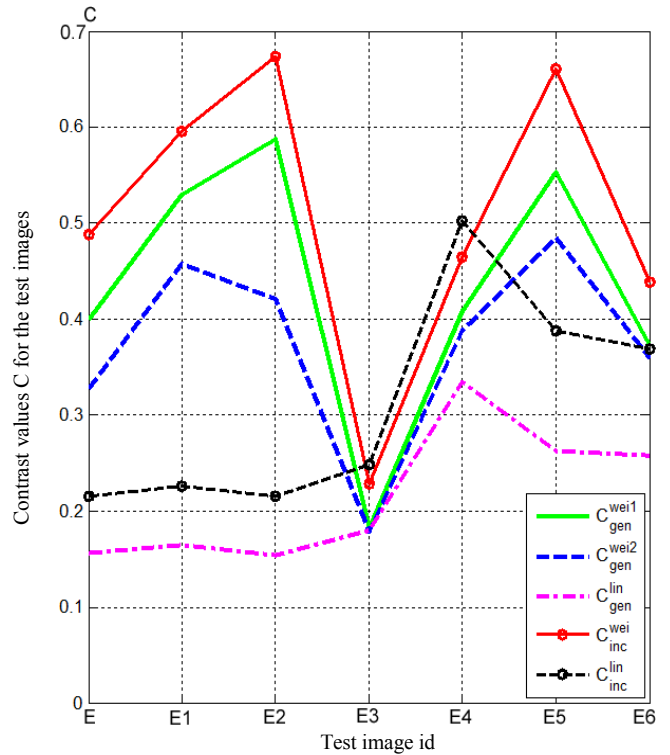


Fig. 12. Contrast for processed images

V. DISCUSSION

The methods of image enhancement by histogram transformation are widely used for images processing to increase their integral contrast and are characterized by high efficiency.

The techniques of image processing by histogram equalization (5) provide the most effective enhancement of image contrast and are widely used to improve the image quality.

At present, histogram equalization is the basic technique of image enhancement.

However, known techniques of histogram equalization have a number of disadvantages that substantially limit their practical use in the automatic mode.

The main disadvantages of the known methods of image enhancement by histogram equalization are an overly increase in the contrast of large-sized objects in the image.

The efficiency of techniques of linear stretching (1), non-linear stretching using the sigmoid function (3) and gamma correction (4) for images enhancement depends significantly on the distribution of image brightness and the values of parameters of the transformation function which significantly limits the use of these methods.

The use of modified equalization of histogram based on the assessment of the two-dimensional probability distribution of brightness allows increasing the efficiency of contrast enhancement.

The proposed method for image processing by modified equalization of histogram increases the contrast of small-sized objects and the integral contrast for all test images.

VI. CONCLUSION

At present, the solution of the task of enhancement of image quality with an acceptable level of computational costs is urgent as never before.

The histogram equalization is currently the basic technique of image processing.

The technique of histogram equalization and its modifications are widely used for images processing to increase image contrast and are characterized by high efficiency.

Histogram equalization techniques provide an effective increase of the generalized contrast and enhance the objective quality of the image.

However, known techniques of histogram equalization have a number of disadvantages that limit their use for image processing in the automatic mode.

Their main disadvantages are an overly increase in the contrast of large-sized objects and also a decrease in the contrast of small-sized objects in the image.

To address these disadvantages, a new technique of modified histogram equalization was proposed.

The proposed technique of modified histogram equalization is based on the assessment of the two-dimensional probability distribution of brightness.

To demonstrate the possibilities of the proposed technique, a new method to increasing the integral contrast of complex monochrome images based on the modified equalization of histogram was proposed.

The proposed method provides an effective increase in the contrast of small objects in the image, and also increases the integral contrast of the complex multi-element image as a whole.

The research of the proposed method of modified equalization of histogram was carried out.

The research was carried out by analyzing the results of the measurement of integral contrast for monochrome test images that were processed using the known and proposed methods of image enhancement.

To measure the integral contrast of the test images, the metrics of contrast were used.

The proposed technique of modified histogram equalization provides effective enhancement of image quality and can be recommended to increase the integral contrast in imaging and image processing in automatic mode.

REFERENCES

- [1] R. C. Gonzalez, R. E. Woods, S. L. Eddins, Digital image processing using MATLAB, Prentice-Hall, Inc., NJ, USA, 2003.
- [2] R.J. Kosarevych, B.P. Rusyn, V.V. Korniy, T.I. Kerod, "Image Segmentation Based on the Evaluation of the Tendency of Image Elements to form Clusters with the Help of Point Field Characteristics", *Cybernetics and Systems and Analysis*, vol.51, issue5, 2015, pp.704-713.
- [3] R. Ghabousian, N. Allahbakhshi, "Survey of Contrast Enhancement Techniques based on Histogram Equalization," *Int. J. Rev. Life Sci.*, vol. 5, no. 8, pp. 901-908, 2015.
- [4] R. Hummel, "Image enhancement by histogram transformation," *Comp. Graph. Image Process.*, vol. 6, pp. 184-195, 1977.
- [5] N. Hassan and N. Akamatsu, "A new approach for contrast enhancement using sigmoid function", *International Arab Journal of Information Technology*, vol. 1, no. 2, pp. 221-225, 2004.
- [6] Y. Kim, "Contrast enhancement using brightness preserving bi-histogram equalization", *IEEE Transactions. Consumer Electronics*, Vol. 43, No. 1, pp. 1-8, 1997.
- [7] V.F. Nesteruk, V.A. Sokolova, "Questions of the theory of perception of subject images and a quantitative assessment of their contrast", *Optiko-electronic industry.* - 1980. - no. 5. - pp. 11-13.
- [8] S. Yelmanov, Y. Romanyshyn, "A Method for Rapid Quantitative Assessment of Incomplete Integral Contrast for Complex Images", in *Proceedings of 14th Int. Conf. on Advanced Trends in Radioelectronics, Telecommunications and Computer Engineering (TSET'2018)*, Lviv - Slavske, Ukraine (2018).
- [9] R.A. Vorobel, Loharyfmichna obrobka zobrazen' [Logarithmic Image Processing], *Naukova Dumka* (2012), Kyiv, Ukraine.

Development and Implementation of Human Face Alignment and Tracking in Video Streams

Yevhen Zadorozhnyi

*Information Control Systems and Technologies Department
Uzhhorod National University
Uzhgorod, Ukraine
MrDindows@gmail.com*

Tetiana Fedoronchak

*Software Tools Department
Zaporizhzhia National Technical University
Zaporizhzhia, Ukraine
t.fedoronchak@gmail.com*

Yevhenii Tverdokhlib

*Software Tools Department
Zaporizhzhia National Technical University
Zaporizhzhia, Ukraine
junta.kristobal@gmail.com*

Natalia Myronova

*Software Tools Department
Zaporizhzhia National Technical University
Zaporizhzhia, Ukraine
natali.myronova@gmail.com*

Abstract—The paper presents a method that allows detection, alignment and tracking of a human face in a real time in video streams. To detect and to align face on an image a face shape regression approach is used. The developed method uses scanning window, a cascade of ensembles of regression and classification trees, and adaptive boosting. The same trees are used for classification whether the given window contains a face and for regression of a face shape. For face tracking a starting position for face search is taken from the found shape on the previous frame. Conducted analysis of the proposed method implementation gave good performance results but revealed shortcomings and directions for future work. Sensitivity of face detection is 78% and accuracy of face alignment is 95%. The implementation can track faces in real time with a speed of at least 20 frames per second.

Keywords—*regressor, cascade, face, search, marking, landmarks, classification, optimization, opencv, machine learning, machine vision.*

I. INTRODUCTION

With the development of information technology, there arises a need to solve new non-standard tasks. Machine vision is one of the areas where a large number of topical challenging tasks is issued. One of the typical tasks for machine vision is human face alignment and tracking. A lot of examples of practical usage of face tracking can be found. Tracking is actively used in the cinema industry, in the production of advertisements, 3D animated cartoons and computer games. Tracking also can be applied to define facial expressions of a person, in image stabilization algorithms, for person identification or even in video chats.

The tracking task implies determination of the location and orientation of a moving object in a virtual environment. To align and to track a face in video stream a method should solve the following tasks: face detection; markup of key points on the face for tracking; monitoring the position of the points tracked. Human face is usually tracked not just by its position as a rectangle or an oval, but also by the face model made up of the key points. The main problem in tracking is to compare the positions of the target object on the sequence of frames, especially if the object moves fast comparatively to the frame rate.

The aim of this work is to develop and implement the method that allows to detect and to track human faces in real time in video streams with a high degree of accuracy of recognition through the use of ensembles of random decision trees. The developed method can be used in many spheres from robots vision to entertainment applications.

The majority of traditional approaches to detection of faces on images is implemented with scanning window and classification. The main difference among these methods is in classification algorithms used. The result of work of such methods is the square or rectangle in which the face is likely to appear. If the algorithm requires speed performance, then the resulting rectangle can only be at a zero angle of rotation. For the most of practical tasks this is enough. But there also are some specific tasks that require more detailed information like the model of the face that is tracked. In this work a decision was made to choose the approach to detect 2D key points of faces.

For automatic alignment the face is encoded with 2D key points called landmarks. The landmarks are tied to the position of face contour, mouth, nose, eyes and eyebrows. The set of landmarks is fixed, each of them has its own number and together they form a face model. By the position of landmarks it is possible to determine whether the eyes are opened, in which direction the person's face is directed and under what angle it is turned, whether the mouth is open, whether there is a smile on his face, and what emotions in general the face reflects. The set of model landmarks in a particular alignment is called a shape. A classic approach for face alignment is given the approximate position of a person's face to shift or align the landmarks to their true positions on the image.

The task of face alignment can be referred to as the class of optimization tasks and it is advantageous to use sequential regression algorithms to solve it. The regression algorithms can be combined with the classification. Such approach will allow solving simultaneously the classification problem for face detection and the alignment problem. The result will represent not only the rectangle that limits the face, but also more detailed information with the coordinates of the landmarks of the face, that determine its contour, angles of rotation, emotions, eye state, etc.

The task of face tracking cannot be separated from the task of automatic alignment, since it is just the same task of automatic alignment, where the starting position of search is taken from the found shape on the previous frame.

II. METHODS USED FOR FACE DETECTION AND ALIGNMENT

Active Shape Model (ASM) is a statistical deformable model of the shape of objects [1, 2]. Shape of the face is presented by the set of points distributed in accordance with the model. Through several iterations the shape is deformed and position is changed to fit a face on a given image. The method learns from labeled examples of faces in the training samples allowable variations of the model. The aim of the algorithm is to match a face model with a new image. ASM algorithm consists of two steps that alternate. The first step is a search in the image around each point for a better position. The search could be performed by looking for strong edges on the image or with statistical model of profile that represents information about the image structure near the landmarks. The second step is an update of the model parameters by maximizing the conditions with the newly found positions.

ASM is closely related to the active appearance model (AAM). In faces analysis the pigmentation properties are of the same importance as the properties of the shape. AAM uses statistical model that combines an object shape model and a model of the appearance variations of shape-normalized face [3, 4]. AAM allows to take into account the texture in the context of the shape of a face. Authors revealed correlations between the shape and grey-level variations in images. Model is built during the training stage using a set of images along with the coordinates of the landmarks on them. The algorithm uses the difference between the current evaluation of the appearance and the target on the image using the optimization process. With the least squares method the method can adjust to new images very quickly.

Both ASM and AAM are widely used for the analysis of facial images, mechanical nodes and medical images.

Face alignment via regressing local binary features [5] is an approach to face shape regression based on training of local binary features. It is proposed to teach the linear regression matrix and face feature mapping function in two consecutive steps. The first step is to train feature mapping function to generate local binary features for each key point. The global mapping function is composed of a set of independent local feature mapping functions learned by independently regressing each landmark in the local region. The second step is to train the linear projection matrix with linear regression to estimate global shape. This process is repeated step by step using a cascade of regressors.

Ensemble of regression trees or random forest [6] is one of the algorithms used for the task of automatic face alignment. Weak classifiers are used at the nodes of the tree. In image processing tasks it can be a comparison of the pixel differences from the current descriptor with a fixed value. At the same time, this algorithm involves creation of a large number of low random trees, the total result of which gives a good overall performance. Such a forest will act as one regressor.

There are two approaches to learn the ensemble of random trees: sequential and parallel. In the parallel approach, each tree is learnt independently, and at the use

stage, the arithmetic average is taken from the output of each tree. Another approach is to sequentially learn trees, so that each tree takes into account the result of the previous one. But these results are taken not fully, but with some small coefficient, otherwise the small size of the trees will increase the variance in the training data, which will worsen future learning. This approach in many tasks gives a better result due to system adaptability. In the case of using the forest for classification tasks, it is supplemented by AdaBoost algorithm [7, 8].

The study of existing methods for detection of human faces and automatic face alignment and real-time tracking allowed us to formulate the goal of research. It is observed that popular and effective methods for solving both problems, such as alignment and detection, are approached based on the cascade of weak structures (classifiers or regressors). Therefore, we can make a conclusion that the idea of solving both problems by a single approach and a single algorithm can be implemented. The idea of this work is to combine the algorithms of detection, tracking and automatic alignment of landmarks of the face.

III. DEVELOPMENT OF FACE TRACKING METHOD

The method to be developed should fulfill certain requirements. The main and obvious requirement is the quality of face detection and recognition. This can be estimated by the percentage of errors on the test data set. The second requirement is speed. The developed method and its implementation should provide the opportunity to solve the tasks in real time. As it is a machine learning task, it must support scaling, change of the type of input or landmarks model, and the like. The method should be able to be configured in accordance with the capabilities of the hardware system or the needs of the system that uses it.

In accordance with the set tasks and the analysis of existing algorithms and methods, it was decided to create a method that combines in its implementation several methods of the described above. Definition of tasks and requirements for the face tracking method allowed to choose optimization and machine learning algorithms that will solve determined problems. The key algorithms are a cascade of ensembles of regression and classification trees, and adaptive boosting, through which the classifiers will be taught.

Let's describe the developed method of detection, alignment and tracking of a face on an image, as well as the algorithm of training the cascade of regressors.

The scanning window is a rather primitive, but efficient algorithm, which consists in consistent checking of rectangles from the image with different offsets and sizes of this window. This will allow to reach a person's face in any part of the image. The image in the current position of the scanning window will be called a sample.

The idea behind the cascade of regressors is to use multiple consecutive regressors. Each regressor initially calculates certain information from the image and the current position of landmarks, and converts this information into the result of classification and transformation, which should be applied to the current positions of the sample landmarks.

The algorithm of work will work as follows.

Requires: image I , initial form S_0 .

Returns: calculated form S_T .

Cycle: for t from 1 to T :

1) $F_T = h(I, S_{T-1})$, – descriptor;

2) $\Delta S = R_T(F_T)$, – transformation, that is calculated by the regressor;

3) $S_T = S_{T-1} + \Delta S$, – update of the form.

End of the cycle.

The face model or the form being transformed was chosen to contain 68 landmarks. It is symmetric with respect to the central axis of the face, as shown on Fig. 1.

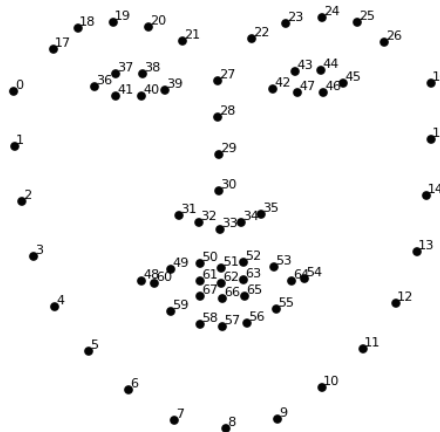


Fig. 1. Schematic image of the model with 68 landmarks

On Fig. 2 the process of iterative transformation of a shape form of a face is presented. It is shown how the initial form is regressed to a true form of the face on the image.

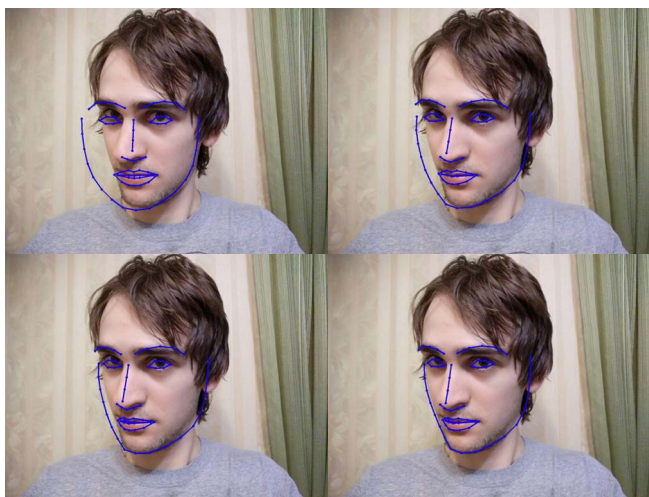


Fig. 2. Iterative transformation of a shape

Given the speed requirements, it was decided to use the values of random pixels from a rectangle that limits the landmarks of the current state as features of the image. It should be noted that there are many algorithms for finding and describing local image peculiarities such as SIFT [9], SURF [10], HOG, GLOH, Gaussian filters, Gabor filters, Canny filter, discrete cosine transformation and others. All these algorithms give much more informative descriptors, but at the same time they require much more time for their

calculation, which makes it impossible to use them with high speed in this task. The set of pixel values used in the current step we will call a descriptor.

Advantage of ensemble of regression trees is that the weak classifiers contained in the nodes of a tree can be just comparisons of the pixel difference from the current descriptor with a fixed value. This allows to achieve high-speed algorithm, because the comparison of two values is a quick and elementary operation. At the same time, due to the speed, it is possible to use a high number of such random small trees, the total result of which gives a good overall quality. The structure of this forest is shown in Fig. 3.

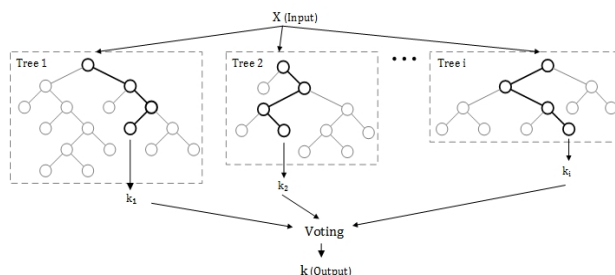


Fig. 3. The structure of forest

The regression itself is calculated by the addition of transformations contained in the leaves of trees, in which the current sample walks into.

In addition to regression, a classification is also required. For classification, the same random decision trees will be used. We can classify into what leaves descriptors of the current object walked. The process of classification will be as follows: each leaf of the tree will have its own value for the classification. During a forest walk, we need to sum up the values in the leaves we visited and draw conclusions according to the value of the sum amount received.

This approach also allows to use the preliminary interruption of the algorithm, in the case where the current accumulated amount already allows us to conclude that the current sample is not a face. As a training algorithm one of AdaBoost modifications called RealBoost [11] is used.

Given that the face on the images can be located in any part of it, at any angle and on any scale, the information about the landmarks should be unified [12] through conversion into some local coordinates. This local coordinate system (Fig. 4) can be found by solving the similarity transformation problem, which converts the points of a typical average model for a given set of landmarks into coordinates such that the sum of distances to the corresponding current landmarks is minimal. This task is reduced to the usual method of least squares.

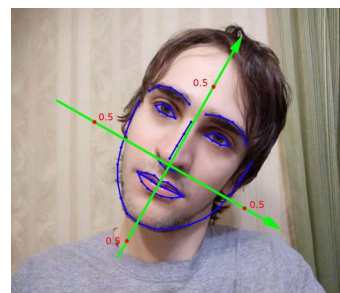


Fig. 4. Form with its local coordinate system

IV. DETECTOR AND REGRESSOR TRAINING

For the implementation of the given method and the training of the detector and regressor, a base with a sufficient number of already marked images containing faces is required. For the whole set of labeled images the average model alignment is counted. This average model is the arithmetic mean of points in their local coordinates. This alignment is used as a starting point for search when performing regression in the probable place of a person's face on an image, as shown on Fig. 5.



Fig. 5. Average model

The training data was prepared with the following structure: true form, guessed form and image. Then on this data the algorithm will learn to transform the form-guess into the true form. To generate a large amount of input data random transformation of forms was used (shift, rotation, zoom). Samples with "negative" images were also prepared. On them the guessed form was not near the person's face or it was not possible to convert guessed form into true alignment.

Then it is necessary to select the macro-parameters of training: the number of regressors, the number of trees in the forest, the number of points for the formation of the descriptor, etc.

The regressors are trained consistently. The training of the regressor takes place as a sequential generation of fixed height trees. The generation of a tree is a recursive search for a weak classifier (rule) that divides the current sample into two children. The regression and classification tasks are to be solved simultaneously, but these tasks are independent. That is why each of the trees with some probability will optimize the classification error or, otherwise, a regression error. Leaves of the tree will receive appropriate transformations and values for classification in terms of optimality. That is, the leaves will minimize the corresponding function of the error. For the regression error is the distance between true alignment and guess. For classification it is the error used in the algorithm of adaptive boosting. For each tree threshold value of the accumulated sum of the classification values is calculated. This value allows to delete negative samples. This threshold is calculated from the logical proposition that almost all positive samples should remain in the training. After the training of the regressor is complete, it must be applied to all training data and continue the training of the next regressor in the cascade.

The process of training of the system takes about 35 to 40 hours on 4-core 3,1 GHz intel core i5 CPU.

V. IMPLEMENTATION OF FACE TRACKING SYSTEM

To implement the system it was decided to use C++ programming language. It is a powerful programming

language that allows to quickly and conveniently process data. And the possibility of rapid processing of numerical data is an integral part of the system to meet the speed requirements. The face tracking method requires the ability to work with images such as downloading, cropping, zooming, colour changing, etc. Thus the decision was made to use an open library of functions and algorithms for computer vision, image processing and general purpose numerical algorithms OpenCV.

The developed system does not require any custom visual interface, since it is only the implementation of face tracking method that solves individual computer vision tasks and can be used in other software systems. Therefore, the regressor cascade training module is executed as a console application, and the demonstration of facial detection and tracking capabilities is accomplished with the help of built-in OpenCV functions to get an image from a webcam or file and display the image on the screen.

A. Preparation of data for the regressor training

The prepared training data consists of true form, guess form of a face. A set of random feature points values is generated to form the initial descriptor. Next, for each pair of training data, the descriptors of the guess form and the corresponding image are calculated. After that, the true form of each pair is translated into the coordinate system of the guess-form. In this local coordinate system the vector-shifts are calculated, which must be added to the local points of the guess-form to get the true form. Since the model used for system implementation contains 68 points, one transformation is encoded by 136 real numbers. It is obvious that the number of degrees of freedom of the normal form is much smaller than this number, since the relative location of many points of the model is constant. Therefore, to accelerate the system, compression of the transformation coordinate system up to 40 numbers is performed. This is done with the singular value decomposition algorithm, or SVD. After that, all transformations are encoded by 40 numbers. Next comes the phase of direct training of the regressor, which purpose is to calculate the desired transformation according to the descriptors.

B. The process of tree and rules generation

Each ensemble tree is generated sequentially. After generation, each of tree leaves has a vector transformation that is subtracted from the residual transformation of all the samples walked in this leaf, and also has a classification value that is added to all the samples. Before generating, it is randomly determined what the tree will optimize: classification or regression error. When generating a tree its nodes are recursively generated. On each iteration, a large number of random potential rules are examined to choose such one that minimizes the value of the error function. If regression is optimized, then the rule is chosen to minimize the sum of the squares of the lengths of the vectors of the residual transformations. The value of the transformation in the leaf is calculated as the arithmetic mean of the residual transformations of all the samples that have reached it. If the classification is optimized, then a rule is chosen that separates the set of positive and negative samples in a best way. The value of the classification in the leaf is calculated as

$$(\log(\text{posWeightSum}) - \log(\text{negWeightSum})) * 0.5,$$

where posWeightSum is the sum of the weights of the positive samples appearing in this leaf, and negWeightSum is the sum of the weights of negative samples. Weight of each sample is recalculated after generation of each tree. After that the weights are normalized so that the sum of their values equals to 1. Also for each tree the threshold is determined to reject the sample as exactly negative. This value is selected so that the trained classifier has high sensitivity.

C. System training configuration

The structure of system training has customizable macro-parameters. Most of the parameters are selected intuitively or manually optimized due to the lack of actual optimization. The most important parameters are the following:

- cascade size;
- the depth of one tree (the number of rules, that is, the tree of depth 4 has 16 leaves);
- the number of rules from which to select the best one when generating one internal node;
- number of feature points generated for a given regressor;
- the size of a given regressor ensemble;
- the coefficient with which the transformation from a leaf of a tree is taken. Another popular name is learning rate;
- the probability that the tree will optimize the classification, not regression;
- the dimension of space to which the initial transformations is reduced.

D. Functional capabilities of the system

The developed system solves problems of detection, tracking and automatic face alignment. Accordingly, the developed system and the trained cascade of regressors provide the following functionalities: the search for one or more faces in the image (Fig. 6), calculation of the coordinates of 68 key points (Fig. 7), face tracking in sequential frames (Fig. 8). The system performs search and alignment in real-time with a speed of at least 20 frames per second. Recognition speed for one frame is 30 to 35 milliseconds average and not more than 50 milliseconds in general.

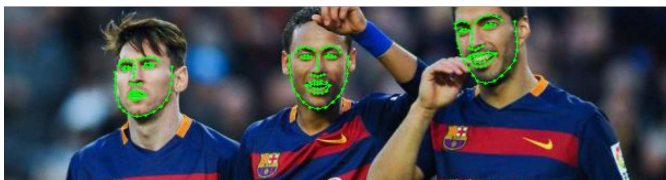


Fig. 6. The result of a face search on the image

E. Comparison of the functional capabilities of the system with known analogues

Among the existing systems analogues for identification and tracking of people faces in real time, one can mention

the following systems: OpenCV, Google Vision API, Dlib Library, Flandmark, Face++. Their disadvantages include a lower rate of accuracy of persons identification, licenses forbidding commercial use, low productivity, little informative results of work. All these disadvantages impose a number of limitations on the possibility of using these systems. Comparison of the functional capabilities of existing similar systems to detect and track the faces of people and the developed system is presented in Table I.

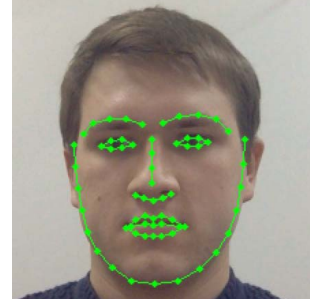


Fig. 7. Alignment of 68 landmarks



Fig. 8. Face tracking in sequential frames

TABLE I. COMPARATIVE ANALYSIS OF SYSTEMS FOR DETECTING AND TRACKING PEOPLE'S FACES

Parameters	OpenCV	Google Vision API	Dlib Library	Flandmark	Face++	Developed system
Face Search	+	+	+	+	+	+
Landmarks alignment	-	+	+	±	± (small number)	+
Face tracking	± (no landmarks)	-	+	±	+	+
Work in real time	+	±	-	±	-	+
Portability	+	± (web API)	+	+	±	+

Based on the analysis, it can be concluded that the developed system, in contrast to existing analogues, allows to search faces simultaneously with their alignment, as well as to perform facial tracking in a real time.

VI. ESTIMATION OF THE ACCURACY OF RECOGNITION

An automated testing was performed to assess the recognition rate and the performance of the developed search algorithms and facial tracking. The open source database of annotated images was used as a test sample FDDB: Face Detection Data Set and Benchmark [13].

Testing on this database, containing more than 4000 images, gave the following indicators of performance. Sensitivity of face detection is 78%, that is, the probability that the algorithm will find a face on an image where it is. Accuracy of face alignment is 95%, that is, the probability that the found form will actually be located on the image of the face. Such numbers are good indicators, since this database contains many different faces in poorly functioning algorithm conditions. At the same time, these indicators are not perfect. Therefore, the algorithm and the process of its learning should be improved.

VII. CONCLUSIONS

In this work, the method of face detection, alignment and tracking is developed and implemented. Although the developed system shows good results and solves the set tasks, analysis and testing allowed to reveal a number of shortcomings and problematic conditions of the work of its implementation. These flaws can be eliminated by improving the method and the learning process directly. We identified a number of extreme conditions in which the system may not work, or show poor results: poor lighting (darkened or illuminated areas), large angles of rotation or inclination of the face; partial covering of the face with objects.

Among the main ideas for improving the system's work, you can highlight the following:

- expansion of the training base. This will improve the generalization of the trained cascade, improve the quality of work in extreme conditions, and reduce the effect of overfitting;
- optimization of macro parameters of the system;
- use of crossvalidation and re-generation of data in the training of each regressor;
- increasing the training time to find the best regression parameters;

- improving the algorithm of the scanning window by dividing it into several phases;
- performing the preliminary classification of the angle for revealing faces with a large angle of rotation;
- acceleration of the system by using vector processor operations.

REFERENCES

- [1] B. Ginneken, "Active Shape Model Segmentation with Optimal Features", IEEE Transactions on Medical Imaging, vol.21, No.8, pp. 924-933, 2002.
- [2] T. F. Cootes, C. J. Taylor, D. H. Cooper, and J. Graham, "Active shape models – their training and application", Computer Vision and Image Understanding, vol. 61, pp. 38–59, 1995.
- [3] T. Cootes, G. Edwards, and C. Taylor, "Active appearance models", IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 23, No.6., pp. 681–685, 2001.
- [4] T. F. Cootes, G. J. Edwards, and C. J. Taylor, "Active appearance models", European Conference on Computer Vision, vol. 2, pp. 484–498, 1998.
- [5] S. Ren, X. Cao, Y. Wei, J. Sun, and S. Ren, "Face Alignment at 3000 FPS via Regressing Local Binary Features", Computer Vision and Pattern Recognition, pp. 1685–1692, 2014.
- [6] L. Breiman, "Random forests", Machine Learning, vol. 45, No. 1, pp. 5–32, October 2001.
- [7] Y. Freund, R. Schapire, and N. Abe, "A short introduction to boosting", Journal of Japanese Society for Artificial Intelligence", vol. 14, No. 5, pp. 771-780, September 1999.
- [8] M. Valstar, B. Martinez, and X. Binefa, "Facial Point Detection using Boosted Regression and Graph Models", Computer Vision and Pattern Recognition, pp. 2729-2736, 2010.
- [9] S. Zhu, C. Li, C. Change Loy, and X. Tang, "Face Alignment by Coarse-to-Fine Shape Searching", Computer Vision and Pattern Recognition, pp. 4998-5006, 2015.
- [10] J. Li and Y. Zhang, "Learning surf cascade for fast and accurate object detection", Computer Vision and Pattern Recognition, pp. 3468–3475, June 2013.
- [11] R. Schapire and Y. Singer, "Improved Boosting Algorithm Using Confidence-rated Predictions," Machine Learning, vol. 37, No. 3, pp. 297-336, 1999.
- [12] B. M. Smith and L. Zhang, "Joint face alignment with nonparametric shape models", 12th European Conference on Computer Vision, 14 p., 2012.
- [13] FDDB: Face Detection Data Set and Benchmark. URL: <http://vis-www.cs.umass.edu/fddb>. Last checked on 10th March, 2018.

Investigation the Ateb-Gabor Filter in Biometric Security Systems

Mariya Nazarkevych
Lviv Polytechnic National University,
Lviv, UKRAINE,
mariia.a.nazarkevych@lpnu.ua

Ivanna Klyujnyk
Technical College,
Lviv Polytechnic National University
Lviv, UKRAINE,
ivanna.klyujnyk@gmail.com

Hanna Nazarkevych
Faculty of Cybernetics Taras
Shevchenko National University of Kyiv
Kyiv, UKRAINE,
h.nazarkevych@gmail.com

Abstract—Biometric information security systems are analyzed. One of the most important characteristics is the high reliability, which corresponds to the ability of the system to distinguish between biometric characteristics that correspond to different people on also it is safe to find a match. To implement this, at the filtration stage, it is proposed to use the Ateb-Gabor filter, which extends the usual filtering. With this filtering, you can improve the gradation characteristics of biometric models. The Ateb-Gabor filter was constructed and studied, graphs of the new function and corresponding frequency graphs are presented. Filtration experiments are performed. The SNR filtration quality is estimated.

Keywords—Image filtering, Ateb-Gabor filter, Ateb-functions.

I. INTRODUCTION

Biometric technologies have become the most significant of the latest achievements in the area of identification and control of the information access. Biometric technologies are based on the biometric characteristics of a single person [1]. These include the unique characteristics that a person receives at birth: DNA structure, eye iris pattern, eye retina, geometry and temperature card of the face, fingerprints, palm geometry. Biometric characteristics include those that are acquired and change over time - signature, voice, walking [2].

Biometric systems are divided by the following biometric indicators [3] - 58% of the are fingerprints, 18% - facial geometry, 7% - retinal eye, 7% - hand geometry, 3% - human vein pattern, 5% - human voice, 2% - other biometric Indexes..

All biometric systems consist of two parts - hardware and specialized software.

There are four stages of identification in any biometric system[4]:Зберігання – фізичний взірець зберігається системою

Selection - the unique information is extracted from the model and the biometric model is compared to it.

The model is compared with available models. Match/mismatch - the system decides whether biometric models match.

The purpose of this work is to review modern biometric methods of identification of a person and the level of their development, as well as the possibilities of using modern technologies in information security systems and the development of a new filtration method.

One of the most important information security characteristics based on the biometric technologies is high reliability. It can be defined as the ability of the system to reliably distinguish between biometric characteristics belonging to different people and to reliably find coincidences [5].

Dactyloscopic recognition method takes half of the market access systems [6]. Nowadays such systems include laptops, keyboards, mice, flash drives, door locks, etc.

Biometric security systems are frequently used in the modern society. Identification of the following characteristics are introduced:

- the probability of a False Acceptance Rate (FAR) [3] is the most undesirable result that needs to be minimized.

- the probability of a False Rejection Rate (FRR) refusal [3], this false result can be corrected.

These characteristics are related. The smaller the first, the bigger the second. The point in which these two errors are equal is called EER (Equal Error Rates) [3]. The lower the EER value, the higher the error rate of the access system. The first number characterizes the probability of denial of access to the person having access, the second is the probability of false coincidence of the biometric characteristics of two people. Forging a papillary pattern of a person's finger or a rainbow of an eye is very difficult. So the emergence of "second kind errors" (that is, granting access to a person who does not have this right) is practically excluded. However, under the influence of some factors, the biological characteristics that make identification of the personality can change. Therefore, the frequency of occurrence of "errors of the first kind" (denial of access to a person having this right) in biometric systems is large enough. The better the system is, the lower the value of FRR with the same FAR values is. EER (Equal Error Rate), which defines the point in which the FRR and FAR graphs cross, are sometimes used. But it is not always representative.

II. ANALYSIS OF RECENT DEVELOPMENTS

Using biometric systems, especially facial recognition systems, even after introducing correct biometric characteristics, the authentication decision is not always correct. This is due to a number of features and, first of all, with the fact that many biometric characteristics can change. There is a certain degree of system error probability. And with the use of different technologies, the error can vary significantly. For access control systems using biometric technologies, it is necessary to determine what it is more important not to miss a "stranger" or skip all "own" examples [3].

Many firms and state institutions are currently developing biometric security systems. Among them, the most well-known are Ekey biometric systems and ZKTeco, which introduced biometric systems for implementation [7].

Ekey biometric systems is an Austrian company, a leader in the development and implementation of biometric systems in Europe. ZKTeco - The Chinese company produces budget devices for access control while taking into account working time, which scans fingerprints and facial geometry [8]. Such equipment is in demand in various financial and public organizations.

The developed Ateb-Gabor filter has several significant advantages compared with the known ones. Among the parameters that you can modify the image, there are two independent ones that accept the values of rational numbers. Additionally, filtering can be done not for the whole image, but for some kind of slabs that look noisy. The filtering developed by the filter showed good results. The Gabor function [9] is a Gauss modulated. Gaussian function with four parameters: displacement t_0 , standard mean square deviation σ , modulation frequency Ω , and phase shift θ :

$$G(t) = e^{i\Omega(t-t_0)-i\theta} e^{-\frac{(t-t_0)^2}{2\sigma^2}}$$

Decomposing on the Gabor functions is a decomposing on the modulated fragments of the sinusoid. The length of the fragments for all frequencies is constant, which gives a different number of oscillations for different harmonics. It follows that a sufficiently well localized in the t and k -space function of the Gabor cannot be a basis of the wavelet transform, since the basis based on it does not have the properties of self-similarity [9].

We will solve the system of differential equation:

$$\begin{aligned} \dot{x} + \beta y^m &= 0, \\ \dot{y} + \alpha x^n &= 0. \end{aligned} \quad (1)$$

where α, β – some real constants and where m, n is:

$$\begin{aligned} n &= (2\theta'_1 + 1) / (2\theta''_1 + 1) \\ m &= (2\theta'_2 + 1) / (2\theta''_2 + 1), (\theta'_1, \theta''_1, \theta'_2, \theta''_2 = 0, 1, 2, \dots). \end{aligned} \quad (2)$$

Let's solve this system graphically and solution for parameters $m=1, n=1, \alpha=1, \beta=1$ are shown on Fig. 1, a $m=7, n=7, \alpha=1, \beta=1$ are shown on Fig. 2 Solutions of this systems are Ateb-functions [10].

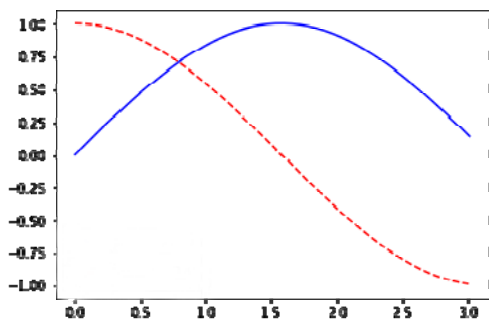


Fig. 1. Ateb-ca (red) and Ateb-sa (blue) with parametres $m=1, n=1, \alpha=1, \beta=1$.

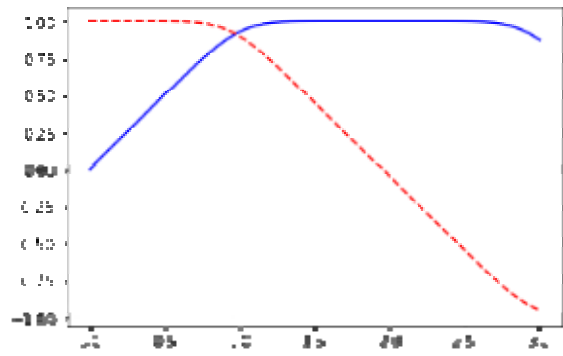


Fig. 2. Ateb-ca (red) and Ateb-sa (blue) with parametres $m=7, n=7, \alpha=1, \beta=1$.

III. TWO-DIMENSIONAL ATEB-GABOR FILTER

Filtration of the two-dimensional Ateb-Gabor is implemented by formula:

$$\begin{aligned} \text{Ateb-G}(x, y, \lambda, \theta, \psi, \sigma, \zeta) &= \\ \exp(-(x^2 + \psi y^2) / 2\sigma^2) \text{Ateb-ca}(2\pi x / \lambda + \zeta). \end{aligned} \quad (3)$$

$$\begin{aligned} x' &= x \cdot \cos(\theta) + y \cdot \sin(\theta) \\ y' &= -x \cdot \sin(\theta) + y \cdot \cos(\theta), \end{aligned}$$

where λ the wavelength of the cosine - multiplier, θ - parallel bandwidth normal orientation, ζ - lagging (phase transmittion; phase shift), ψ - data compression ratio.

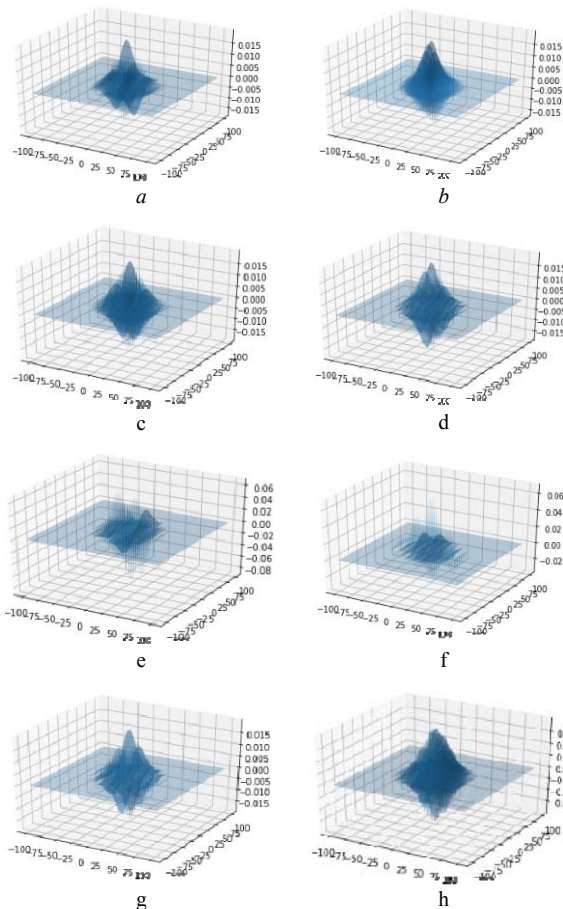


Fig. 3. Graphic representation of two-dimensional Ateb-Gabor for a) $m=0.1, n=1$; b) $m=0.5, n=1$; c) $m=1, n=1$; d) $m=3, n=1$; e) $m=2, n=1$; f) $m=4, n=1$; g) $m=5, n=1$; h) $m=1, n=5$.

Fig. 3 shows graphs of the Ateb-Gabor two-dimensional filter for various Ateb-function parameters. As you can see from the drawings, you can pick up the values that would most closely match the figures with biometric data. Experiments (except Fig. 3h) are performed with the parameter $n = 1$. The parameter m changes its value to the side of the increase. In graph c, we can observe the parameters $m = n = 1$, which corresponds to the well-known Gabor filter.

Fig. 4 shows the frequency graphs corresponding to the charts of the two-dimensional Ateb-Gabor. The experiments were carried out with the same parameters of the Ateb function as shown in Fig. 3. On the graphs depicted in Fig. 3 it can be observed that the maximum of the Ateb-Gabor function in Fig. 3 are displayed on the frequency chart with white curves. When filtered, this will allow you to draw out the contours of the figure with a biometric. On graph 4c we can observe the parameters $m = n = 1$, which corresponds to the frequency pattern of the Gabor filter.

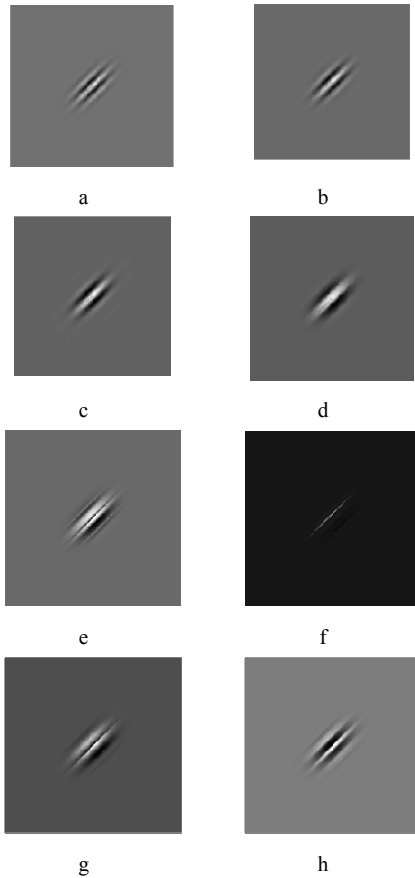


Fig. 4. Graphic representation of frequency two-dimensional Ateb-Gabor for a) $m=0.1, n=1$; b) $m=0.5, n=1$; c) $m=1, n=1$; d) $m=3, n=1$; e) $m=2, n=1$; f) $m=4, n=1$; g) $m=5, n=1$; h) $m=1, n=5$.

IV. EXPERIMENTAL DATA

Fig. 5 shows the representation of the Ateb-Gabor filtration of the fingerprints $Ateb-G(x,y,\lambda,\theta,\psi,\sigma,\zeta)$ at different values of the sinusoidal wave frequency θ , see (3). The experiment was carried out with values $m = n = 1$. The value of σ of the mean-deflection of the Gabor rotation is selected as 1, in order to see the visual differences in the filtration. As it can be seen from Fig. 5e and 5f look more contrasting than

other images. In Fig. 5f and 5g on the papilla of the fingers small details are visible more clearly.

We are evaluating the effectiveness of using different filtration and filtration with different characteristics, comparing characteristics that describe the invisibility (distortion level) and bandwidth. The study of image quality based on enhanced image resolution is devoted [11]. Method image superresolution from two frames is based on the aggregate divergence matrix elements of the theory and genetic algorithms [12].

Evaluation of the image distortion level is based on PSNR [13]:

$$P_{SNR} \quad (4)$$

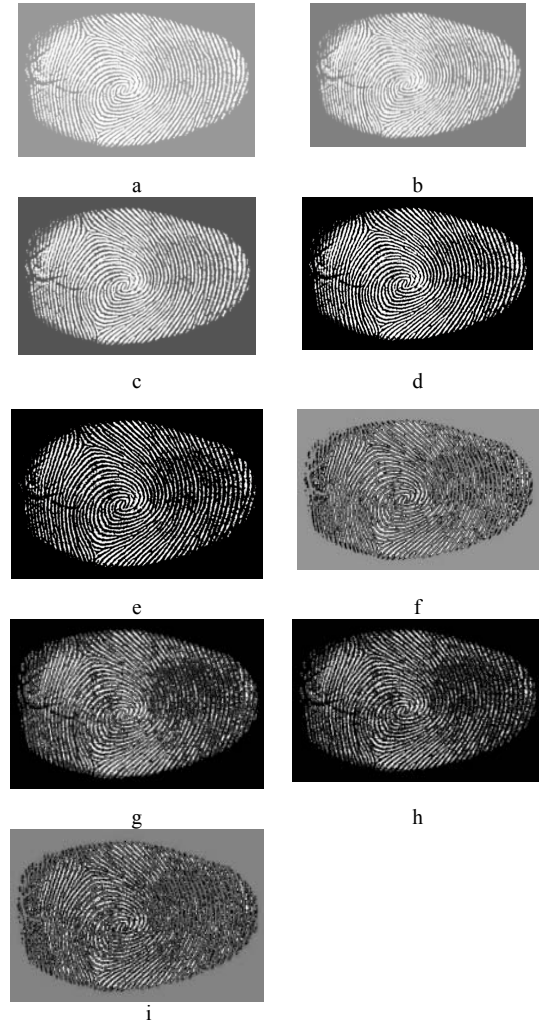


Fig. 5. Results of the image processing by the Ateb-Gabor filter at different frequency values of the sinusoidal wave θ : a) $0.1/p$, b) $0.3/p$, c) $0.5/p$, d) $1/p$, e) $2/p$, f) $3/p$, g) $4/p$, h) $5/p$, i) $6/p$.

During the research, images of the jpg format, which was undelivered and with a size of 466x311 pixels, were used. The study was conducted in the following way. Two images were compared. The first image is called the original, and the second image is the filtered image. In order to make the equal-sized images which vary slightly among themselves, both of them are filtered out. However, the first image was changed based on the habitual Gabor filter, and the second one was replaced with the Ateb-Gabor filter. Then, on the basis of (4), these two images were compared. The

comparison results are presented in Table I. From the comparison results, it can be noted that the larger the parameter m, n , the more the filtered image differs from the same picture filtered with Gabor filter.

TABLE I. EVALUATION OF BIOMETRIC IMAGE DISTORTION WITH ATEB-GABOR FILTERING

Original image	Filtered image	SNR
m1 n1	m3 n3	20,63 dB
m1 n1	m7 n7	19,04 dB
m1 n1	m11 n11	16,01 dB

Models of systems and pattern recognition using the SNR signal-to-noise ratio are described in [14]. Mathematical approaches to the recognition of biometric images are taken from [15]. The graphical representation of the change in the gradation characteristics of biometric images during filtration by the Ateb-Gabor is shown in Fig. 6. As can be seen from the graph, with the increase of the parameters of the Ateb-Gabor, the image substantially changes, this corresponds to a curve which value decreases with increasing x coordinate.

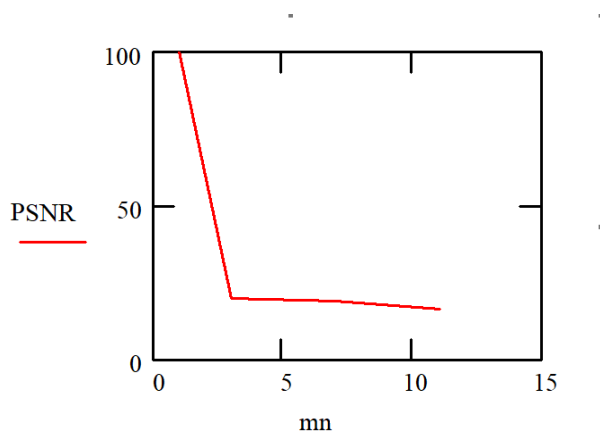


Fig. 6. Estimation of changes in the characteristics of the PSNR image from parameters m, n Ateb-Gabor

V. CONCLUSIONS

The properties of the Ateb-Gabor filter are studied. Usually the images which come to the system of processing have low quality due to noise action. Filtration of Ateb-Gabor is to reduce effects of noise and obstacles, This is actually expand methods of filtration.

The Ateb-Gabor filter has been developed, it extends the functions of controlling the conventional Gabor filter through. This allows developing a new image processing method.

The change of the parameters m and n provides different values of the period, which gives a possibility to expand the number of filter options.

Ateb-Gabor function gives a possibility to solve the problem with identification of finger's papilloma by improving the identification process. And on its basis, it is possible to conduct the image filtration with a big amount of combs. All this guarantees better characteristics than usual one-dimensional Gabor filter.

The efficiency-proving experiments of SNR-based filtration based on signal-to-noise ratio are carried out. It is shown that when entering the parameters m and n of the Ateb-garbage, it is possible to modify and extend the results of the filtration.

REFERENCES

- [1] S. Pankanti, R.M. Bolle, and A. Jain. "Biometrics: The future of identification [guest editors' introduction]," Computer society, 33.2 pp. 46-49, 2000.
- [2] J. Wayman, et al. "An introduction to biometric authentication systems," Biometric Systems. Springer London., Pp.1-20, 2005.
- [3] A. Ross, and A. Jain. "Information fusion in biometrics." Pattern recognition letters 24.13 (2003): 2115-2125.
- [4] U. Umut and A.K. Jain. "Attacks on biometric systems: a case study in fingerprints," Security, Steganography, and Watermarking of Multimedia Contents VI, vol. 5306. International Society for Optics and Photonics, 2004.
- [5] X. Liu, M. Tanaka, and M. Okutomi. "Noise level estimation using weak textured patches of a single noisy image," Image Processing (ICIP), 2012 19th IEEE International Conference on. IEEE, 2012.
- [6] W.F. Lane, "Self-authenticating identification card with fingerprint identification," U.S. Patent No. 5,623,552. 22 Apr. 1997.
- [7] J.E. Thurman, "Biometric security now and in the future," East Carolina University, 2016.
- [8] T. Murakami and K. Takahashi. "High security biometric authentication system," U.S. Patent Application No. 13/706,854.
- [9] J.P. Jones, and L.A. Palmer. "An evaluation of the two-dimensional Gabor filter model of simple receptive fields in cat striate cortex." Journal of neurophysiology, 58.6, pp. 1233-1258, 1987.
- [10] M. Nazarkevych, R. Oliamiyk, H. Nazarkevych, O., Kramarenko, and I. Onyshchenko, "The method of encryption based on Ateb-functions," IEEE First International Conference Data Stream Mining & Processing (DSMP), pp. 129-133, August 2016.
- [11] Y. Rashkevych, D. Peleshko, O. Vynokurova, I. Izonin and N. Lotoshynska, "Single-frame image super-resolution based on singular square matrix operator," 2017 IEEE First Ukraine Conference on Electrical and Computer Engineering (UKRCON), Kiev, pp. 944-948, 2017. doi: 10.1109/UKRCON.2017.8100390.
- [12] D. Peleshko, T. Rak, M. Peleshko, I. Izonin and D. Batyuk, "Two-frames image superresolution based on the aggregate divergence matrix," 2016 IEEE First International Conference on Data Stream Mining & Processing (DSMP), Lviv, Ukraine, pp. 235-238, 2016. doi: 10.1109/DSMP.2016.7583548
- [13] O. K. Yudin, O. G. Korchenko, G. F. and Konahovich, Information security in data networks. K: TOV "NVP" INTERSERVIS, 2009.
- [14] O. Riznik, I. Yurchak, E. Vdovenko and A. Korchagina, "Model of stegosystem images on the basis of pseudonoise codes," 2010 Proceedings of Vth International Conference on Perspective Technologies and Methods in MEMS Design, Lviv, pp. 51-52, 2010.
- [15] O. Riznyk, I. Yurchak and O. Povshuk, "Synthesis of optimal recovery systems in distributed computing using ideal ring bundles," XII International Conference on Perspective Technologies and Methods in MEMSDesign (MEMSTECH), Lviv, pp. 220-222, 2016. doi: 10.1109/MEMSTECH.2016.7507545.

Applying the Neuronetchic Methodology to Text Images for their Recognition

Bohdan Durnyak
Department of Automation and Computer technology
Ukrainian Graphic Arts Academy
Lviv, Ukraine
durnyak@uad.lviv.ua

Oleksandr Tymchenko
Department of Safety Engineering,
University of Warmia and Mazury
Olsztyn, Poland
o_tymch@ukr.net

Oleksandr Tymchenko Jr.
Department of Computer science and Information technology
Ukrainian Academy of Printing
Lviv, Ukraine
bratokokok@gmail.com

Bohdana Havrysh
Department of Preprinting Technology
Ukrainian Academy of Printing
Lviv, Ukraine
dana.havrysh@gmail.com

Abstract — There is considered the methodology for recognizing text images based on neural networks, methods and algorithms for building a neuro-fuzzy system for recognizing text images, in particular methods for improving the quality of text images and reducing noise through linear and nonlinear filtration. Features of binarization of such images, fuzzy processing of images to allocate boundaries and segmentation of symbols, and the ability to implement grammar for the structural recognition of text images is shown. The simulation of the developed system is also carried out.

Keywords — OCR system, fuzzy image processing, neural networks.

I. INTRODUCTION

With the development of information technology it has become possible to facilitate, accelerate and improve the quality of recognition of printed or handwritten texts. The first element of the letter recognition system is a scanner or a digital camera that inserts text images into the computer. To create a text document, you need to recognize individual characters in this image. There is a range of software that have virtually automated the process of recognizing texts. However, it is not always possible to ensure a satisfactory result in the case of distortions of printer or handwritten text images of various types (geometric, noise, etc.) [1]. The problem of effective text recognition plays an important role in the areas of informatization of various processes of human activity. The textual presentation of information, in comparison with graphic, allows significantly reduce the costs of storing and transmitting information, and also allows us to implement all methods of using and analyzing electronic documents. Therefore, the greatest interest from a practical point of view is precisely the transformation of information from paper carriers into a text electronic document.

II. GENERAL CHARACTERISTICS OF TEXT RECOGNITION SYSTEMS

The OCR system (for example, FineReader, OmniPage, ReadIris, etc.) receives a digital image of a scanned or photographed document and forms the text containing this

image in one of the formats of electronic text documents (Fig. 1) [2].

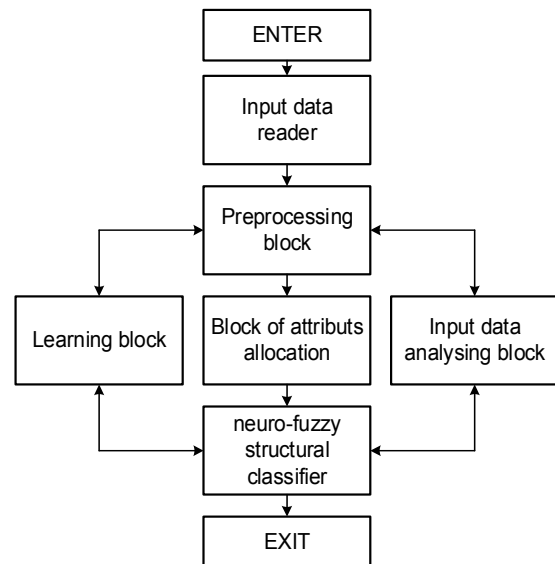


Fig. 1. The generalized structure of the neuro-fuzzy OCR system

Block of attributes allocation has a different complexity depending on the nature of the image being recognized, as well as the methods that are used. Each pre-processed object must be represented in the form of a language-type structure, such as a chain. This process consists of two subprocesses: segmentation and allocation of non-intrusive elements. The main purpose of the segmentation process is to select individual elements from the image to highlight the features or compare the differences with the standard samples in memory. That allow the recognition to obtain a text that is almost identical to the original.

The classification block generates signs of matching elements of the image with reference samples. The classification efficiency is evaluated by the number of features that will be used for this compliance [3].

Preliminary processing of digital images is considered as the result of eliminating the various types of interference and the effects of distortions on which this image was formed. Since methods for improving the quality of images

in the frequency domain require a high computational complexity, it is advisable to use spatial in this case.

Most of these operations of OCR systems due to the fuzziness and blurriness of the parameters it is expedient to execute methods of fuzzy logic.

Thus, the purpose of this work is to develop methods and algorithms for constructing a neuro-fuzzy recognition system for text images that contain [4]:

1. reduce noise of text images;
2. binarization of images;
3. allocation of boundaries and segmentation of characters;
4. realization of grammar for the structural recognition of text images.
5. Verification of proposed methods.

III. DECREASE OF DIGITAL IMAGES NOISE

To reduce noise, there are used linear and nonlinear filters that store sharp changes and edges of objects while eliminating noise. Most often noise is considered as impulse noise, Gaussian and mixed pulse and Gaussian noises [5].

In the case of mixed noise, linear and nonlinear filters can be used sequentially.

Gauss filter (linear) averages pixels around the point of the image according to Gauss's law $(x,y) \in Z^2$:

$G(x,y) = \frac{1}{2\pi\sigma^2} \exp\left(-\frac{x^2 + y^2}{2\sigma^2}\right)$. This filter is separable, filtering can be performed in rows and columns of the

image matrix, since it decomposes into two independent filters at different coordinates.

Median filter (nonlinear) best deals with impulse noise. The work of the median filter consists in choosing the median from a set of pixels around: $Im_{i,j} = \text{med}[Im_{i+s,j+t}, (s,t) \in W]; (i,j) \in Z^2$. All pixel values around (s, t) relative to the pixel of the image $(i,j) \in Z^2$ are sorted in order of magnitude, then the median value is selected, which replaces the central pixel around.

Note that the processing of images due to fuzziness and blurring of data is today one of the key in the theory and practice of developing information systems. One of the simplest algorithmic methods of image processing is "soft computing" with the use of fuzzy logic. This will improve its visual quality by improving the detail differentiation and increasing detail in general for further classification and image recognition.

Fuzzy image processing is a set of various fuzzy approaches, which are understanding, representation, processing of images, their segmentation and classification. In the process of recognition, the process of pre-fuzzy image processing is extremely important, because a quality of the data that arrives at the inputs of the neural network depends on it. The algorithm of the previous fuzzy processing can be represented in the form of a sequence of steps (Fig. 2): image acquisition; converting the resulting color image into a grayscale image; fuzzy image processing (fuzzyfication, fuzzy output system, defuzzyfication) [6].

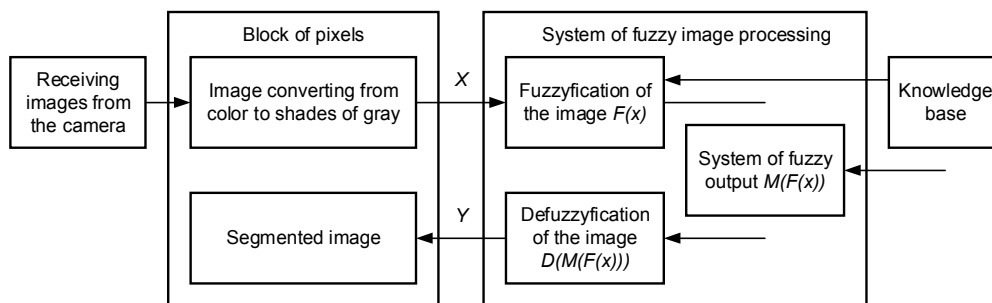


Fig. 2. Algorithm of the previous fuzzy processing of images for their segmentation

A. Binarization of the image

In image binarization the pixel value is conventionally considered equal to zero or one: if its value is above a certain threshold, it corresponds to a white color if it is below the set threshold, that is black. If $P(x, y) > PT(x, y)$ then the pixel in the binary image is white, otherwise black.

The threshold surface of the PT is a matrix whose dimension corresponds to the dimension of the original image [7]. All methods of binarization are divided into two groups based on the principle of constructing a threshold surface – these are methods of global and local processing of binarization.

In the methods of global processing, the threshold surface is a plane with a constant brightness value, it means that the threshold value is the same for all pixels of the original image. Global threshold processing has a significant disadvantage – if the original image has non-uniform lighting, the areas that are illuminated worse are classified as black.

In local methods, the threshold value varies for each point based on the features of the domain belonging to the vicinity of a given point.

In the developed algorithm, the original half-tone image, given in digital form, is divided into square matrix S

with size $h \times h$ and elements $S(x,y) \in [0,1]$. Each matrix S is transformed into a binary matrix r of the same size. The transformation $S \rightarrow r$ is carried out under the condition that the matrix is equal to the brightness:

$$\text{Ent} \sum_{x,y} S(x,y) = \sum_{x,y} r(x,y) = b, \quad (1)$$

where brightness is the sum of the elements of each matrix. The brightness of the binary matrix is equal to the number of units b . During each $k = \overline{1, b}$ appealing to the elements of the matrix S in it, the position of the element with the maximum value is determined. In place of this position in S zero is written, and in the binary matrix r a unit is written [8].

The criterion for constructing a binary matrix is the minimum of the Euclidean distance between the binary and the halftone matrices.

Then the task is to choose from $2n$ matrices. The algorithm searches for the maximal element of the matrix

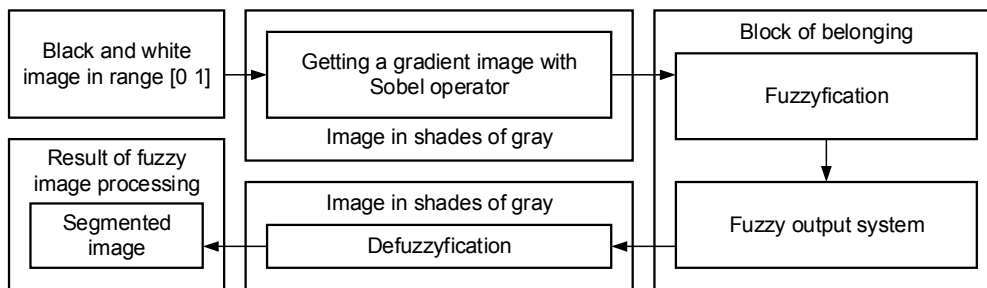


Fig. 3. Algorithm of fuzzy processing of images to allocate boundaries

An image of X -size $M \times N$ with L gray levels $g = 0, \dots, L-1$ can be defined as a fuzzy single-point set that specifies the value of each pixel attribute relative to the image property (for example, brightness, smoothness, etc.).

$$X = \bigcup_{m=1}^M \bigcup_{n=1}^N \frac{\mu_{mn}}{g_{mn}}, \quad \mu_{mn} \in [0,1] \quad (2)$$

where μ_{mn} and g_{mn} – assignments of the mn pixel in the fuzzy set. Determining the values of affiliation depends on the specific application requirements and knowledge base.

Since the symmetric parts of the bonding site of the membership function are not always effective, attention in using fuzzy logic to improve the quality of the half-tone images leads to the need to construct S-like fuzzy functions. That have a changed center of symmetry [9]: the function of belonging is described as two glued parabolic links with a continuous derivative in the place of gluing, that is, functions of the second order.

The specific choice of methods of defuzzifications is carried out depending on the desired behavior of the fuzzy output. It is advisable to use the center weighting function of belonging fuzzy set.

$$Y = D(M(F(X))). \quad (3)$$

S , so the number of steps in the order of bn is required. Unlike pseudo-tone [8] methods, the method of locating zeros and units in the investigated algorithm allows you to convert a semitone unit based on its brightness histogram. Visually this is manifested in a more accurate reproduction of the details and boundaries of objects in a binary image.

B. Algorithm of fuzzy processing for boundaries and segmentation of images allocation

After binarization, the image enters the system of fuzzy image processing (Fig. 3).

Fuzzy image processing consists of three main phases: F-image, fuzzy-output system (M) on the values of affiliation, and defuzzifications of the (D) images. After transferring the image from the gray-level image to the phase-out, the fuzzy output system is determined by the value of the accessory function.

C. Construction of grammar for the structural recognition of text images

To justify the method of constructing the grammar of nonstructural image recognition and character classification, consider the following recognition methods.

In the comparison method, the selected element is compared with the database, where each object is represented by different angles, scales, displacements, deformations. For letters in the database, you should also specify the font, font properties, etc.

The second approach is an analysis of the characteristics of the image. In the case of optical text recognition, this may be the definition of the geometric parameters of individual characters [10].

Methods, that use artificial neural networks, require a large number of examples in training, but have higher efficiency and productivity. In these methods, the symbol image is reduced to some standard size, for example, 16×16 pixels. The values of brightness in the nodes of the normalized raster are used as input parameters of the neural network. The number of output parameters of the neural network is equal to the number of recognizable characters. The result of the recognition is a symbol that corresponds to the highest value of the source vector of the neural network.

Note that structural description is:

- understandable for a person who solves the problem of object recognition;
- suitable for computer realization;
- free from complexive computing and information loss.

Structural features used in the structural description are non-unique (elementary unique) elements (symbols), primitives of the image. Structural methods store information not about the spontaneous character writing, but about its topology. That is, the standard contains information on the mutual placement of individual components of the symbol [11]. In this case, the size of the distinguished letter and the font that it is printed is not important [12].

The recognizable symbol is given to some sample that has reference dimensions and positions and is smoothed. That means, that an exclusion from the image of all elements that led to its distortion, by replacing the group of image elements (usually adjacent ones) with an element equivalent to them.

The obtained image is subjected to the procedure of skeletonization (reduction) [13]. The contour of the skeletal representation is described in the form of a series of special points of the special points and a circuit code bypassing the circuit clockwise, consisting of an anchor point, a number of codes, and an array of directions from the next point to the next. Special points are end-points and branch points (triods). These are the points whose neighbors form at least three connected areas. By renumbering special points and changing the path start, there is made the contour identification with one of the main types. Operating a limited number of atomic (non-derivative) elements (primitives), you can get a description of various objects.

As a result of connections from non-derivative elements (structural features) there will be formed an object similar to the way the sentences of the language are constructed by combining words that consist of letters. In this structural methods have an analogy with the syntax of natural language.

For recognition we use sentences, each of which describes the structure of an object from non-derivative (elemental) elements. Structural or linguistic classification of an object in such case is performed by comparing the sentence of an unknown object with standard sentence classes.

D. Characteristics of texts` image

The peculiarity of these images [14] is that they consist of a large number of interconnected parts. Therefore, it is expedient to analyze such images using methods of structural recognition. The result of this recognition is not attribution of the symbol image to the prototype, but the list of characters and relations between them. The complexity of recognizing such images is that the symbol is not defined uniquely by its image. The name of the character image depends not only on the image of this fragment, but also on its place, environment.

Structural recognition of symbols in the conditions of random noises is reduced to finding the optimal image of the characters that are recognized. The function of quality is to search for the most probable set of hidden image parameters. For example, in order to recognize the line of the text, the requirement to find the most to find the most consistent number of letters is the same as requiring the minimization of the number of incorrectly recognized characters.

Images that contain texts, tables, drawings, are created and read in accordance with certain rules, which can be formalized as a grammar. Obviously, algorithms for the recognition of such images should be based on the use of the rules of this grammar.

Methods for creating such grammar are as follows:

- usage of graphs theory. The image of the text is presented in the form of a well-defined graph. The tasks of recognition are presented as the problem of finding an isomorphism of the reference and input graphs, or of the isomorphism of their subgraphs.
- methods of the theory of formal languages and grammar. The image is considered as a word in some formal language, which is given using constructs that are generalizations of the Khomsky grammar. Recognition is to find the best in a certain meaning of the output of an image in a given grammar [15].

The grammar considers the image as an object, consisting of certain rules from a large number of elementary parts. These parts and rules can differ significantly from each other (for example, recognition of notes [16] or recognition of mathematical formulas [17]).

Let T be a certain rectangular subset of a two-dimensional integral lattice: $T = \{(i, j) \mid 0 \leq i < H, 0 \leq j < W\}$. The set T is called the field of view, the numbers H and W — its height and width, and the elements of the field of view will be called pixels. Pixel colors are selected from a finite Y set. The reflection $x: T \rightarrow Y$ is called the image, the value $x(t)$ determines the color of the image x in the pixel $t \in T$.

The two-dimensional context-free grammar will be $G = \langle E, V, P, \varepsilon \rangle$, where E is the set of terminal images containing one-pixel images (images determined by the size of 1 pixel) of all colors from the set Y, V .

Y, V is the set of nonterminal names (metacharacters) assigned to parts of the image in the process of generating it using grammar, $\varepsilon \in V$ is an axiom, used to name all, completely generated images, P – set of rules of output. It contains the rules of three types: the rules of horizontal concatenation, vertical concatenation (association of image blocks) and the rules of substitution. Each separate set of these rules will be marked Ph, Pv, Ps accordingly.

The rules for the substitution of Ps have the form $v \rightarrow e$, where $v \in V$ – nonterminal name, and $e \in E$ – terminal image.

The set of rules of horizontal concatenation Ph contains triples of non-terminal names of the form $v \rightarrow v|vr$. That is, any rectangular image fragment can get the name v if it can

be broken up by vertical lines on such two rectangular fragments that the left already has the name vl , and the right one is vr . The symbol $|$ is used to divide a pair of nonterminal names in the rules of horizontal concatenation. Similarly, the set of rules for vertical Pv concatenation contains triples of nonterminal names of the form $v \rightarrow vt/vb$. That is, any rectangular image fragment may get a name v , if it can be split by a horizontal line on such two rectangular fragments that the upper one already has the name vt , and the lower one is vb .

Grammar G is composed of images that can be assigned a name ε . The sequence of rules applied to the image x , which results in the assignment of the name ε to the entire image, is called the output of the image x in the grammar G . The algorithm [8] for solving the problem is a direct generalization of the Cocke-Younger-Kasami algorithm [18] to determine the relevance of the language of a certain context-free grammar of the Khomsky. t consists in the fact that in the sequential review of all rectangular fragments of an image for each of them it is determined what names can be assigned to him in this grammar. At the same time for reviewing fragments are arranged in size.

The introduction of a two-dimensional context-free grammar and the formulation of a task for exact collision obviously have a number of disadvantages that significantly limit their practical application:

- the formalism of two-dimensional context-free grammar does not always allow you to find a real image that can be split into rectangular fragments, and thus the given fragments do not intersect;
- excessive detailing of the rules of grammar: for each image it is necessary to indicate how it consists of smaller parts up to the level of individual pixels. Obviously it should be used larger fragments of the image, for example, fragments that correspond to individual text lists;
- setting the task for exact collision requires an exact match of the color values of individual pixels of the image with the colors of the terminal images. This condition is not fulfilled if there is a noise characteristic of the recognition.

The basic context-free construct introduced in [19] allows us to eliminate these disadvantages, as it:

- describes the image as being composed of fragments of an arbitrary, not just rectangular shape;
- the process of constructing images is set to the level of a set of term fragments, whose sizes can be much larger than the pixel size;
- a fine is imposed for the assignment of names to fragments of an image, which for a given fragment is equal to the sum of the fines of the fragments from which it was formed;
- the fine for the term fragments is determined by an arbitrary function that is not related to the formalism of context-free grammar. This allows us to use heuristic reasoning in its construction. In the task of recognizing texts, it can have any function that defines the similarity of an arbitrary letter and a fragment of an image.

E. The neural network functioning

To construct a non-structural character recognition system, there is used a neural network of reverse error distribution, which consists of several layers of neurons, and each neuron of the previous layer is associated with each neuron of the next layer. In such networks, after determining the number of layers and the number of elements of each layer, it is necessary to calculate the values of the weights and thresholds of the network in such a way to minimize the forecast error. For network teaching is used an algorithm for reverse error spreading. It calculates a vector gradient surface error. Then it moves to a certain value in the direction of the vector (it will indicate the direction of the fastest descent), where the error value will be less. This gradual progression will gradually lead to a minimization of the error. Denote the matrix of weight coefficients from the inputs to the hidden layer W , and the matrix of weights connecting the hidden and output layer — V . The entries will be numbered only by the index i , the elements of the hidden layer — the index j , and the outputs — by the index k . The number of network inputs is n , the number of neurons in the hidden layer is m , the number of neurons in the output layer is p . If the network is studying on a sample (X^t, D^t) , $t = \overline{1, T}$. Then the learning algorithm for multilayer perceptron will look like [20].

Step 1. Initial initialization. For weighted coefficients we give small random values, for example, from the range $(-0.3, 0.3)$; set: ε — learning accuracy parameter, $\cdot \approx 0.1$ — learning speed parameter (may be decreased in the learning process), N_{max} — the maximum number of iterations.

Step 2. Calculation of the current output signal. At the network entrance we submit one of the images of the training sample and determine the values of the outputs of all neurons of the neural network

Step 3. Calculation of the weight change for the source layer of the neural network:

$$V_{jk}^{N+1} = V_{jk}^N - \alpha \frac{\partial E}{\partial V_{jk}}, \quad (4)$$

$$\text{where } \frac{\partial E}{\partial V_{jk}} = \delta_k y_j^c, \quad \delta_k = (y_k - d_k) y_k (1 - y_k)$$

For a hidden layer:

$$W_{ij}^{N+1} = W_{ij}^N - \alpha \frac{\partial E}{\partial W_{ij}}, \quad (5)$$

$$\text{where } \frac{\partial E}{\partial W_{ij}} = \left(\sum_{k=1}^p \delta_k V_{jk}^{N+1} \right) y_j^c (1 - y_j^c) X_i.$$

Step 4. Steps 2-3 repeating for all learning vectors. The training ends after achieving for each of the learning images the value of the error function which does not exceed ε or after the maximum allowable number of iterations.

Below is a review of the work of the developed neuro-fuzzy system (with a 5x5-pixel window) with the commercial product ABBYY FineReader 11 Corporate

Edition (the image of the page text is 702 characters, the Gaussian noise is 0.03 from the black level). For practical implementation, the Matlab Simulink software environment with built-in Fuzzy Logic Toolbox fuzzy logic elements is selected.

TABLE I. RESULTS

Developed system		ABBYY FineReader 11	
<i>Work time</i>	<i>% errors</i>	<i>Work time</i>	<i>% errors</i>
40 s	16	< 2 s	2

Not enough high quality recognition of the developed system compared with the commercial product due to the small size of the selected window.

IV. CONCLUSIONS

The main stages of processing of digital images for the tasks of character segmentation and subsequent recognition of texts are considered. The algorithm of preliminary processing using fuzzy logic and the process of binarization of the image is considered in detail. A fuzzy processing algorithm is constructed to draw the boundaries of characters in the image.

It is shown that the algorithm for solving the problem of syntactic analysis in the chosen grammar is a generalization of the corresponding algorithm for two-dimensional context-free grammars and consists in a sequential calculation for each fragment of a fine for assigning to it each name [19, 20].

The time and spatial complexity in the case of two-dimensional context-free grammars is very height ($O(H^2W^2(H + W))$ and $O(H^2W^2)$). This complexity limits the application of algorithms in practice.

The time and spatial complexity of these algorithms is determined first of all by the number of fragments that are reviewed in the course of their work. Therefore, the reduction of this number is the main way to reduce the complexity of the recognition algorithms.

The methods and algorithms that are considered allow us to approximate the methods of OCR systems to those that are used by people, because despite the great achievements in this area, there are no systems that could equalize the recognition of the text with the person.

ACKNOWLEDGMENT

The authors are appreciative to colleagues for their support and appropriate suggestions, which allowed to improve the materials of the article.

REFERENCES

- [1] S. Kumar, S. Chandrakar, A. Panigrahi and S. K. Singh, "Muzzle point pattern recognition system using image pre-processing techniques," 2017 Fourth International Conference on Image Information Processing (ICIIP), Shimla, India, pp. 1-6, 2017.
- [2] J. R. Balbin, M. P. Sejera, C. O. A. Martinez, N. A. M. Cataquis, L. M. H. Ontog and J. K. Toribio, "Cloud based color coding scheme violator plate detection through character recognition using image processing," 2017 7th IEEE International Conference on Control System, Computing and Engineering (ICCSCE), Penang, pp. 253-257, 2017.
- [3] S. Antoshchuk, "The automatized systems with the visual information processing design," International Conference Modern Problems of Radio Engineering, Telecommunications and Computer Science, Lviv-Slavsko, Ukraine, 2004, pp. 268, 2004
- [4] Y. Cao, T. Zhang, S. Zhang and B. Luo, "Forward scattering bistatic radar imaging method and practice data processing," in Journal of Systems Engineering and Electronics, vol. 22, no. 2, pp. 206-211, April 2011.
- [5] S. Stomiński, "Potential resource of mistakes existing while using the modern methods of measurement and calculation in the glare evaluation," IEEE Lighting Conference of the Visegrad Countries (Lumen V4), Karpacz, pp. 1-5, 2016.
- [6] I. Kalaykov and G. Tolt, "Fast fuzzy signal and image processing hardware," 2002 Annual Meeting of the North American Fuzzy Information Processing Society Proceedings. NAFIPS-FLINT 2002 (Cat. No. 02TH8622), pp. 7-12, 2002.
- [7] L. Dorosinskiy and F. Myasnikov, "Radarsignal classification algorithms synthesis and analysis," IEEE 2nd Advanced Information Technology, Electronic and Automation Control Conference (IAEAC), Chongqing, pp. 119-122, 2017.
- [8] P. Marino, V. Pastoriza, M. Santamarfa and E. Martinez, "Fuzzy image processing in quality control application," Sixth International Conference on Computational Intelligence and Multimedia Applications (ICCIMA'05), pp. 55-60, 2005.
- [9] M. I. Chacon, L. Aguilar and A. Delgado, "Definition and applications of a fuzzy image processing scheme," IEEE 10th Digital Signal Processing Workshop, 2002 and the 2nd Signal Processing Education Workshop., pp. 102-107, 2002.
- [10] P. Melin, C. I. Gonzalez, J. R. Castro, O. Mendoza and O. Castillo, "Edge-Detection Method for Image Processing Based on Generalized Type-2 Fuzzy Logic," in IEEE Transactions on Fuzzy Systems, vol. 22, no. 6, pp. 1515-1525, Dec. 2014.
- [11] R. Chellappa et al., "Towards the design of an end-to-end automated system for image and video-based recognition," 2016 Information Theory and Applications Workshop (ITA), La Jolla, CA, pp. 1-7, 2016.
- [12] D. Peleshko, T. Rak, and I. Izonin, "Image Superresolution via Divergence Matrix and Automatic Detection of Crossover," International Journal of Intelligent Systems and Applications (IJISA), vol.8, no.12, pp.1-8, 2016. DOI: 10.5815/ijisa.2016.12.01
- [13] N. Ito and M. Hagiwara, "Image description generation without image processing using fuzzy inference," IEEE International Conference on Fuzzy Systems, Brisbane, QLD, pp. 1-8, 2012.
- [14] L. Vivona et al., "Unsupervised clustering method for pattern recognition in IIF images," 2016 International Image Processing, Applications and Systems (IPAS), Hammamet, pp. 1-6, 2016.
- [15] Z. Qu, G. Xiao, N. Xu, Z. Diao and H. Jia-Zhou, "A novel night vision image color fusion method based on scene recognition," 2016 19th International Conference on Information Fusion (FUSION), Heidelberg, pp. 1236-1243, 2016.
- [16] J. Calvo-Zaragoza, A. H. Toselli and E. Vidal, "Handwritten Music Recognition for Mensural Notation: Formulation, Data and Baseline Results," 2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR), Kyoto, pp. 1081-1086, 2017.
- [17] E. Tapia and R. Rojas, "Recognition of on-line handwritten mathematical formulas in the E-chalk system," Seventh International Conference on Document Analysis and Recognition, pp. 980-984, 2003.
- [18] F. Rousset, N. Ducros and F. Peyrin, "A Semi Nonnegative Matrix Factorization Method for Pattern Generalization in Single-Pixel Imaging," in IEEE Transactions on Computational Imaging, vol. PP, no. 99, pp. 1-1, 2018.
- [19] F. Alvaro, J. A. S'nchez and J. M. Benedi, "Recognition of Printed Mathematical Expressions Using Two-Dimensional Stochastic Context-Free Grammars," 2011 International Conference on Document Analysis and Recognition, Beijing, pp. 1225-1229, 2011.
- [20] K. G. Subramanian, M. Geethalakshmi, A. K. Nagar and S. K. Lee, "Two-dimensional Picture Grammar models," 2008 Second UKSIM European Symposium on Computer Modeling and Simulation, Liverpool, pp. 263-267, 2008.

Forest Fire Monitoring System Based on UAV team, Remote Sensing, and Image Processing

Vladimir Sherstjuk
Kherson National Technical University
Kherson, Ukraine
vgsherstyuk@gmail.com

Maryna Zharikova
Kherson National Technical University
Kherson, Ukraine
marina.jarikova@gmail.com

Igor Sokol
Maritime Institute of Postgraduate
Education
Kherson, Ukraine
kherson.sokol@gmail.com

Abstract—This work presents the fire monitoring and detecting system for tactical forest fire-fighting operations based on a team of unmanned aerial vehicles, remote sensing, and image processing. The idea of such a system and its general parameters and possibilities are described. Functions and missions of the system, as well as its architecture, are considered. The image processing and remote sensing algorithms are presented, a way for data integration into a real-time DSS is proposed. The results of experimental research of the prototype system are presented. The combination of multi-UAV-based automatic monitoring, remote sensing and image processing techniques provides required credibility and efficiency of the fire detection.

Keywords—unmanned air vehicles; forest fire monitoring; remote sensing; image processing; fire detection.

I. INTRODUCTION

The first decade of XXI century is characterized by activation of forest fires whose effects are damaging to the environment and which require intervention methods and techniques adapted to the conditions and needs of each incident. A number and intensity of forest fires are significantly increasing year by year due to a growth of human activity and climate changes. Forest fires response operations are becoming increasingly important but challenging because they are traditionally based on visual observations and decision-makers estimations that are made under the high responsibility conditions in a lack of time. Thus, the problem of forest fire detection, monitoring and forecasting is very relevant for many countries. It stipulates the development of real-time decision-support systems (DSS) for the forest fire monitoring, detection and response.

However, developing such kind of DSS is a complex and non-trivial task because the forest fire is a process with unpredictable behavior. Taking into account inaccurate or missing data describing fire, incomplete scientific understanding of fire behavior, a forest fire is poorly modeled and predicted. This excludes the use of well-studied classical decision support approaches based on models, rules, etc. In such conditions, the efficiency of forest fire operations strongly depends on the availability of online fire monitoring and detecting tools. To build such tools, we can use a suite of modern methods and techniques, such as remote sensing, image processing, unmanned air vehicles (UAVs), etc., which should work synergistically. Thus, the research of ways of developing an UAV-based online forest fire monitoring and detecting system is the topic of our current interest.

II. LITERATURE REVIEW

The most important problem of each forest fire response operation is a timely fire detection. The success of a fire suppressing depends essentially on its early detection, which makes it possible to start suppression as quickly as possible while the fire is still small and well-controlled [1]. For this purpose, forest fire monitoring can usually be carried out as the observation of a wildland area in terms of fire activity. The monitoring process generally includes such activities as searching for new forest fire starts (i.e., from satellite imagery, watchtowers or aerial patrols) and detecting of fires, which is the task of determining that a fire exists and triggering an alarm to start a fire response operation. Thus, fire detection task can be reduced to resolving uncertainty about whether a fire exists or not [2]. Obviously, to start a response operation on the fire it also needs to be localized and characterized, so the fire monitoring task can also be aimed at the real-time computation of an evolution of the most important parameters related to the fire propagation based on the online observations [3].

Traditionally, satellite and airborne systems are used in order to have a broad overview of the forest fire evolution, but the monitoring activities are still carried out mainly by people. Watchtowers need to be carefully placed to ensure adequate visibility and are expensive and inflexible; thus they are usually allocated for monitoring high-value or high-risk situations [4]. A ground-based system for fire monitoring using static cameras has been presented in [5]. Given the fact that large-scale forest fire management requires a great number of cameras, the price and complexity of such system are unacceptable. Satellite-based systems [6] have been proposed for forest fire detection and monitoring, but the temporal and spatial resolutions of such systems are very low for the requirements of forest-fire fighting. Manned aircraft are large and expensive; their use highly depends on the weather conditions and requires the presence of aerodromes [7].

One of the new approaches to forest fire monitoring is online UAV-based monitoring using remote sensing techniques [8]. Using remote sensing data gives some advantages: the data acquiring is often less expensive and faster than from the ground; remote sensing allows capturing data across a wider spectrum that can be seen by the human eye; it can cover large areas including far away and inaccessible areas; remote sensing provides frequent updates. Recent years have seen a great progress in the field of using UAVs for forest fire monitoring, detection, and even fighting [9]. UAVs can perform long-time,

monotonous and repeated missions beyond human capabilities. However, uncertainty and distortions of received image frames due to vibrations and turbulence, as well as the inability to measure directly the parameters required for decision-making are significant drawbacks of this approach [10].

Thus, the integration of UAVs with remote sensing techniques can provide rapid, mobile, and low-cost powerful solutions for the forest fire monitoring and detection tasks [11]. Therefore, many researchers pay increased attention to using UAVs. The use of a single and complex UAV with sophisticated complex sensors has been investigated in FIRE project, while the cooperative use of a simpler UAVs' team was explored in European COMETS project and in a number of other projects surveyed in [7]. Many works related to the detection, modeling and forecasting of forest fires were carried out at the National University of Civil Protection (Kharkiv, Ukraine). As well, over the past few years, a project has been implementing at the Ukrainian Research Institute of Civil Protection (Kyiv, Ukraine) concerning the monitoring of fires and providing operational communication during the emergencies' response based on UAVs. Despite the positive results, which have shown the possibility of using UAVs in the forest fire response operations [12], many issues related to UAV-based forest fire monitoring and detection systems, including their remote sensing and image processing techniques, still remain insufficiently investigated, so need further research.

III. PROBLEM STATEMENT

Forest fire-fighting operation should start as early as possible and should be done as fast as possible to minimize the damage caused by fire. By means of the modern sensory technologies, a forest fire can be quickly and accurately detected. For this, UAVs can provide a full range of multi-sourced data for fire monitoring [13], which have the form of streams of great volumes and can be characterized by the following features [14]:

- 1) *volume*: the data are characterized by their great volume;
- 2) *variety*: the remote sensing data are multisource, multitemporal, and multiresolution;
- 3) *velocity*: the data are generated and processed at a high rate and should be analyzed in a real time;
- 4) *continuity*: the data come from sensors on a continuous basis.

We assume that the combination of automatic monitoring system based on UAVs and remote sensing techniques with an approximate model of forest fire spreading [15] can provide the required credibility and efficiency of the fire detection synergistically. The aim of this work is to investigate remote sensing and image processing techniques for a multi-UAV-based monitoring system for tactical forest fire-fighting operations considering the joint use of UAVs, remote sensing of different types and GIS-based common terrain model, and their integration with the approximate fire-spreading model.

IV. FIRE DETECTION PATROL MISSION

The multi-UAV-based tactical forest fire monitoring system should perform the fire detection in order to find

potential ignitions, detection of fire, triggering an event, and initializing further monitoring of the fire. In general, the forest fire monitoring system must provide real-time information to decision makers for the forest fire response operation.

The fire detection task can be generally broken down into two successive stages: fire search and fire confirmation. We assume that each UAV will perform a certain mission at each stage. Thus, in this paper we consider two types of missions:

- patrol mission (surveillance over the large region and performing the fire detection);
- confirmation mission (resolving uncertainty about whether a fire exists or not).

In the patrol mission, each involved UAV has its own flight plan that contains a pre-planned path as a sequence of waypoints. Flying along the pre-planned path, UAV observes terrain using onboard sensors and tries to identify fire automatically. It is clear that the flight plan generally is strictly bounded due to some limitations of UAVs' capability, such as duration, range, altitude, sensor resolution, etc. Depending on the size and characteristics of the surveillance region, various number of UAV can be involved in patrol mission simultaneously along their own pre-planned paths.

After the fire is detected, confirmation mission begins. In one of the options, the patrol UAV detected the fire can continue its patrol mission further, but it must trigger the event. Other UAVs with hovering capabilities should be sent to the detected fire location to hover at a safe distance and make confirmation. Another option is to change pre-planned paths of the patrolling UAVs to fly by a circle around the detected fire location in order to confirm it. If the detected fire is not confirmed, the patrolling UAVs resume their missions to surveillance of the region. Given the fact that performed missions differ in goals and requirements, we need the UAVs of different types equipped with the sensors of different types with a single ground command center. The basic structure of the multi-UAV-based forest fire monitoring system is illustrated in Fig. 1.

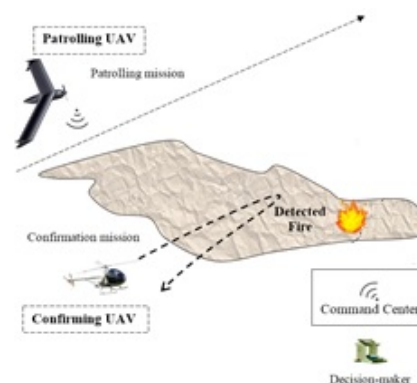


Fig. 1. Structure of the forest fire monitoring system.

The system includes the following components:

- 1) a multitude of UAVs (for the patrolling and confirming missions) equipped with the onboard sensors;

- 2) an infrastructure for the UAV ground support (launch and landing, maintenance) and equipment for the UAVs control;
- 3) specific algorithms/techniques for remote sensing and image/signal processing;
- 4) a dedicated ground command center that includes communication/computation equipment, geographic information system (GIS), and decision support system;
- 5) specific algorithms solving the fire detection, tracking, diagnosis, and prediction tasks.

V. VEHICLES AND SENSORS

Consider the used types of UAVs and sensors. The common requirements on the used UAVs are the following:

- *all-weather suitability*: all UAVs should perform their functions in both night-time and day-time even in the most difficult weather conditions;
- *self-localization*: a common reference terrain model should be used by all involved UAVs for automatic geo-localization of their spatial positions;
- *navigational autonomy*: sophisticated sensors (e.g. GPS receivers, inertial measurement units (IMU)) should be used by all involved UAVs for automatic flying along paths given by ground command center;
- *cooperation*: the UAVs should be able to coordinate their behavior and to cooperate with each other in order to solve their tasks optimally ;
- *payload*: the UAVs should carry all required sensors for fire perception purposes.
- *availability*: all UAVs should be equipped with onboard communication devices that guarantee receiving commands from a ground command center, sending information back to it, as well as exchanging information with the other UAVs.

The general structure of the used UAVs is illustrated in Fig.2. In accordance with the different missions' tasks to be performed, different types of UAV are used for those missions.

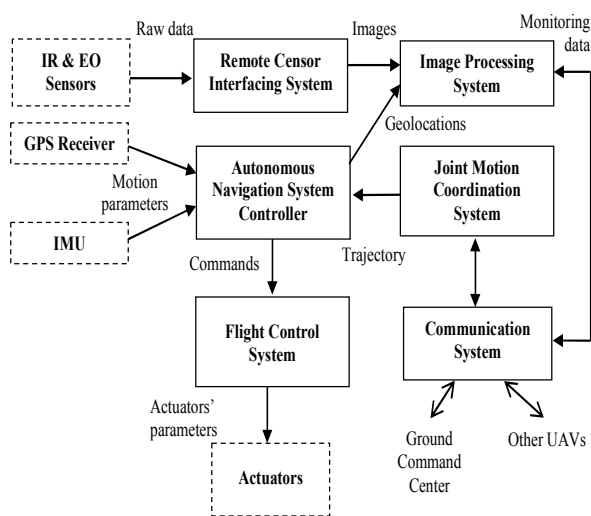


Fig. 2. Vehicle onboard control system

For the patrol mission in order to minimize the solution cost we use fixed-wing micro-UAVs with electric propulsion system having flight ceiling up to 2000 m, cruise speed about 90 km/h, takeoff weight up to 5 kg. This type of UAV provides the flight duration about 2.5 – 3 hours, and flight range with an online connection up to 75 km. It is equipped with a low-cost non-thermal (5-13 μm band) infrared micro-camera and a simple 12-megapixel optical camera with electronically adjustable focus.

We can use the same type of UAV for the confirmation mission. However, the confirmation mission can be performed much easier, faster and more efficient if we send the UAV with hovering capabilities to the detected potential fire location. Equipped with more precision (and, thus, more expensive) infrared and optical sensors, such UAV can hover for some time at a safe distance from the potential fire site and monitor it for verification purposes. Therefore, we use rotary-wing micro-UAVs with electric propulsion system having flight ceiling up to 180 m, cruise speed about 50 km/h, takeoff weight up to 7 kg. This type of UAV provides the flight duration about 1 hour, and flight range with the online connection up to 55 km. It is equipped with a precision gimbal that carries a wideband infrared camera and a 16-megapixel optical camera. A relatively small, but safe enough hovering altitude provides the good applicability of this type of UAV. Due to the higher resolution of the sensors and the closest approach to the observed point, we can achieve a much greater observation accuracy. In addition, this type of UAV can change the point of view relatively quickly, providing a higher efficiency of fire detection and confirmation.

VI. REMOTE SENSING AND IMAGE PROCESSING

The main goal of the forest fire monitoring is to continuously obtain information about the fire activity over a large wildfire region. The fire detection task performing is based on the common terrain model, which is a part of GIS-based DSS.

A. Common Terrain Model

The forest fires arise and spread through the certain area, represented by GIS as a certain terrain. The terrain model is based on the pre-built Digital Elevation Model (DEM) [12].

Firstly, the terrain is divided into a finite set of disjoint spatial objects presented as geometric shapes, which outline boundaries of the certain areas. Such spatial object is named as geotaxon and represents a geo-referenced limited natural part of the terrain with the same physical (or other) characteristics. For example, areas with the same features of the soil can be described as geotaxons of a certain kind. GIS can contain an unlimited number of geotaxons' layers. Secondly, a grid of isometric square cells $D = \{d_{ij}\}$ approximates the terrain and constitutes a certain GIS layer. Thus, each spatial object's location is discrete and bounded to a specific cell. The size $\delta \times \delta$ of each cell d_{ij} can vary, so the terrain scale can also change.

B. Image Processing During the Patrol Mission

During the patrol mission, the main sensor is the non-thermal infrared camera because obtained infrared images are not affected by smoke (the smoke is transparent for the used far infrared wavelengths). Besides that, the infrared camera is workable under either weak or no light conditions.

It does not provide temperature measures but only estimations of the radiation intensity, represented by colors. If a certain cell does not have any radiation, its pixels in the image are black. The appearance and increase of radiation change the pixels' color from black to white. Thus, if there are no white pixel in the image, it is very likely that there is no fire in the corresponding cell. However, some pixels can have grey colors.

At first, the brightness data is averaged within each cell d_{ij} , so the whole cell takes a certain average brightness B_{ij} . Then we use a partial order of grey colors (Fig. 3), which implies the ordinal color scale from black to white mapped on a numerical scale from 0 (black) to 1 (white). Thus, based on the brightness B_{ij} of each cell d_{ij} , we obtain a value μ_{ij} expressing the degree of ignition/burning possibility at this cell.

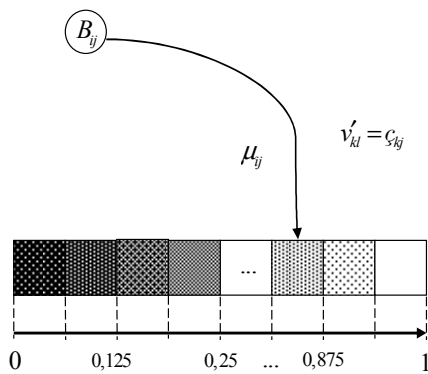


Fig. 3. Partial order of grey colors

Since various superheated or supercooled rocky areas, soils, and water surfaces can distort data obtained from the non-thermal infrared camera, we should simultaneously get data from the optical camera. An obvious feature of the presence of a forest fire is the smoke, observed as a light gray figure like a cone elongated in the wind direction. Thus, the processing of color images obtained by the optical camera is aimed at searching smoke or flame within them.

For this, firstly we define the discriminating interval in RGB space $LG = [RGB(60\%, 60\%, 60\%) - RGB(95\%, 95\%, 95\%)]$, which exclude some distortions caused by lighting conditions. We use texture-based classification method with the color-diffusion evaluation that differentiates smoke and non-smoke based on the counting of the number of pixels, which color belong to the interval LG , relative to the total number of pixels in the cell. These values could be averaged over a certain time interval (for example, 5 sec) for each cell d_{ij} and returned as a degree of ignition/burning possibility η_{ij} ranged from 0 to 1.

During the onboard automatic processing (Fig. 4), on the first step, we perform stabilization of the images. On the next step, we process the infrared and visual images by the above-mentioned methods. On the third step, we perform geo-localization, then geo-rectification. The UAVs' locations can be obtained using the GPS. The positions and orientation of the infrared and optical cameras can be computed based on their orientation angles and IMUs' data. To determine the position of each image pixel we use the

photogrammetric projective transformation that projects all points on the ground described by the DEM. DEM is also used to define the geospatial context (e.g., latitude/longitude/elevation) and timestamps for the geo-rectification. Thus, each image pixel becomes geo-referenced and can be mapped onto the grid. If both cameras are calibrated and DEM is available, we obtain the geo-referenced images in the common terrain model approximated to the cell level.

On the final step, we perform image fusion based on geo-referenced points. The values of the ignition/burning possibility μ_{ij} and η_{ij} that refer to the same cell d_{ij} for both images are combined by absorbing smaller values by larger: $v_{ij} = \max(\mu_{ij}, \eta_{ij})$. Then the array of resulting values v_{ij} should be transmitted to the command center. On the receiving side, we use a training-based threshold selection method to reduce false alarms. Using the given threshold v^* , we consider the fire detected only for the cells, which have greater or equal values of the ignition/burning possibility degree v_{ij} . The training stage could be performed by an experienced decision-maker on a set of training images identifies the conditions, under which the uncertain assessments v_{ij} can be considered as positive.

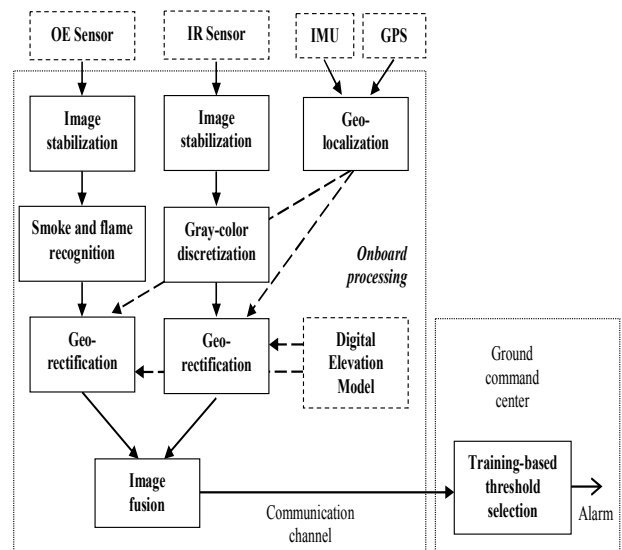


Fig. 4. Image processing during the patrol mission

VII. EXPERIMENT RESULTS

UAV onboard control system prototype was implemented using embedded microcontroller STM32F429 (180 MHz Cortex M4, 2Mb Flash/256Kb RAM internal, QSPI Flash N25Q512). Control center prototype used two servers HP ProLiant ML350 (Intel Xeon E5-2620, 8 cores up to 3 GHz).

The system has been tested with the multi-UAV team (3 drones for the patrol missions, 1 helicopter for the confirmation mission). All UAVs' cameras were precisely calibrated.

The performance of the developed system was studied in the laboratory conditions. It depends mainly on the cell size. The simulation experiment has been conducted varying the cell size from 5 m to 25 m as well as varying a number of discretization level in image processing algorithms.

Obtained results reflect that the bottleneck is the significant computational load of the UAV onboard control system. Further increase in processing power is limited by the available characteristics, and it requires to move some image processing to the ground command center.

However, the developed system has demonstrated the satisfactory probability of correct forest fire detection ($\approx 92\%$) in near-real-time conditions (processing time less than 2 min) with 10 m cell size and 16 levels of the color discretization.

Using the developed system, we achieved a good accuracy (up to 96%) of the fire spreading prediction for various terrains and weather conditions. Thus, the result of the experiment has shown that the developed system can provide required credibility and efficiency of fire prediction and response.

VIII. CONCLUSION

It can be concluded that the remote sensing techniques, which offer a variety of ways to detect and monitor forest fires, provide a great promise in solving the problems of forest fire management in real-time decision support systems.

It is shown how UAVs can be very helpful in fire-fighting response operations participating in fire monitoring and detection tasks. The combination of the multi-UAV-based automatic monitoring system and remote sensing techniques with an approximate model of forest fire spreading can provide the required credibility and efficiency of a fire detection and response. The developed system has demonstrated the satisfactory probability of the forest fire detection ($\approx 92\%$) in near-real-time conditions with processing time less than 2 min, 10 m cell size and 16 levels of the color discretization.

REFERENCES

- [1] M. C. Arienti, S. G. Cumming, and S. Boutin, "Empirical models of forest fire initial attack success probabilities: The effects of fuels, anthropogenic linear features, fire weather, and management," *Can. J. For. Res.*, vol. 36, pp. 3155–3166, 2006.
- [2] V. Ambrosia and T. Zajkowski, "Selection of Appropriate Class UAS/Sensors to Support Fire Monitoring: Experiences in the United States," in *Handbook of Unmanned Aerial Vehicles*, Springer Netherlands, 2015, pp. 2723–2754.
- [3] M. Zharikova and V. Sherstjuk, "Development of the Model of Natural Emergencies in Decision Support System," *Eastern European Journal of Enterprise Technologies*, vol. 4(73), no. 1, pp. 62–69, 2015.
- [4] R. S. Allison, J. M. Johnston, G. Craig, and S. Jennings, "Airborne Optical and Thermal Remote Sensing for Wildfire Detection and Monitoring," *Sensors*, vol. 16, no. 8, p. 1310, 2016.
- [5] J. Martínez de Dios, B. Arrue, L. Merino, A. Ollero, and F. Gómez-Rodríguez, "Computer vision techniques for forest fire perception," *Image and Vision Computing*, vol. 26, no. 4, pp. 550–562, 2007.
- [6] H. Olsson, M. Egberth, J. Engberg, J. Fransson, T. Pahlén et al, "Current and Emerging Operational Uses of Remote Sensing in Swedish Forestry," in *Proc. of the 5th Annual Forest Inventory and Analysis Symposium*, US Forest Service, Portland, USA, pp. 39–46, 2005.
- [7] C. Yuan, Y. Zhang, and Z. Liu, "A Survey on Technologies for Automatic Forest Fire Monitoring, Detection and Fighting Using UAVs and Remote Sensing Techniques," *Canadian Journal of Forest Research*, vol. 45, no. 7, pp. 783–792, 2015.
- [8] D. Kolaric, K. Skala, and A. Dubravic, "Integrated system for forest fire early detection and management," *Period. Biol.*, vol. 110, no. 2, pp. 205–211, 2008.
- [9] C. Yuan, Y. Zhang, and Z. Liu, "UAVs-based forest fire detection and tracking using image processing techniques," in *Int. Conf. on Unmanned Aircraft Systems*, pp. 639–643, 2015.
- [10] J. Martínez de Dios, B. Arrue, L. Merino, A. Ollero, and F. Gómez-Rodríguez, "Computer vision techniques for forest fire perception," *Image and Vision Computing*, vol. 26, no. 4, pp. 550–562, 2007.
- [11] H. Olsson, M. Egberth, J. Engberg, J. Fransson, T. Pahlén et al, "Current and Emerging Operational Uses of Remote Sensing in Swedish Forestry," in *Proc. of the 5th Annual Forest Inventory and Analysis Symposium*, US Forest Service, Portland, USA, pp. 39–46, 2005.
- [12] L. Merino, J. Martínez de Dios, and A. Ollero, "Cooperative Unmanned Aerial Systems for Fire Detection, Monitoring, and Extinguishing," in *Handbook of Unmanned Aerial Vehicles*, Springer Netherlands, 2015, pp. 2693–2722.
- [13] G. Chen, J. Zhao, L. Yuan, Z. Ke, M. Gu, and T. Wang, "Implementation of a geological disaster monitoring and early warning system based on multi-source spatial data: a case study of Deqin Country, Yunnan Province," *Nat. Hazards Earth Syst. Sci. Discussions*, vol. 2017, pp. 1–15, 2017.
- [14] M. Chi, A. Plaza, J.A. Benedittson, Z. Sun, J. Shen, and Y. Zhu, "Big data for remote sensing: challenges and opportunities," in *Proc. of the IEEE*, vol. 104, no. 11, pp. 2207–2219, 2016. DOI: 10.1109/JPROC.2016.2598228
- [15] M. Zharikova, V. Sherstjuk, and N. Baranovskij, "The Plausible Wildfire Model in Geoinformation Decision Support System for Wildfire Response," in *Proc. of the 15th Int. Multidisciplinary Sc. Geoconference SGEM-2015*, Albena, Bulgaria, vol. 2, book 3, pp. 575–583, 2015.

Evaluation of Objects Recognition Efficiency on Mapes by Various Methods

Yuriy Furgala
*department of electronic and computer
technologies*
Ivan Franko National University of
Lviv
Lviv, Ukraine
furgala@mail.lviv.ua

Yuriy Mochulsky
*department of electronic and computer
technologies*
Ivan Franko National University of
Lviv
Lviv, Ukraine
y.mochulsky@ukr.net

Bohdan Rusyn
*Karpenko Physico-Mechanical Institute
of the NASU*
Lviv, Ukraine
University of Technology and
Humanities
Radom, Poland
rusyn@ipm.lviv.ua

Abstract—The paper analyzes the efficiency of image recognition on terrestrial photographs by SURF, SIFT and ORB methods. It has been shown that for high-quality images, the highest probability of recognition in the application of the SIFT method. In the case of identifying fragments of images on noisy and blurred images, the best results are obtained using the ORB method, which, together with this, has the highest performance among the methods used.

Keywords —SURF, SIFT, ORB, recognition efficiency.

I. INTRODUCTION

The task of recognizing images in their arbitrary orientation in the image is solved by various methods, the effectiveness of which is evaluated differently for different types of images. When recognizing, it is usually not a problem if the rotation angle of the image is relatively small. Instead, at angles of more than 20 degrees, the recognition efficiency drops sharply. In order to solve the problem of recognition of images with high reliability it is necessary, first of all, to eliminate the dependence of the proposed method or algorithm on affine transformations, namely: parallel transfer, zoom and spatial rotation of the image, which is subject to classification. In addition, in the process of recognition, there are and distorting factors. First and foremost, this is a considerable noising and blurring of investigated images, which are typical distortions during their registration, caused by atmospheric phenomena and imperfect scanning means. The most commonly used methods for solving such problems are SURF (Speeded Up Robust Features), SIFT (Scale-Invariant Feature Transform) and ORB (Oriented Features from Accelerated Segment Test and Rotated Binary robust independent elementary features) [1-7]. In addition, it is known that in order to obtain accurate estimates of the probability of recognition, one can use the approach described in [8-10]. Comparison of the authenticity of recognition and the performance of these methods was carried out by many authors on various objects [3, 11-18]. However, unambiguous conclusions were not drawn, most likely, given the peculiarities of the implementation of methods in specific software solutions.

II. RESULTS AND DISCUSSION

The paper analyzes the efficiency of object recognition on images of the Earth's surface, in particular in city photographs. The quality of recognition is performed for both original high-quality images and for noisy and blurred

images. The templates used five fragments of the original image size 256*256 pixels, which were randomly selected on the original image of the map area of 3600*2120 pixels. The recognition program was created using SURF, SIFT and ORB methods implementations in the OpenCV library. The investigated image has a normal histogram. The three methods of specific points detection (SURF, SIFT and ORB) were determined for the range of angles between the orientation of the template and the image from 0 to 90 degrees at different levels of white noise and the size of the blurring matrix at different scaling ratios.

The effectiveness of the methods was determined as the ratio of special point's number on the investigated image that coincided with the corresponding points in the template, to the total number of special points in the template. We denote this value in work as the recognition efficiency a fragment of an image. The dependence of the recognition efficiency on the size of the investigated image was carried out in a manner where the initial image was successively reduced by half until no fragments were identified. The results of this study are shown in Fig. 1, from which we see that all methods give a 100% result for a half-sized image, but for a reduced 4-times only SIFT gives satisfactory results, SURF works well only for parallel and perpendicular patterns of the pattern and image, and the ORB reliably recognizes fragments for orientations close to 45 degrees between the template and the image. In the case of a reduction of the image 8 times, none of the used methods do not recognize changes in any fragment.

The study of the dependence of the recognition efficiency on the noise level was investigated on full-size images by applying a white noise with a sequential increase in the dispersion of the normal distribution. The image of the image at different levels of white noise is shown in Fig. 2, and the recognition results are shown in Fig. 3. The figure shows that for the range of dispersion values from $\sigma = 2$ to $\sigma = 16$, all methods have recognized all fragments. For case $\sigma = 32$, only 100% results were obtained using the ORB method, and the efficiency of recognition by SURF and SIFT is respectively about 90% and 80% respectively. For $\sigma = 64$, none of the methods could identify at least one fragment.

The study of the dependence of the recognition efficiency on the level of image blurring was carried out by using a blurring matrix whose size was $(2n + 1) * (2n + 1)$ pixels for $n = 1..5$. The image of the image at different levels of white noise is shown in Fig. 4, and the recognition results are

shown in Fig. 5. From the figure it can be seen that the recognition results for these methods differ markedly from one another. The worst result was obtained by applying the SURF method, which gave the probability of recognition at 70% for a 9*9 pixels blurring matrix and a zero result for a 17*17 pixels matrix. The best results are obtained for the SIFT method, in particular for the 9*9 pixels matrix there is almost 100% recognition, for 17*17 pixels, about 70%, and the lack of recognition for a matrix of 33*33 pixels. Instead, using the ORB method gives 100% for almost all images blurred by matrices up to 17*17 pixels. An exception is the parallel and perpendicular orientation of the image and the pattern for which the probability of recognition is noticeably lower. This behavior of the efficiency dependence of the ORB recognition method from the angle of the mutual orientation of the image and the template correlates with the corresponding dependence when the zoom factor is changed (Fig. 1).

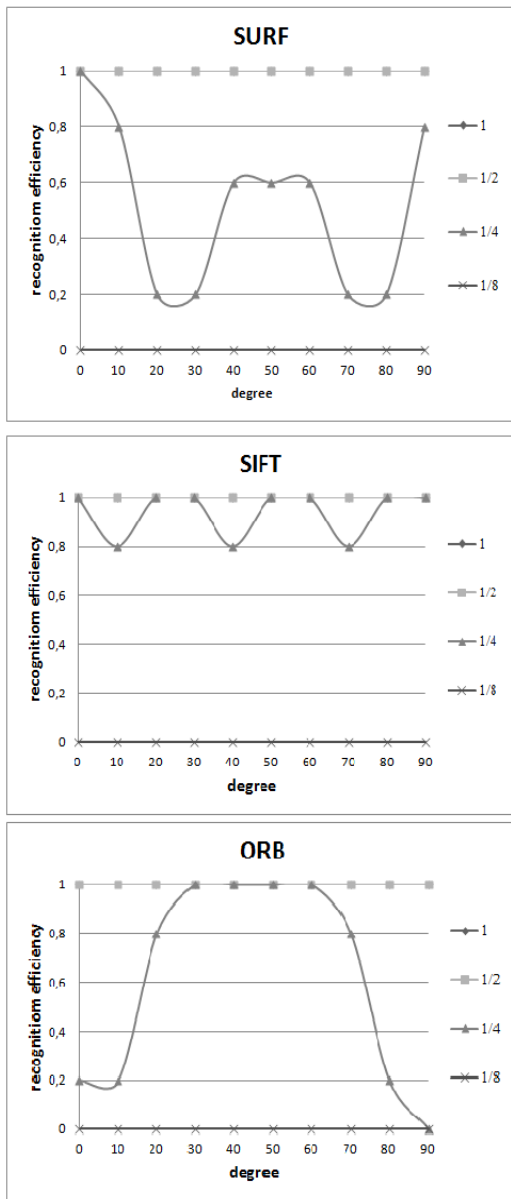


Fig. 1. The dependence of the recognition efficiency on the angle of rotation of the pattern relative to the image for different sizes of the image (1 is the image of the original size, 1/2 - the image is reduced twice, 1/4 - the image is reduced 4 times, 1/8 - the image is reduced 8 times) when used different recognition methods.

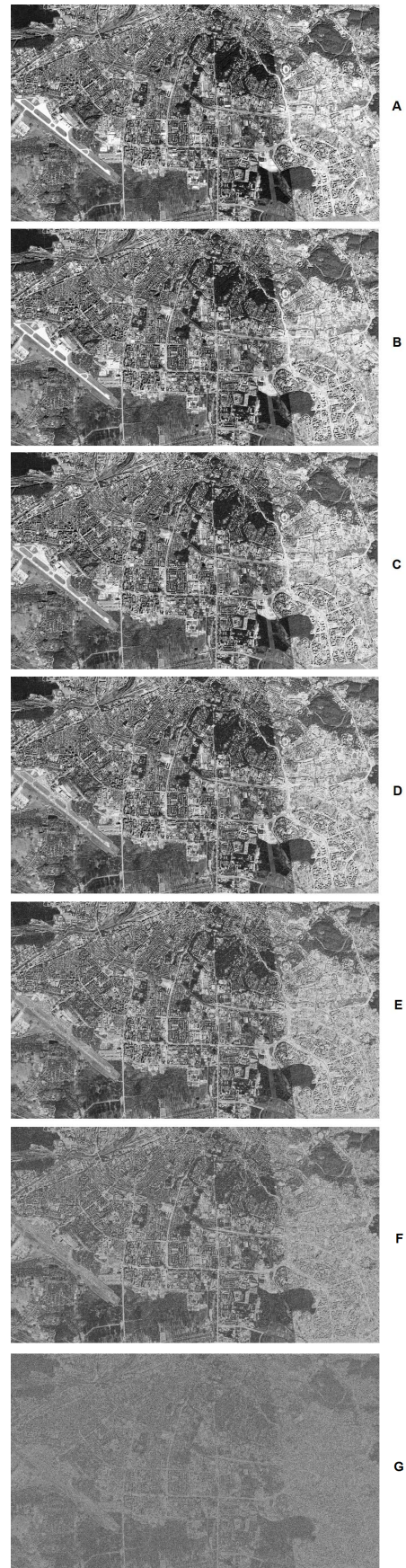


Fig. 2. Photos of terrain with different levels of additional noise (A – source, B – $\sigma = 2$, C – $\sigma = 4$, D – $\sigma = 8$, E – $\sigma = 16$, F – $\sigma = 32$, G – $\sigma = 64$).

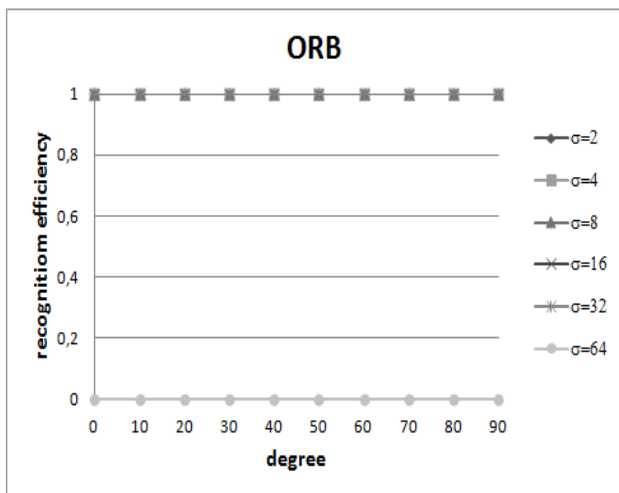
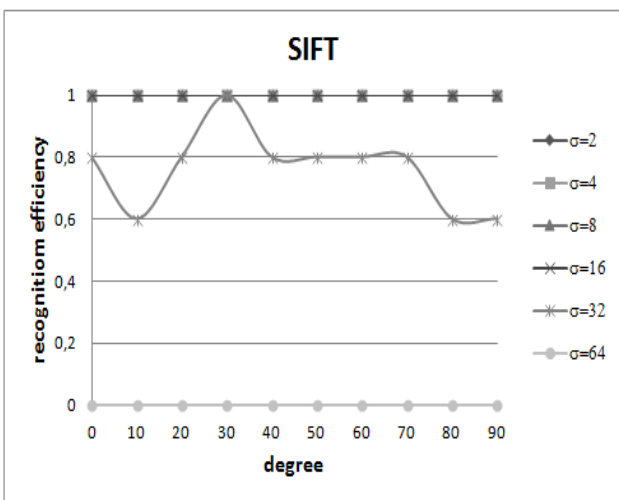
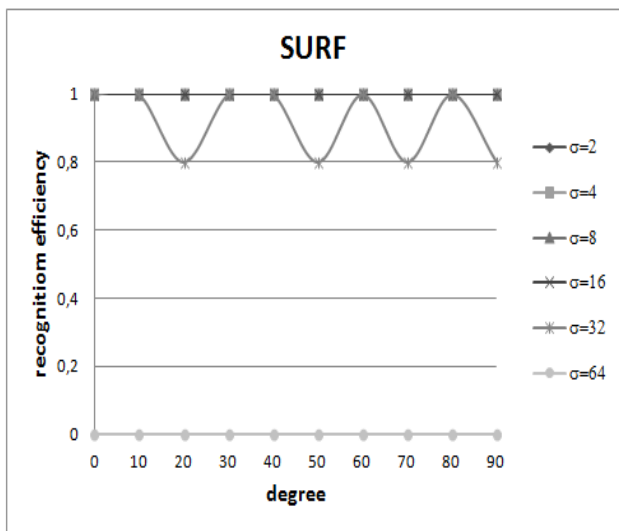


Fig. 3. The dependence of recognition efficiency on the angle of rotation of the pattern relative to the image for different levels of white noise (A – source, B – $\sigma=2$, C – $\sigma=4$, D – $\sigma=8$, E – $\sigma=16$, F – $\sigma=32$, G – $\sigma=64$) using different recognition methods.

Comparing the working times with these three methods indicates a certain advantage of the ORB method, for which the operating times was approximately two times less than the recognition time using the SURF and SIFT methods.

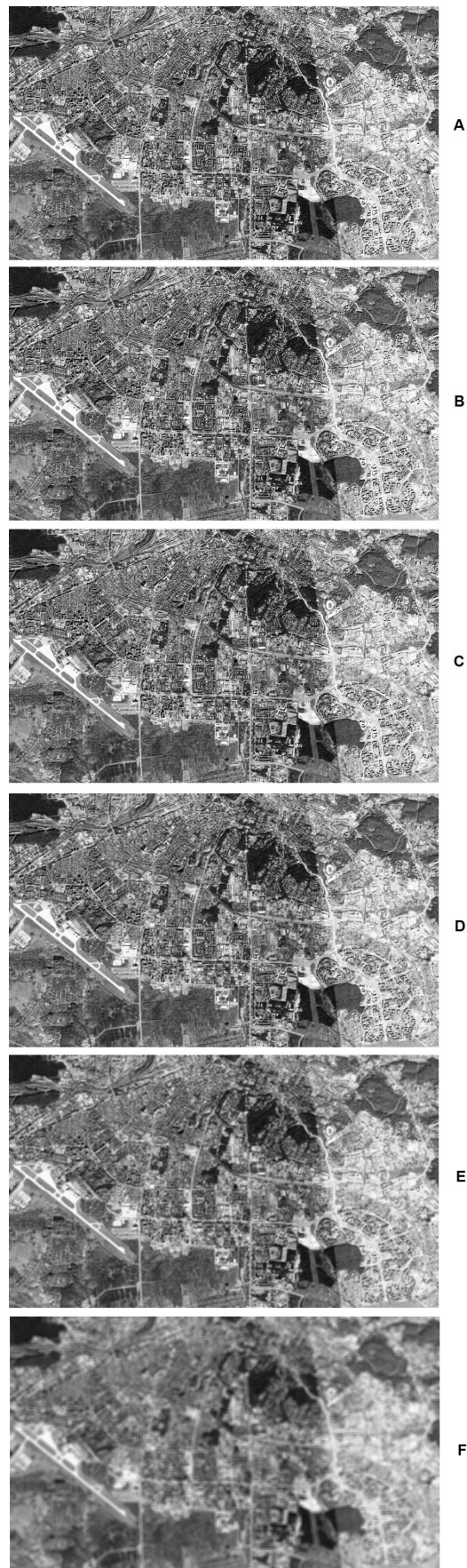


Fig. 4. Photos of terrain with various size blurring mask (A – source, B – 3*3, C – 5*5, D – 9*9, E – 17*17, F – 33*33)

III. CONCLUSIONS

In this work, the dependence of the efficiency of recognition of fragments of the image on the photographs of the area due to the influence of factors that impair the image quality and complicate the obtaining of reliable results is investigated. It is shown that for the ideal conditions for obtaining images, the best result is obtained using the SIFT method. At the same time, the recognition time among the methods used is one of the largest. In the presence of distortion of images such as noise and blur, the best results are given by the ORB method, which works more efficiently at lower image quality. Significantly higher is the efficiency of SURF and ORB recognition at a 45 degrees angle between the original and the investigated image compared to the parallel and perpendicular orientation, and the SIFT method is practically invariant to the angle of rotation of the pattern relative to the image. In general, it can be argued that when recognizing low-quality terrain maps, the best results should be expected when applying the ORB method.

REFERENCES

- [1] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, vol.60, issue 2, pp. 91-110, 2004.
- [2] Y. Ke and R. Sukthankar, "Pca-sift: A more distinctive representation for local image descriptors," *Computer Vision and Pattern Recognition*, pp. 506-513, 2004.
- [3] Luo Juan, and Oubong Gwun, "A Comparison of SIFT, PCA-SIFT and SURF," *International Journal of Image Processing*, vol.3, iss. 4, pp.143-152, 2010.
- [4] Ethan Rublee, Vincent Rabaud, Kurt Konolige, and Gary Bradski, "ORB: an efficient alternative to SIFT or SURF," 2011 IEEE International Conference on Computer Vision, pp.2564-2571, 2011.
- [5] Herbert Bay, Andreas Ess, Tinne Tuytelaars, and Luc Van Gool, "SURF: Speeded Up Robust Features," *Computer Vision and Image Understanding*, vol. 110, no. 3. – pp. 346-359, 2008.
- [6] P. M. Panchal, S. R. Panchal, and S. K. Shah, "A Comparison of SIFT and SURF," *International Journal of Innovative Research in Computer and Communication Engineering*, vol. 1, no. 2, pp. 323-327, 2013.
- [7] P. Sykora, P. Kamencay and R. Hudec, "Comparison of SIFT and SURF Methods for Use on Hand Gesture Recognition based on Depth Map," *AASRI Procedia*, vol. 9, pp. 19-24, 2014.
- [8] B. O. Kapustiy, B. P. Rusyn, and V. A. Tayanov, *The pattern recognition systems in small data base*. Lviv: SPOLOM, 2006. (in Ukrainian).
- [9] B. O. Kapustiy, B. P. Rusyn, and V. A. Tayanov, "A new Approach to Determination of Correct Recognition Probability of Set Objects," *Upravlyayushchie Sistemy i Mashyny*, iss. 2, pp.8-12, 2005.
- [10] B. P. Rusyn, *Structurally linguistic methods for pattern recognition in real time*. Kyiv: Naukova dumka, 1986. (In Ukrainian).
- [11] Seema Asht, and Rajreshwar Dass. "Pattern Recognition Techniques: A Review," *International Journal of Computer Science and Telecommunication*, vol.3, iss. 8, 2012.
- [12] C. Michael, V. Lepetit, S. Christoph, and F. Pascal, "BRIEF: Binary Robust Independent Elementary Features," *CVLab, EPFL, Lausanne, Switzerland*, p. 14, 2009
- [13] Rahul Das Gupta, Jatindra K. Dash, and Sudipta Mukhopadhyay, "Rotation invariant textural feature extraction for image retrieval using eigen value analysis of intensity gradients and multi-resolution analysis," *Pattern Recognition*, vol. 46, pp. 3256–3267, 2013.
- [14] Bin Xiao, Gang Lu, Tong Zhao, and Liang Xie, "Rotation, Scaling and Translation Invariant Texture Recognition by Bessel Fourier moments," *Pattern recognition and image analysis*, vol. 26, issue 2, pp. 302-308, 2016.
- [15] Frank Y. Shih, *Image Processing and Pattern Recognition: Fundamental and Techniques*. Wiley-IEEE Press, 2010.
- [16] Christopher M. Bishop, *Pattern Recognition and Machine Learning*. Springer, 2006.
- [17] Bernhard Zeisl, Pierre Fite Georgel, Florian Schweiger, Eckehard G. Steinbach, and Nassir Navab, "Estimation of Location Uncertainty for Scale Invariant Feature Points," *BMVC*. pp. 1-12. 2009.
- [18] Florian Schweiger, Bernhard Zeisl, Pierre Fite Georgel, Georg Schroth, Eckehard G. Steinbach, and Nassir Navab, "Maximum Detector Response Markers for SIFT and SURF," *VMV*, pp. 145-154, 2009.

Music Content Selection Automation

Tetiana Gladkykh
Data Science Group, SoftServe
Lviv, Ukraine
thlad@softserveinc.com

Taras Hnot
Data Science Group, SoftServe
Lviv, Ukraine
thlad@softserveinc.com

Roman Grubnyk
Data Science Group, SoftServe
Lviv, Ukraine
thlad@softserveinc.com

Abstract— We are proposing the solution for musical content recommendation, which is based on assessment of tracks similarity with taking into account tree factors - genre description, sound and rhythm patterns and user preferences. We have introduced the music compositions distance measure based on their representation as mel-spectrograms, and deep-learning approach to high-level (tags) music description, based on the extracted acoustic and rhythmic patterns from their spectra.

Keywords—music recommender; tracks similarity; mel-spectrogram; deep-learning; tags recognition

I. INTRODUCTION

If you look at a relatively close past, it becomes clear how much easier music lover life has become. Search of the right artist, composition or concert has become several orders of magnitude easier. Dozens of different services at our disposal that can satisfy the needs of even the most sophisticated music lover. Nevertheless, even with such music content variety, the problem of repertoire updating does not cease to be relevant - new performers working in the genre that is interesting right now, perhaps - in the mood; compositions that sound similar, have similar drums to something heard or liked. Existing services make the audio composition search directional and, certainly, ease the task of music content selection. They make it easier, but do not meet the challenge completely, because they work with a number of limitations. For example, the most popular service for music content search, Shazam [1, 2], uses prints of the spectra comparison as the core of the "search engine", so it turns out to be useless in case of, for example, a cover version of a famous work or new music content. Last.FM recommendation service [3, 4] allows each registered user profiling in order to perform his positioning regarding other registered participants and predict what else might be interesting for him based on the auditions general history. Among the weaknesses a fairly high secondary content percentage in the recommendations can be noted. The inability to recommend in the case of a new work and / or a new user. The Pandora service [4] based on comparison, which rests on the "contents" of a music piece evaluation that is expressed in the several hundred attributes set, provided by professional musicians. Yandex.Music [6], uses only user's listening history, it does not allow to segment songs by genre and directions, therefore, a fairly significant history of the user's activity is required to get adequate recommendations, and, as with two previous services, the problem of a "cold start" remains. If you take a whole galaxy of similar services, like TuneGlue, Music Roamer, Music-Map and others, they simply build a tree of compositions similarity, relying only on the musical works metadata, and can hardly be classified as a recommendation system.

We offer functionality that in many respects repeats existing services capabilities, in fact, combining them within

a single product. But the distinctive feature of our solution is the search for works that sound alike, relying on the sound and rhythmic pattern, even if musical fragments don't match exactly.

II. RELATED WORKS AND OUR CONTRIBUTION

In the context of music recommendation, we can mention three general approaches: 1) recommendation based on musical compositions metadata (like set of tags that describe musical genre, direction, artist, etc.), 2) recommendation based on context, like playlist, web-based co-occurrences, etc. 3) recommendation based on music feature extraction.

There are a lot of works are related to the first approach, it's general advantage – relatively simple realization in the context of huge songs datasets and possibility to use user-based descriptions. But these methods require extremely detailed description in order to increase the results relevancy and can't be applied in the case of new composition. The second approach allows to assess songs similarity based on the principle that two songs should be considered as similar if they are mentioned in the same context, so recommendations, in this case, may include so-called user-based similarity – similarity that is based on the user's rates. The main advantage of this approach – the only information about song, that we should get, is the context, but this approach has significant restriction – we need some historical information about composition, so it also can't be applied to some new song.

Recommendations that are based on music feature extraction also divided into two groups – high-level low-level features based recommendations. The low-level features [7] describe any audio signal in the form of well defined and determined acoustic features like: loudness, spectrum powers, brightness, bandwidth, pitch and cepstrum. The main disadvantage of these features – they can't be easily used for understanding so-called structure of music to users without technical skills in this subject. This disadvantage is not specific for, so-called, high-level features – composite music characteristics like melody or harmony. This features describes the type of knowledge that a listener may extract, recognize and understand from one or other piece of music. There are works related to high-level feature extraction based on the chromagram analysis and estimation of the basic frequency corresponding to the pitch of the predominant melody with different modifications [8, 9]. All of these algorithms allow to extract complex music characteristics, based on the generalizations of the low-level music features processing, and representation them in more understandable form. So, high-level features may be considered as high-level interpretation of the set of low-level music features and, according to the recent works, low-level features are indispensable in the context of machine-learning approaches to music processing, understanding and tracks similarity assessment. For example, in work [10, 11] the similarity

between spectrum are used for assessment similarity between corresponded audio-contents, in [12,13] the same representation was used for genres and artists' recognition.

In our approach we proposed algorithm of music compositions similarity assessment based on acoustic and rhythmic patterns that can be extracted from musical tracks' spectral representation. Moreover, we proposed deep-learning approach to high-level music description based on the the same initial representation. In addition to the above-mentioned, in our solution we combine several approaches to music similarity assessment – based on the songs' metadata and on the registered users' preferences analysis, so, finally we provide an opportunity to give recommendations by the extraction of the acoustic perception of the composition user like, which is supported by automatic identification of the genre, style and direction results. This, on the one hand, allows giving accurate estimates of whether a certain composition of previously unknown artist, will be liked by some registered user, taking into account their personal preferences. And, on the other hand, to select adequate content for new users by analyzing their audio library. Below we will consider the approaches underlying the musical works similarity evaluation based on the genre description, sound and rhythm pattern, registered users' preferences, automatic genre and stylistic affiliation of the music content determination, and the music content selection automation.

III. SIMILARITY ASSESSMENT OF MUSICAL COMPOSITIONS

To obtain a compositions list that can be recommended for listening to one or another user, a comprehensive method to evaluate similarity of musical works was developed. It includes a similarity assessment in three areas: 1) based on genre description - two compositions are considered close when they are described by a close set of genre tags; 2) based on sound and rhythm patterns - two compositions are considered close when they are characterized by close sound and rhythm patterns; 3) based on the registered users' preferences analysis - two songs are considered close when users with close preferences like them.

A. Similarity evaluation based on genre description

When we assess the compositions similarity, it is necessary to take into account many different factors, including the track description in the form of a tags set stamped by individual users. The more users provide this kind of characteristics, the more likely they match to track. Tags that are characterized by a large number of matches can be considered the most significant composition characteristics, since people with different, in most cases, preferences, were solidarity with their descriptions. The separate tag importance as a track's characteristic depends on the following factors: 1) **popularity** - depends on the number of users who in their track description indicated this characteristic; 2) **uniqueness** - a value indicating how this characteristic distinguishes a track against other tracks. So, the most popular tags are tags related to genres - general musical directions, such as rock, hip-hop, jazz, etc. Unique tags include, first of all, the performers' names, the album name, specific sub-genres, etc. For example, if we select compositions in which some genre direction is mentioned, $track = \{ \langle tag_i \rangle; \langle genre_j \rangle \in track, a \text{ composite tag can be considered as a sub-genre that contains the genre addition: } \langle sub_genre_{jk} \rangle = \{ \langle genre_j \rangle, \langle description_k \rangle \} \in track.$

Subgenre can be interpreted as a clarifying tag if it occurs much less frequently than the corresponding genre. So, if we consider 2 genre directions, we get the following frequency distribution for genres and sub-genres (Fig. 1).

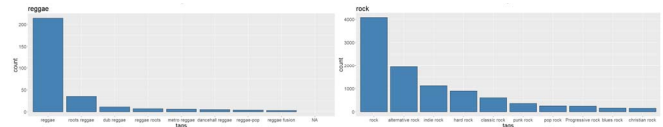


Fig. 1. Genres and Sub-genres distribution

You can see that subgenres may occur 4 times less frequently than the corresponding global directions, although such subgenres as "heavy metal", "alternative metal" and "alternative rock" can be considered as separate genres. Accordingly, when describing compositions, we can distinguish three tags' categories:

- General - genre affiliation. Tags are characterized by a high mention frequency in the context of a large compositions number
- Clarifying - subgenre affiliation. Tags are mentioned in a much smaller songs number than common tags, but they accompany the main genre tag
- Unique - artist or album characteristics. Tags are mentioned in the context of an extremely small tracks number, but each is one of the most popular particular work description

To rank the tags, we used a well-known statistical measure to evaluate the word importance in the context of a document that is part of the corpus - TF-IDF [14]. In our case, TF is the relative number of users who described a certain track with a certain tag, and IDF is a uniqueness measure, depending on the relative number of documents in which the tag was mentioned. Thus, each track can be described by the set VTrack of TF_IDF coefficients of tags mentioned at least once in the compositions' description. For this purpose, we used a bag-of-words model for documents presentation (in our case, tracks). Similarity between compositions S_x и S_y is estimated as the cosine distance between the corresponding vectors:

$$Dist_{S_x S_y} = 1 - \cos(VTrack_x, VTrack_y) \quad (1)$$

The proposed approach result is shown in the fig. 2. The figure shows a tag cloud describing the source composition – "India Arie – Beautiful" and one of the recommended songs – "Erykah Badu - Bad Lady". As you can see, the compositions perfectly correspond with each other in terms of genre and style descriptions.

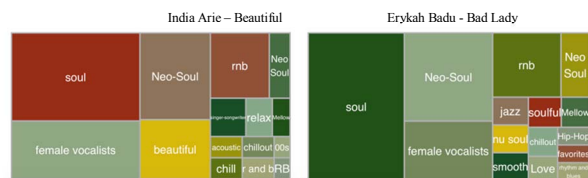


Fig. 2. Tags-based similarity

B. Similarity evaluation by sound

The second approach to similarity evaluation of compositions is based on the sound and rhythmic pattern formalization of individual tracks and their subsequent comparison.

Sound and rhythm patterns

At the core of the proposed solution is based on attempt to take into account the sound perception features by the organ of human hearing. This perception is estimated by a psychoacoustic value - the pitch, unit of which is "Mel". By definition, "pitch" [15] is "the sound quality determined subjectively by a person using ear." Mel is an off-grid pitch unit, and to quantify system uses the results of data statistical processing on sound subjective perception [16-18]. The audio signal (sound) can be described by a set of so-called mel-cepstral coefficients - a representation of the spectrum power in the mel-frequency, obtained with separation of individual spectrum frequency range. To do this, windows that are evenly spaced on the mel-axis are used. To represent the music track, it is preliminary divided into fragments of short duration - about 23 ms, in order to describe the signal spectrum change character in time (the cases of signal non-stationary in the sections under consideration may be disregarded). The track fragment spectra combination makes it possible to describe the input signal in the form of a spectrogram, a two-dimensional function that displays the spectral power density of the signal dependence on time. In order to take into account the auditory perception peculiarities of sound to humans, we proceed to the mel-cepstral coefficients - the orthogonal logarithm mapping of the energy spectrum square at certain frequencies for a certain period of time:

$$S[n] = \sum_{m=0}^{M-1} P[m] \cos(\pi n (m + 0.5)/M), 0 \leq n < M$$

where,

$$P[m] = \ln(\sum_{k=0}^{N-1} |X_a[k]|^2 H_m[k]), 0 \leq m < M$$

$$X_a[k] = \sum_{n=0}^{N-1} x[n] e^{-\frac{2\pi i}{N} kn}, 0 \leq k < N$$

$$H_m = \frac{(k-f[m-1])}{(f[m]-f[m-1])}, f[m-1] \leq k < f[m]$$

For each 23 ms fragment, we obtain a sequence of 40 cepstral coefficients, combination of which is a mel-spectrogram. The diagram (fig.3) shows visualizations of distinctive representatives of different musical genres.

Compositions similarity assessment

With a musical composition compact representation in which the person's auditory perception features are laid, we used it to assess the compositions similarity by sound. The core of the proposed method is the algorithm usage of the time scale dynamic transformation, which allows finding the optimal correspondence between time sequences.

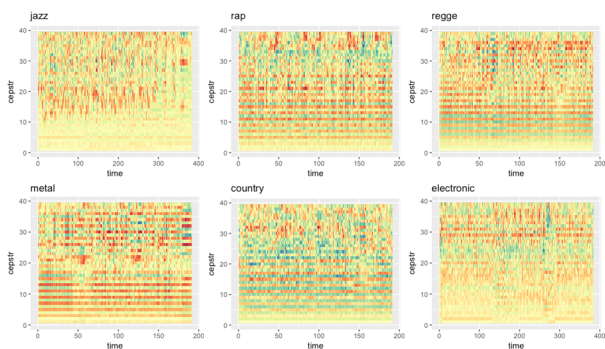


Fig. 3. Different genres spectrogramms

This algorithm is often used to compare time series and allows you to find a variant of their best alignment in order to level the error from an incorrect estimate of the distance

associated with the possible rows displacement to each other. In its implementation, the algorithm is close to estimating the editorial distance when comparing two lines. The core of it is to build a distances matrix between all points pairs of the analyzed sequences A and B, after which the so-called transformation matrix is constructed:

$$D_{ij} = d_{ij} + \min(D_{i-1,j}, D_{i-1,j-1}, D_{i,j-1})$$

The distance between the sequences is the last element of the matrix. Number of points is m:

$$DTW(A, B) = D_{mm}$$

In our case **time sequence** - ordered in time set of mel-cepstral coefficients calculated for a fragment with a duration of 23 ms. Accordingly, each such set is interpreted as a **sequence element**. The distance matrix is filled with cosine distances between the elements of 23 millisecond sequences of two compositions **A** and **B**:

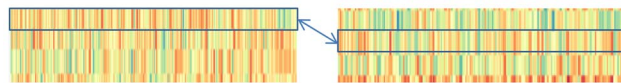


Fig. 4. Fragments comparison visualiaztion

$$d_{ij} = 1 + \cos(A_i, B_j)$$

Two compositions, represented by spectrograms S_z and S_x , are considered as close when they have quite a lot of matches in rhythmic pattern and sound. This means that we need to segment the tracks into sections that are long enough to catch this picture, but at the same time short enough to make the number of cases, when segment contains two or more sound patterns not too big. Each composition is characterized by a set of N "control" segments, therefore, their comparison consists in a complex estimation of the DTW distance between pairs of all control observations:

$$\text{Dist}_{S_z S_x}(N) = \text{quantile}_{25}(\{DTW_{r,k}\})$$

where $DTW_{r,k} = DTW(S_{zr}, S_{xk})$, $r, k = \overline{1, N}$, N - number of control segments.

The optimal segment duration - 3 sec, was established empirically, by comparing clustering fragments results of 10 compositions with the corresponding classes (each composition is characterized by one class). It is expected that with properly estimated segment length, each of the resulting clusters will consist almost entirely of one composition fragments. The evaluation was carried out in accordance with the Rand index:

$$\text{Rand} = \frac{SS+DD}{DD+DD+SD+DD}$$

where SS - the number of elements pairs belonging to the same class and to one cluster, DD - the number of elements pairs belonging to different classes and different clusters, SD - the number of elements pairs belonging to the same class and to different clusters, DS - the number of elements pairs belonging to different classes and one cluster. Target optimization function in the context of segment length and criterion $Rand$: $\min_i \left(\text{len}_i, \frac{1}{\text{Rand}_i} \right)$.

The following figure shows an example of visualizing the t-SNE transformation of 10 different compositions. The compositions belong to different genres and are described by

a mel spectrograms combination for different window sizes (1 and 3 seconds, respectively).

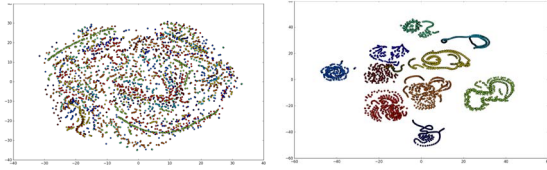


Fig. 5. T-SNE transformation of 10 different compositions

As can be seen, in the second case the compositions were divided into compact, well-divided groups, which confirms the result validity. The second parameter is the number of control fragments. The DTW distance between composition fragments estimating procedure is laborious enough - the computational complexity of $O(N^2)$, hence the computational complexity of estimating the distance between compositions is - $O((M*N)^2)$, by a factor of 1.5. The optimal number of segments was established on the basis of significance estimating of the discrepancy between the inter-composition distance for songs of one and different genres. The objective function should ensure the number N minimization of fragments with discrepancy maximization within and between genre distances:

$$\min_N \left(N, \text{Dist}_{S_z S_x}(N) / \text{Dist}_{S_z S_y}(N) \right)$$

where S_z, S_x - one genre, S_z, S_y - different genera

The following graphs show inside inner- and outer-genre distances, which depends on the number of tracks fragments:

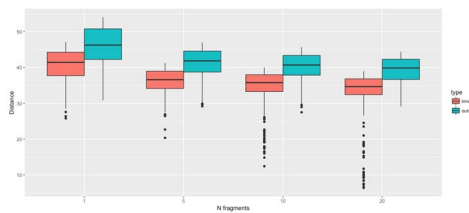


Fig. 6. Inner- and Outer- genre distances

As can be seen on the given diagram with the fragments number equal to 10, we get a satisfactory separation of the compositions inside and between the genre subgroups. With the number increase of picture fragments, the result improves slightly, accordingly number increase of segments is not justified. The similarity evaluation result of compositions by sound is shown in the Fig. 7. As in the previous example, the figure shows a tag cloud for the original song style and genre - Miles Devis - "Move" and one of the recommended songs - Charles Mingus - "Tonight at Moon". As it is not difficult to see, the recommended composition belongs to the same direction as the original composition.

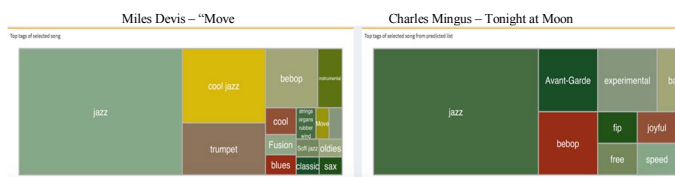


Fig. 7. Sound and rhythm based similarity

Similarity evaluation based on the preferences of registered users analysis

In addition to the objective characteristics of musical tracks, such as genre, rhythmic and sound patterns, which were discussed earlier, the closeness between individual compositions can be assessed on the basis of individual users' subjective preferences. This technique is used in one of the methods, that is used in recommendation systems construction - item-to-item collaborative filtering, which is based on two objects similarity evaluation based on user estimates. Objects can be considered similar if they are liked by the same user group, or by users with similar preferences. In our case, we are dealing with a listening number matrix of users' songs. The value of non-zero matrix elements can be interpreted as a kind of track rating according to the user - the higher the number of plays, the higher the confidence that the composition is included in the list of his preferences.

$$\mathbf{R} = (r_{ij}), \quad i = \overline{1, N_u}, \quad j = \overline{1, N_s}, \quad r_{ij} \geq 0$$

Because the plays number is not limited from above, unlike the objects rating, the matrix should be normalized with taking into account the user's activity. As the popularity rating of a song, we can correlate the plays number of a song with the upper bound of the corresponding $\alpha_q(R_i)$ distribution:

$$\text{rate}_{ij} = \max\{\alpha_q(R_i) | r_{ij} \leq \alpha_q(R_i)\}, \quad \alpha = 0.1n, \quad n = \overline{0, 1}$$

Since the tracks number is characterized by a histogram with right-hand asymmetry ($Sk(\text{listening}) > 1.5$) (Fig.8, left), relation between rating and listening number can be represented by a logarithmic function (Fig.8, right). Parameters a and b depend on the user activity degree.

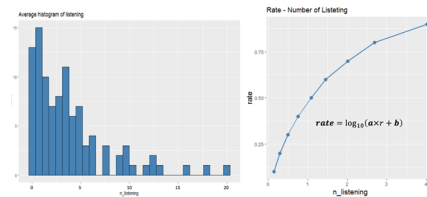


Fig. 8. Distribution of listening number and Rate function

Because the listening matrix is extremely sparse (about 1% of non-zero cells), the most optimal approach to collaborative filtering implementation is a model-based approach. As estimation model for users, a latent semantic model was chosen that allows to describe the system (user - composition) through a set of latent links. The solution is based on the listening matrix LU-factorization R' , which allows describing it as:

$$R' = \begin{matrix} & L^T & \times & U \\ N_u \times N_s & N_u \times d & & d \times N_s \end{matrix}$$

where d - significant factors number, $U_j \in U_{d \times N_s}$ vector of j composition in latent factors space, $L_i \in L_{N_u \times d}$ vector of i user in latent factors space

Users and compositions representation in the latent factors space, their optimal number was evaluated based on the weighted estimates original matrix restoration error. The error weight is a non-decreasing function of the listening songs number and is given by the following expression:

$$w(r_{ij}) = \begin{cases} \alpha, & r_{ij} = 0 \\ r_{ij}^\beta, & r_{ij} > 0 \end{cases}$$

where α – error weight with zero plays number, β - scaling factor

Optimization functions looks like this:

$$\sum_{ij} (w(r_{ij})(r'_{ij} - L_i^T \times U_j)^2) + \lambda (\sum_i \|L_i\|^2 + \sum_j \|U_j\|^2) \rightarrow \min$$

With such a representation, the similarity estimation between two compositions S_z and S_y can be reduced to distance estimation between corresponding vectors in the latent factors space:

$$\text{Dist}''_{s_z s_y} = 1 - \cos(U_z, U_y)$$

The result of the proposed method is shown on Fig.9. As you can see, the recommended composition shows well the original work stylistics, and, consequently, it can be expected that the recommendation will be relevant to the user's preferences.



Fig. 9. User preferences based similarity

IV. NEW COMPOSITION AUTOMATIC TAGGING

In advisory systems, the recommendation of new, previously unknown content, is particularly difficult, since there is often not enough information for its positioning among already existing objects. To solve this problem, we have developed a system of automatic musical work description (tagging), relying only on the analysis of its rhythmic and sound pattern. The solution is based on a model that allows the composition to be assigned to one of the 100 predefined stylistic classes based on the mel-spectrograms analysis of its fragments sets.

Stylistic classes are compositions collections that are close to each other in a genre descriptions combination expressed as a set of tags, and were obtained on the basis of more than 10.000 musical works in different genres and directions clustering. Clustering was performed using the method of agglomerate clustering with the optimal cluster number estimation using the Duda-Hart method. The closeness between the compositions is based on the genre description, and is given by (1). For example, the first class (cluster) consists of 288 works with the dominant genre of punk rock. The largest cluster contains 628 elements and is formed with compositions related to the set of R&B, POP and Hip-Hop. One of the small clusters (36 elements) refers to an alternative musical direction, uniting musicians from Iceland. The resulting clusters can be characterized by common (intersecting) tags, but their set is unique. Thus, each composition can be associated with one cluster (a unique set of describing tags) that reflects its genre and stylistic features and can be recognized on the basis of rhythmic and musical image analysis of this work.

The proposed solution is based on the recognition model of the genre and stylistic class by analyzing the mel-cepstral coefficients set, that describe each of the 3-second fragments of musical work. The general scheme of the automatic tagging system for the musical composition is shown in the Fig.10.

To model input $G_{\text{genre}}(x)$ of genre and stylistic recognition mel-spectrogram $S_{T x_i}$ is submitted. It describes 3 second fragment x_i of musical composition T. Model returns

100-component vector $V_{T x}$, each element of which contains a confidence degree that the recognizable fragment belongs to one of 100 classes:

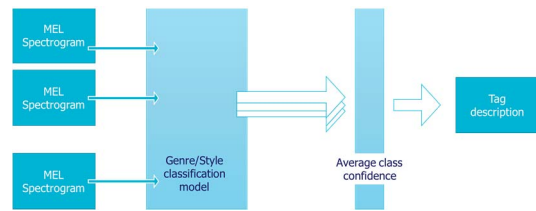


Fig. 10. General scheme of the automatic tagging system

$$V_{T x} = G_{\text{genre}}(S_{T x})$$

The composition genre is determined by averaging its fragments recognition results:

$$V_T = \frac{1}{n} \sum_{i=1}^n V_{T x_i}$$

Final recognition of song's cluster is the set of top-n clusters, based of their confidence levels. The classifier model is based on an artificial neural network consisting of three convolutional and two fully connected layers. For model training we used 10'000 compositions of 30 sec lengths. Model accuracy was assessed like percent of correct clusters recognitions, when final recognition is the top-n clusters is equal to 65%, 72% and 86% for top-1, 2 and 3 clusters correspondingly. So, final set of tags is detected as set of n dominant tags from three clusters-winners:

$$\text{Sel_Tag} = \underset{n}{\text{argmax}} \left(\left| \left(\sum_{j=1}^3 \sum_r \sum_i \text{TF}_{\text{IDF}_{irj}} \right) \right| \right)$$

V. PLAYLIST COMPOSING

Approaches and functionality described above are the core of solution that allows to propose to user the list of compositions, which are relevant to their preferences, based on the analysis of several uploaded compositions. General flow as follows:

- Detection n songs, which are most close to the analyze composition / compositions based on rhythmic and sounds patterns (top-n (MFCC))
- For each song from top-n (MFCC): detection top-m songs, which are close by tags descriptions (top-m(Tags))
- For each song from top-n (MFCC): detection top-r songs, which are close by user preferences(top-r (Users))
- Play list consists of songs from all three lists – { top-n (MFCC), top-m(Tags), top-r (Users)}

First stage is extremely time-consuming, because it requires assessment of MFCC distance between analyzed song and all songs from service database. In order to decrease the searching space, preliminary stage with detection the tags clusters is included to the flow. So, top-n MFCC distance based closest songs, are detected within songs from the clusters-winners (top 3 tags cluster). Final result is represented on the Fig. 11.

Analyzed song – “Love, love, love” by Monsters and Men was recognized like song that is characterized by following tags: Alternative Rock, Indie, Pop, Folk and Acoustic. Top 10 songs are represented on the figure (top-right part). This list –

is the list of the most relevant songs, songs that are close to the initial song by rhythmic and sound patterns. Below we can see additional songs, which were added based on tags and user based similarity. The most of songs belong to the artist from the first list, but there are some other compositions, which differ from initial song by sound, by very likely to be interesting to user.

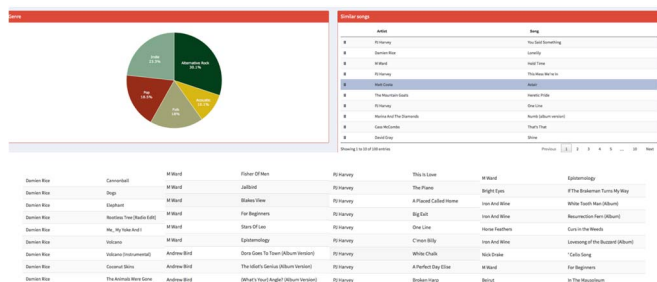


Fig. 11. Playlist composing

VI. CONCLUSIONS

In this work we have described the complex solution for musical content recommendation, which is based on tracks similarity assessment according to tree aspects: genre description, sound and rhythm patterns and the results of registered users' preferences analysis. We have proposed the distance measure between music compositions, which allows to assess the similarity between two and more tracks based on their representation as mel-spectrograms, and deep-learning approach to high-level (tags) music description, based on the extracted acoustic and rhythmic patterns from their spectra. Proposed solution allows to extract and describe user preferences based on perception of the compositions their like and provides the recommendations, which are supported by tags-based and user-based songs similarity. This, on the one hand, significantly improves recommendations that are based on users listening history and content-based tracks similarity only due to the possibility to put rhythmic and sound patterns of preferable compositions to the center of recommendations, but, on other hand, allows to solve so-called "cold start problem" – recommendation for new user, which doesn't have listening history or recommendations of new content (unknown genre and/or artist).

VII. REFERENCES

[1] "Shazam Launches Resonate TV Sales Platform," Billboard. 5 August 2014. Retrieved 15 June 2015.

[2] Bryan Jacobs, "How Shazam Works To Identify (Nearly) Every Song You Throw at It". Gizmodo. Retrieved 13 June 2017.

[3] Elia Alovisi, Last.fm: Was the Only Music Social Network That Made Sense, December, 2017, [https://noisy.vice.com/en_us/article/a37x9g/lastfm-was-the-only-music-social-network-that-made-sense]

[4] "Pandora and Last.fm: Nature vs. Nurture in Music Recommenders," Words & Numbers, A blog by Steve Krause, January, 2006 [http://blog.stevekrause.org/2006/01/pandora-and-lastfm-nature-vs-nurture-in.html]

[5] George Lawton, "How Pandora built a better recommendation engine," August 2017, [http://www.theserverside.com/feature/How-Pandora-built-a-better-recommendation-engine]

[6] Recommendation Technology 'Disco', [https://yandex.com/company/technologies/disco/]

[7] Florian Eyben, Real-time Speech and Music Classification by Large Audio Feature Space Extraction. Springer, 2016.

[8] Justin Salamon, "Tonal Representations for Music Retrieval: From Version Identification to Query-by-Humming," International Journal of Multimedia Information Retrieval, vol. 2, iss. 1, pp 45–58, March 2013.

[9] J., Salamon, and E. Gómez, "Melody Extraction from Polyphonic Music Signals using Pitch Contour Characteristics," IEEE Transactions on Audio, Speech and Language Processing, vol. 20, iss. 6, pp 1759–1770, 2012.

[10] E. Allamanche and B. Froba, "Content-based identification of audio material using mpeg-7 low level description," In in Proc. of the Int. Symp. of Music Information Retrieval, pp 197–204, 2001.

[11] J. Wood and J. Dykes, "Spatially ordered treemaps," IEEE Transactions on Visualization and Computer Graphics, vol. 14(6), pp 1348–1355, 2008.

[12] J.-J. Aucouturier and F. Pachet, "Music similarity measures: What's the use?," In Proc. Int. Conf. Music Information Retrieval (ISMIR), Paris, pp. 157-163, 2002

[13] B. Logan and A. Salomon, "A music similarity function based on signal analysis," In Multimedia and Expo, 2001. ICME 2001. IEEE International Conference on, pp. 745–748, 2001.

[14] Christopher D. Manning, Prabhakar Raghavan and Hinrich Schütze, Introduction to Information Retrieval. Cambridge University Press. 2008.

[15] Anssi Klapuri, "Introduction to Music Transcription," in Signal Processing Methods for Music Transcription, edited by Anssi Klapuri and Manuel Davy, New York: Springer, 2006, pp. 1–20. ISBN 978-0-387-30667-4.

[16] Stanley Smith Stevens, John Volkman and Edwin Newman, "A scale for the measurement of the psychological magnitude pitch". Journal of the Acoustical Society of America, vol. 8 (3), pp 185–190, 1937.

[17] Douglas O'Shaughnessy, Speech communication: human and machine. Addison-Wesley, 1987. ISBN 978-0-201-16520-3.

[18] W. Dixon Ward, "Musical Perception," In Jerry V. Tobias. Foundations of Modern Auditory Theory. 1. Academic Press. 1970, pp. 405-447

Areal Multistart Method of Optimization for Image Recognition

Galyna Shcherbakova
Odessa National Polytechnic University
Odessa, Ukraine
Galina_onpu@ukr.net

Svitlana Antoshchuk
Odessa National Polytechnic University
Odessa, Ukraine
asgonpu@gmail.com

Anatoly Sachenko
Ternopil National Economic University
Ternopil, Ukraine
sachenkoa@yahoo.com

Maksym Gerganov
Odessa National Polytechnic University
Odessa, Ukraine
max.at942@gmail.com

Marina Polyakova
Odessa National Polytechnic University
Odessa, Ukraine
marina_polyakova@rambler.ru

Victor Krylov
Ternopil National Economic University
Ternopil, Ukraine
viktor.krylov@gmail.com

Abstract—The multistart method of optimization for image recognition tasks has been developed. This method allows receiving an optimal solution in the form of the convergent sequence of areas to area pragmatic (semantic) sufficiency (APSS). The application of the method in the framework of the information-statistical approach will allow the formation of information technologies in visual information processing systems with the possibility of changing the error and speed of procedures based on optimization when the conditions for obtaining visual information change.

Keywords—component; formatting; style; styling; insert (key words)

I. INTRODUCTION

Modern methods of processing and recognizing signals and images widely use methods of optimization and classification in automated systems (AS) for visual information processing (ASVIP) [1-3]. In this task generally goal function is basis of information of which are characterized by multidimensionality, redundancy, a high level of uncertainty. The efficiency and quality of functioning the AS depends on many factors: resolution capability, the angular size and contrast of the image, etc. ASVIP in case of recognition of objects face execution of contradictory requirements: need of recognition of objects, that are invariant to projective conversions (scale, shift and turn) on images with noise, i.e. with distortions of the form, and the requirement of carrying out process of recognition in real time. It determines the level of noise immunity and high-speed performance as the AS required for tasks of processing in general, and its separate procedures.

With need of an acceleration of development process and extension of a scope of already developed systems in the conditions of fast enhancement of production technologies, change of conditions of observation etc., the important direction in development such ASVIP is creation of the systems capable to change the parameters. By development of ASVIP in general and its separate modules optimization methods are widely applied. As it was shown in lots of works, the purpose of optimization is hit of figures of quality and efficiency of ASVIP and its separate modules in an area pragmatic (semantic) sufficiency (APSS), that is receiving a "soft" optimal solution. Such decisions significantly increase flexibility of ASVIP, that is, allow to change system parameters in case of change of properties of a class of processed images without loss of efficiency.

However, the existing optimization methods are mainly intended for obtaining "point" optimal solutions.

In work [6] multi-start methods of "point" optimization are developed, in which wavelet transforms are used to estimate the direction of search. The use of wavelet transformation allowed to increase probability of finding of a global extremum, noise immunity, to reduce sensitivity to starting point of search and local extremums, to increase convergence speed in case of target functions like "ravine".

However, the properties of the wavelet transform (the variable sizes of a window of processing, the adjustable detail of the result) can also allow us to obtain the optimal solution in the form of a convergent sequence of regions (areas) whose parameter is the scale level serving as criterion of break, that is, go to soft optimization based on the wavelet transform. This will allow a "soft" optimal solution in the form of a convergent sequence of areas to the APSS ASVIP.

The purpose of the work is development and a research of an "areal" method of determination of an optimal solution in the form of a convergent sequence of areas to the area pragmatic (semantic) sufficiency ASVIP, on the basis of wavelet transform.

II. RESEARCH OF AREAL METHOD ON DETERMINATION OF AN OPTIMAL SOLUTION

As a rule, in practice iterative methods of optimization are used for determination of coordinates and/or the ranges of coordinates of an extremum. Basis of these methods – assessment of a gradient or subgradient of a target function. In the specified applications target function by optimization is often multimodal, can have a rough, piecewise linear, noisy surface. The optimization methods based on gradient assessment in such cases differ in low noise immunity, and at subgradient methods – a high error. For the decision of such tasks, with reduction of influence of the listed shortcomings, authors offered a multi-start method of optimization on the basis of the wavelet-transform (WT). The areal method is based on multi-start assessment of the direction of search of the APSS and also assessment of the sizes of areas forming the sequence converging to the APSS. A hit in the APSS is the criterion of break of areal method. Assessment is executed with use of the wavelet function (WF) of Haar which carrier depends on the calculated scale level. The analysis of other methods processing by means of Haar WF revealed their advantages and shortcomings: in particular

increase in noise immunity, growth of an error in case of an asymmetrical functional of quality. Because of the multistage assessment of the direction of the search for an extremum by means of WF, the implementation of these methods can require considerable computational costs. This reduces high-speed performance of methods and restricts the field of their application.

However for a row of applications high-speed performance of optimization shall be increased. It is possible to reach it, having defined APSS with the required error as optimization on the basis of WT allows localizing search a span by determination of the sequence of areas on processing stages of WF of Haar. The areal method is implemented according to the iterative diagram:

$$\mathbf{c}[n] = \mathbf{c}[n-1] - \gamma[n]WT(\mathbf{Q}(\mathbf{x}[n], \mathbf{c}[n-1])). \quad (1)$$

Suppose, that exist $j, n \in [1; N]$, j is performed before iteration n , where

$\text{sign}(WT(\mathbf{Q}(\mathbf{x}[n], \mathbf{c}[j]))) = \text{sign}(WT(\mathbf{Q}(\mathbf{x}[n], \mathbf{c}[n-1])))$
and
 $\text{sign}(WT(\mathbf{Q}(\mathbf{x}[n], \mathbf{c}[j]))) \neq \text{sign}(WT(\mathbf{Q}(\mathbf{x}[n], \mathbf{c}[n])))$ for
 $j \leq n-1, j < n$, then $\mathbf{c}[j] = \mathbf{c}[n-1]$ and check
 $\|\mathbf{c}[n] - \mathbf{c}[j]\| \leq D_{pr}$

if true formed new $A_k[p-1, p] = \{c_{p-1}[k], c_p[k]\}$, where
 $c_{p-1}[k] = \mathbf{c}[j], c_p[k] = \mathbf{c}[n]$, $p=1$ for $k \in 0, \dots, K$ the
end of the search.

else
 $\text{sign}(WT(\mathbf{Q}(\mathbf{x}[n], \mathbf{c}[j]))) = \text{sign}(WT(\mathbf{Q}(\mathbf{x}[n], \mathbf{c}[n-1])))$
and

$\text{sign}(WT(\mathbf{Q}(\mathbf{x}[n], \mathbf{c}[j]))) = \text{sign}(WT(\mathbf{Q}(\mathbf{x}[n], \mathbf{c}[n])))$
then $j = n$, $\mathbf{c}[j] = \mathbf{c}[n]$ and $c_p[k] = \mathbf{c}[j]$ or
 $c_{p-1}[k] = \mathbf{c}[j]$ depends on a side coming, then $n = n+1$.
This refines a value of border of an old areal $A_k[p-1, p]$.

Where the sign of (1) can be changed depending on a search (gradient or antigradient); the assessment norm $\|\mathbf{c}[n] - \mathbf{c}[j]\| \leq D_{pr}$ is a hit condition in the APSS; D_{pr} – the size of area APSS, $\mathbf{Q}(\mathbf{x}, \mathbf{c})$ – the functional, which depends on the vector of coefficients $\mathbf{c} = (c_1, \dots, c_N)$ and from $\mathbf{x} = (x_1, \dots, x_R)$; $\gamma[n]$ – the step; n – the iteration number, N – the number of iterations; $A_k[p-1, p]$ – the convergent sequence of the values forming areas; p – the address maker of the boundaries of areal in the sequence of areas converging to the APSS; k – the areal start number (the number of area), K – the number of area starts, j – the iteration number under the given conditions, defining appearance of areas, where $j, n \in [1; N]$ and j is performed before iteration n , WT defines the direction of movement to the APSS and it is calculated according to:

$$WT(\mathbf{Q}(\mathbf{x}[n], \mathbf{c}[n-1])) = \{G_{1j}, G_{2j}, \dots, G_{Rj}\}. \quad (2)$$

G_{lj} – the result of processing on the l -th variable:
 $l \in 1, \dots, R$ – the dimension of the parameter vector at the
 j -th iteration, (the description of the areal method is given
for the R -dimensional case);

Suppose, that $j \in n$, that for each j -th iteration find
 $\mathbf{G}[j]$, where

$$\mathbf{G}[j] = \frac{1}{s_j} \sum_{\substack{i=-\frac{s_j}{2} \\ i \neq 0}}^{\frac{s_j}{2}} \mathbf{Q}(\mathbf{x}[n], c_j + ia) \cdot \Psi_i(j). \quad (3)$$

Then if $\text{sign } \mathbf{G}[j] > 0$ it means that the amount of the elements of the convolution including the function value $\Psi_i(j) = 1$ more than with $\Psi_i(j) = -1$, what specifies on a spatial provision $\mathbf{c}[j]$ relatively the APSS (in a case with a test function Waves [5], by search of a gradient, – at the left) and therefore, the direction of movement, taking into account the situation, towards approximation to the APSS. Also under a condition $\text{sign } \mathbf{G}[j-1] = \text{sign } \mathbf{G}[j]$, $s_j = s_{j-1}$, remain in the area of the previous areal start with a number $k \leq j$. Then if $\text{sign } \mathbf{G}[j-1] \neq \text{sign } \mathbf{G}[j]$, a new areal start is started with the number $k = k+1$ and length of the carrier $s_j = s_{j+1}$, where the constraints of the second kind are imposed on the computation s_{j+1} and $s_j \gg s_{j+1}$.

But if $\text{sign } \mathbf{G}[j] < 0$ it means that the amount of the elements of the convolution including the function value $\Psi_i(j) = -1$ more than with $\Psi_i(j) = 1$, what specifies on a spatial provision $\mathbf{c}[j]$ relatively the APSS (in a case with a test function Waves, by search of a gradient, – at the right) and therefore, the direction of movement, taking into account the situation, towards approximation to the APSS. Also under a condition $\text{sign } \mathbf{G}[j-1] = \text{sign } \mathbf{G}[j]$, $s_j = s_{j-1}$, remain in the area of the previous areal start with a number $k \leq j$. Then if $\text{sign } \mathbf{G}[j-1] \neq \text{sign } \mathbf{G}[j]$, a new areal start is started with the number $k = k+1$ and length of the carrier $s_j = s_{j+1}$, where the constraints of the second kind are imposed on the computation s_{j+1} and $s_j \gg s_{j+1}$.

Therefore, $\mathbf{G}[n]$ a sequence defining the spatial position of $\mathbf{c}[j]$ with respect to the APSS and the direction of movement towards approach to the APSS at each n iteration for R variables. s_j – the carrier length WF at the j -th iteration k -th areal start (s_j – an even number); a – the sampling step WF; $\Psi_i(j)$ – WF at j -th iteration; k – the areal start number in case of execution the conditional restrictions for j -th iteration.

In case of the method research, Haar WF was selected for assessment of the direction of search of the APSS, also in this case, taking into account persistence $\Psi_i(j)$

further $\Psi(j)$. Determination of the sequence of areas coming to the APSS is made on the basis of the areal method with the following basic data: D_{pr} – the APSS size is defined at a stage of prior researches of the functional of quality; the length of carrier s_j of the WF and initial approximation to the APSS are defined in the analysis of target function, taking into account: a type of the researched function, it's value of noises and the restrictions of the 2nd kind for the sequence of areas $A_k[p-1,p]$ converging to the APSS.

TABLE I. WF PARAMETERS FOR AREAL METHOD FOR THE TEST WAVE FUNCTION [5].

USED NAMES	VALUE OF THE APSS SIZE $D_{pr} = 0.08$			
areal number start k	0	1	2	The sequence of areas converging to APSS: $A_0[-2;2], \sqrt{(-2)^2 + 2^2} > 0.08,$ $A_1[-0.44;0.03],$ $\sqrt{(-0.44)^2 + 0.03^2} > 0.08,$ $A_2[-0.07;0.03],$ $\sqrt{(-0.07)^2 + 0.03^2} < 0.08.$
carrier length S_k	180	90	18	

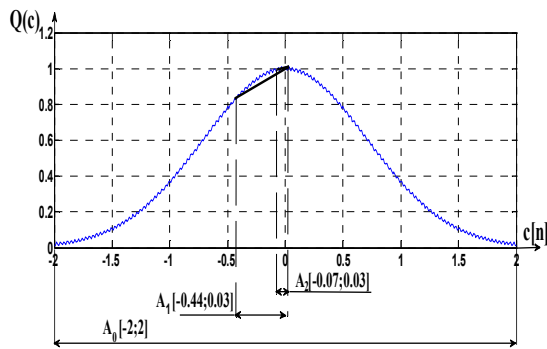


Fig. 1. Hit in the APSS by the areal method for the test Wave function [5].

Step 1. Set: $\mathbf{c}[1] = -0.97$ – the initial approach to the APSS; initial areal $A_0[-2,2]$ with number $k=0$ for making of the sequence of areas $A_k[p-1,p]$ converging to the APSS; $\gamma[1]=1$ – the step; a type of WT and WF; $a=0.01$ – the sampling step; s_0 – the WF carrier length (see tab. 1); $p=1$ – the address maker of the boundaries of areal in the sequence of areas converging to the APSS, the number of iteration $n=1$, the value of the APSS size $D_{pr}=0.08$.

Step 2. It is estimated on (3) a choice of the direction of search for achievement of the APSS. Estimated by (2) the position of the approach point relative to the APSS (in the point $\mathbf{c}[1]$ in case of $n=1$). The length of carrier s_1 of the WF (see tab. 1) is defined in an analysis of the target

function. The integral character of such WT allows to reduce sensitivity to local extremums and to select a segment of target function, where, as showed researches, with high probability there is APSS. (This step repeats for each n).

Step 3. The following $\mathbf{c}[n]$ point is calculated according to (1), where $n=n+1$, for receiving k -th of an areal for $n=2$, $\mathbf{c}[2] = -0.44$. (This step repeats for each n).

Step 4. If from (2) on iteration of n is execution of a condition $\text{sign } \mathbf{G}[n-1] \neq \text{sign } \mathbf{G}[n]$ where $n \neq 1$ in this case $k=k+1$ – the new transition through the APSS. In this case of change of the sign, based on the known property of estimates of the direction of search on the basis of a gradient; the sign will change upon transition through an optimum [4]. Thus, the elements of the convergent sequence: $A_k[p-1] = \mathbf{c}[n-1]$, $A_k[p] = \mathbf{c}[n]$, where $\mathbf{c}[j] = \mathbf{c}[n-1]$, $j=n-1$, consequently, for $n=3$ $j=1$, $p=1$, $A_1[1] = \mathbf{c}[2] = -0.44$, $A_1[2] = \mathbf{c}[3] = 0.03$ the area with number $k=1$ is found (same for $k=2$), transition to step 5; otherwise to step 7.

Step 5. The condition of achievement of the APSS $D_{pr}=0.08$ is checked, the result of hit in the APSS is estimated $\|\mathbf{c}[n] - \mathbf{c}[j]\| \leq D_{pr}$. This assessment is also a condition of break (end) of the search for the areal method. If the APSS is not reached go to step 6.

Step 6. It is made changes of length s_j of the WF carrier (see tab. 1), also taking into account the superimposed restrictions of the 2nd kind, go to step 2.

Step 7. At this stage taking into account $\text{sign } \mathbf{G}[n] = \text{sign } \mathbf{G}[n-1]$ specify boundary of a required area from earlier defined approximation side $\mathbf{c}[j] = \mathbf{c}[n]$ where, for $n=3$, $j=n$, $\mathbf{c}[3] = 0.03$ next go to step 2. (This step is repeated for each n).

Thus, in the process of searching by means of WF Haar, the convergent sequence is formed of a row of nested areas (see fig. 1). The initial, from these areas, there is the given area (see tab. 1) for optimization of the function $\mathbf{Q}(\mathbf{c})$

where $s_k = 180$ – length of the WF carrier of the initial start $\Psi(1)$. Also at all stages of computation, length of the WF carrier $\Psi(j)$, for each k -th areal start, changes taking into account: type of the WF $\Psi_i(j)$, a type of the researched function $\mathbf{Q}(\mathbf{c})$, the amount of interference imposed on the function, as well as the imposed constraints of the second kind on a convergent sequence of areas $A_k[p-1,p]$.

The basis of the areal method is the immediacy of the soft achievement of the APSS, where the stopping criterion at the stage of each next approximation is the entry into the APSS. However also because of lowering of the length of the carrier $s_{k-1} \gg s_k$ for the WF $\Psi(j)$ in case of approximation to APSS, at a stage of determination of the following area, noise immunity, in comparison with the initial areal start, decreases. This, under certain conditions, can make it difficult the hitting into the APSS the given size.

Features of the procedure of achievement of APSS were checked experimentally on test function of DeJong $f(x) = x^2$ (for $x \in (-5; 5)$) with adding of the noise distributed under the Gaussian distribution with a zero average and mean squared deviation from 0 to 6, the maximum function value 25) and the Wave function $f(x) = e^{-x^2} + 0,01 \cdot \cos(200 \cdot x)$ for $x \in [-2; 2]$ (see fig. 1).

Procedure parameters: the step $\gamma = 1$, the initial length of the carrier WF $\Psi(1)$ $s_1 = 180$, the sampling step WF $a = 0.01$. The sequence of the converging areas to the APSS (see tab. 1).

III. SUMMARIES

Thus, the areal multi-start method of optimization for image recognition corresponds to the requirements of optimization tasks of clustering and classification.

It was found, that the areal method in case of hit in the APSS, showed good results of speed. For the test DeJong function the runtime [0.013473c; 0.090658c], for the test Wave function the runtime 0.064796c. Calculations are made on a PC with a configuration (CPU Intel Core 2 Dual P8400 2.26GHz, RAM DDR2 6400 2GB).

The results of the research of efficiency of the areal method showed, that for the test Schwefel function [5] the

area of the global extremum with probability 0.84 is reached, and for the DeJong function [5] with probability 0.92 is reached. For the DeJong test function [6] with a signal-to-noise ratio from 64 to 28, the result of the speed showed an increase of 17% compared to the previously proposed method on the wavelet transformation base [6]

These results allow recommending the areal multi-start method of optimization for image recognition tasks.

REFERENCES

- [1] Ya. Z. Tsyarkin, "Optimality in optimization problem and algorithms under uncertainty," "Avtomat. i Telemekh.", no. 1, pp. 75-80, 1986. (In Russian).
- [2] P. G. Katys "The Systems of machine vision with intellectual video sensors", Informacionnye tehnologii, no. 10, pp. 28 - 33, 2001. (In Russian).
- [3] M. Moganti, F. Ercal, C. Dagli, and S. Tsunekawa, "Automatic PCB inspection algorithms: a survey," Computer vision and image understanding, vol. 63, no. 2, pp. 287-313, 1996.
- [4] E. Polak, Computational Methods in Optimization. A Unified Approach. London: Academic Press, 1971.
- [5] A. B. Sergienko, Test functions for global optimization v.1.32 Krasnoyarsk: Reshetnev Siberian State University Publisher, 2015. (In Russian).
- [6] G. Yu. Scherbakova, V. N. Krylov, and O. Yu. Babylonha, "Research of the extremum region using a multi-start optimization method based on the wavelet transform", Electrotechnical and Computer Systems, no. 18 (94), pp. 86-91, 2015. (In Russian).

Development of Real-time Face Recognition System Using Local Binary Patterns

Maksym Kovalchuk

*Department of Information Computer Systems and Control
Ternopil National Economic University
Ternopil, Ukraine
kow.max7@gmail.com*

Vasyl Koval

*Department of Information Computer Systems and Control
Ternopil National Economic University
Ternopil, Ukraine
vko@tneu.edu.ua*

Anatoliy Sachenko^{1,2}

¹*Department of Informatics
Kazimierz Pulaski University of Technology and Humanities in
Radom, Poland*
²*Research Institute for Intelligent Computer Systems
Ternopil National Economic University
sachenkoa@yahoo.com*

Diana Zahorodnia

*Research Institute for Intelligent Computer Systems Ternopil
National Economic University
Ternopil, Ukraine
dza@tneu.edu.ua*

Abstract— This paper describes the development of real-time human recognition system in video streams with the help of Local Binary Patterns (LBPs). The description of the system architecture, face detection process, additional methods for recognition accuracy increase and the method of image comparison based on center-symmetric LBPs are given.

Keywords— *video analytics; face detection; face recognition;*

I. INTRODUCTION

Video analytics is a widely used technology that uses computer vision techniques to collect various information based on the sequence of frames received from video cameras online or from by using optic flow. This technology can be used in video surveillance, security systems, restricted access systems, pay systems, criminal identification, etc [1]. One of the tasks solved by video analytics is the recognition of faces in video streams. The solution of this problem primarily has a direct application in access control and identification systems [1].

Face recognition is a difficult task to implement due to the variable conditions for visualizing the face such as lighting, the position of the head in relation to the camera, facial expressions and other factors [2-4]. When designing systems, developers try to avoid the negative impact of these factors by imposing severe limitations on the process of acquiring images of the individuals, but the most practical problem is the problem of recognizing faces fast enough to process them with high rate of correct recognitions [2]. In recent years, there has been significant progress in this area, largely due to the improvement of the hardware and computer vision libraries [2].

There are several methods for face detection and face recognition [2-4], which were considered as possible options for implementation.

A. Methods of Face Detection:

- The Viola-Jones Method;
- AdaBoost (Adaptive Boosting);
- Support Vector Machine;

- Convolutional Neural Network Methods (YOLO);
- SNoW (Sparse Network of Winnows).

Each of the methods has its main strong points and drawbacks [5]. The Viola-Jones Method: high detection rate due to the use of the cascading classifier, but imposes restrictions on the position of the face upon detection [5]. AdaBoost: high speed of work, but is sensitive to noise and data outliers [5]. SNoW: high speed of work due to sifting the components of the vector of signs, sensitive to noise [5]. Neural Networks: computational complexity and sensitivity, but high detection accuracy with proper network settings [5].

The Viola-Jones method was selected as a method of choice as it includes an improved variation of AdaBoost, has very high detection speed and high detection accuracy of faces in particular compared to the competitors [5], significantly reduces the computation and is famous for face detection with very low false positive rate [5].

B. Methods of Face Recognition

- Methods based on pixel brightness values
 - a) *Eigenfaces* ;
 - b) *Fisherfaces*;
 - c) *Local Binary Patterns* .
- Methods based on characteristic points
 - a) *Feature-based (structural) methods*.
- Methods based on neural networks.

Although neural networks may be preferable in terms of maximum accuracy the use of neural networks does not justify itself in projects where a 3-5% percent difference in recognition accuracy is not critical [6], the processing capabilities of the hardware is limited, and a trade-off between speed and accuracy leans toward speed [6,7]. Each method is designed to solve a specific problem and has its own peculiarities when working under different conditions [6]. In this regard, it is necessary to select such components in the developed system, which will be more universal and will show good results in various tests for identifying and

recognizing faces under conditions most often encountered in practice. Local Binary Patterns method is a method of choice considering its high-speed computation capabilities, rotation invariance, effectiveness in image segmentation, tolerant against monotonic illumination changes, conservativeness with resources and ability to recognize a moving object via background subtraction compared to the functionality of other methods [7-9]. A goal of this paper is to create the face recognition system with the capability of real-time processing using a modification of Local Binary Patterns that would improve the accuracy of face recognition compared to the use of classical LBP, as well as traditional methods of face recognition.

II. SYSTEM ARCHITECTURE

The development of the technical system for processing of the frames from video stream includes two main stages. For the solution of the problem, the following architecture is proposed:

- Detection of faces by the Viola-Jones method;
- Recognition of the detected faces with the help of histograms of local binary templates and the method of the nearest neighbor.

In addition to the detection and recognition stages (see Fig. 1), it is reasonable to use the following additional intermediate stages of processing:

- Processing of detected faces by the Gaussian filter;
- Applying a mask of significant areas;
- Calculating the histograms of the detected faces.

As a result, the program obtains a list of monitored individuals with their characteristics and coordinates of the rectangular areas of the frame in which they are located.

Gaussian filter eliminates noise. A Gaussian filter is an image blur filter that uses a normal distribution (also called a Gaussian distribution) to calculate the transformation applied to each pixel in the image [10].

Images of faces obtained after the detection procedure are square. However, the face does not occupy the entire space of such an image. Therefore, it would be logical to exclude the influence on the solution of the classifier of image areas in which there is no face. [11]

The obtained local binary pattern (LBP) histograms are then classified by the nearest neighbor method. The method of the nearest neighbor is a classification algorithm, the essence of which is that the object belongs to the class to which it is closest to the element [12].

After applying the LBP operator, the image is divided into rectangular areas. For each histograms of such areas it calculates how often pixels of different brightness values occur in the given region. The resulting histograms are normalized, concatenated, and subsequently used as classification features [12]. There are several types of LBP operators: classical, uniform and center-symmetrical. Fig. 2 illustrates the LBP transformation.

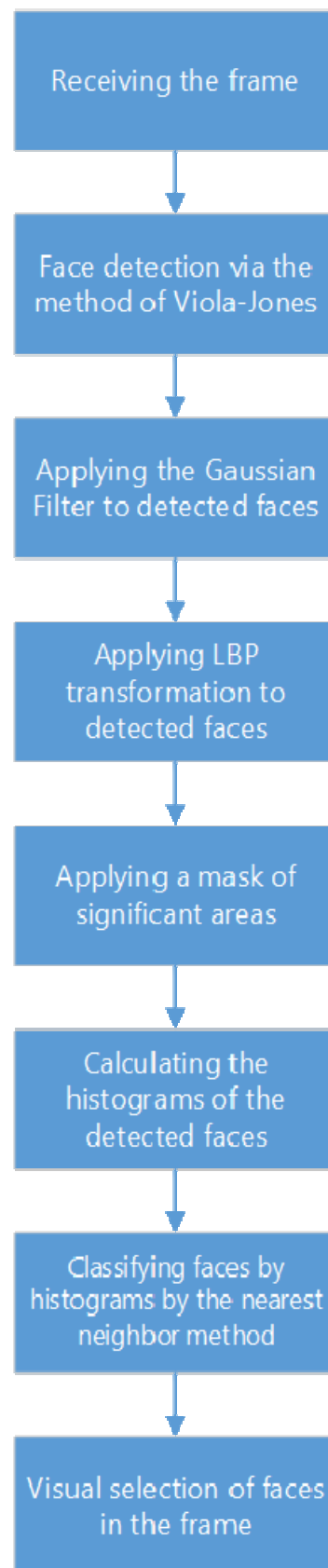


Fig. 1. Processing stages flowchart

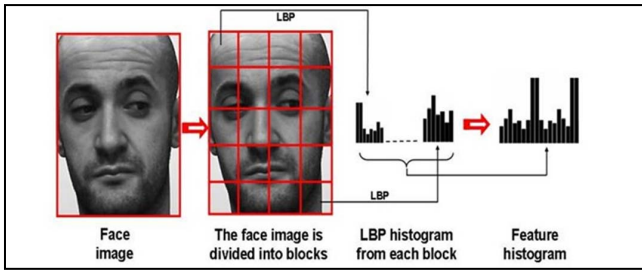


Fig. 2. Local Binary Pattern transformation

Center-symmetric Local Binary Pattern (CSLBP) is selected as a main method as it decreases the memory usage and reduces computational complexity. The essence of the modification is that the threshold value for each area pixel is not the brightness value of the central pixel of the area, but the brightness value of the opposite area of the pixel relative to the center [13]. The comparison of classical and CSLBP operators is shown in Fig. 3.

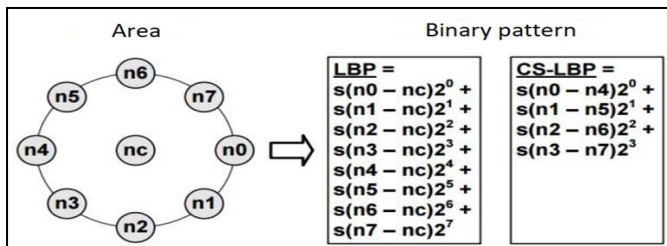


Fig. 3. Calculation of the value of conventional and center-symmetric LBP (CSLBP)

As a result it is possible to derive some specific properties of the considered approach:

A. Limitations:

- Only monotonous change of illumination is admissible;
- The training and test sample should be taken under the same or similar lighting conditions;
- Frontal or close to frontal face positions must be used;
- Neutral expression of faces in images;
- Individuals must not overlap with other objects significantly.

B. Functionality:

- Processing of a video stream with the camera connected to the computer in real-time;
- Ability to customize the work parameters used to detect and recognize algorithms;
- Outputting the information about the identifiable person, including the measure of association with a certain class, graphic display of the histogram and LBP representation of the face;
- The ability to learn and add classes of people using the camera through the application interface (only one person in the camera view) or via photo/video upload.

III. SYSTEM IMPLEMENTATION

The system grounds on the object-oriented programming language C # in the development environment of Microsoft Visual Studio 2013. The presence of the "garbage collection" mechanism in C # effectively organizes the work with the list structures [14]. The project uses OpenCV library. This library is developed in C / C ++ and has interfaces for Python, Java and other languages, including the wrapper for .NET languages - EmguCV [15], which was used in the project. A large collection and articles of theoretical materials on the use of computer vision make the OpenCV library an ideal solution for use in projects devoted to solving problems in this field [15].

Fig. 4 illustrates only the associative links of classes without their contents.

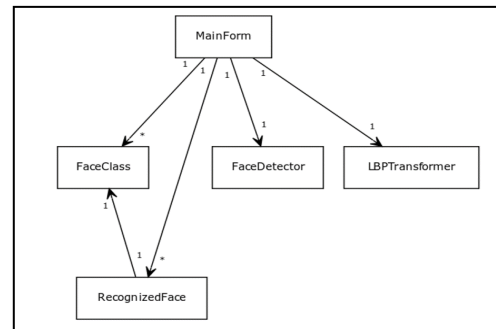


Fig. 4. UML class diagram of the application

RecognizedFace

The RecognizedFace class describes the face recognized by the classifier. Elements of this class are stored in the updated list of recognized faces, based on which data is displayed information on the form.

LBPTransformer

The LBPTransformer class describes a CSLBP converter. This class is a static class, has no attributes, and contains only one method that performs a centrally symmetric LBP image transformation. It implies the addition of other LBP transformation methods with further development of the development.

FaceDetector

The FaceDetector class describes a face detector using the Viola-Jones method.

FaceClass

This class describes the class (category) of persons. Each class of persons corresponds to a particular person being recognized.

MainForm

The MainForm class describes the main application window and implements the main application module.

Fig. 5 illustrates part of a starting sequence of methods presented in MainWindow Class.

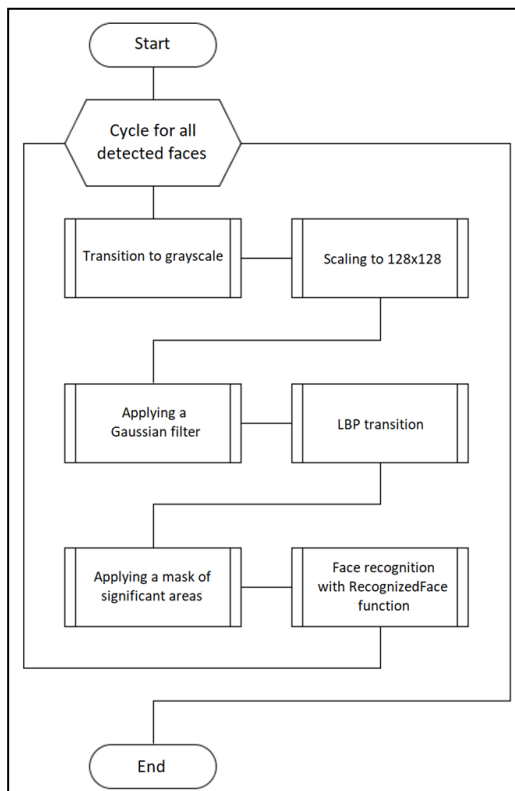


Fig. 5. Core method's unifying processing functions flowchart

The output area of the video stream is designed to display the processed frames. In addition, this area includes buttons for setting the source of the video stream, the button for opening the video file dialog box, and the button that controls the start and stop of the video stream capture. The output area of the information about the user selected is intended for displaying the data of faces recognized by the application. The source of the data to display is the face class selected by the user in the Face Class list and the recognized person corresponding to that class. In this form area, the image of the selected class is displayed, the rectangular area of the current frame corresponding to the person of this class, and also the LBP transformed face image. In addition, the current distance between the histogram of the recognized face and the histogram of its class and the graphical representation of the LBP histogram of the face image are displayed. The application settings area includes two groups of settings. The first group is the settings for face recognition. These include the Gaussian blur level and the recognition threshold. The recognition threshold is the maximum distance between the LBP histogram of the face and the histogram of the class. The second group of settings in this area contains the settings for the face detector. Fig. 6 illustrates the interface and the working process of the developed program.

IV. EXPERIMENTAL STUDIES

For performance, the center-symmetric local binary patterns (CSLBPs) are the ideal choice for classification criteria in a real-time recognition system [15]. Nevertheless, it makes sense to make sure that it does not lose much to other LBP variations [16] in the accuracy of the classification. To solve this problem, three algorithms for calculating LBP histograms were tested on two different datasets. At the same time, the best way of dividing the

images into local areas was also evaluated. Created in the process of research, the implementation of LBP transformations and the method of the nearest neighbor were used in the final version of the program, as well as in testing the speed of this system using various variations of LBP conversion.

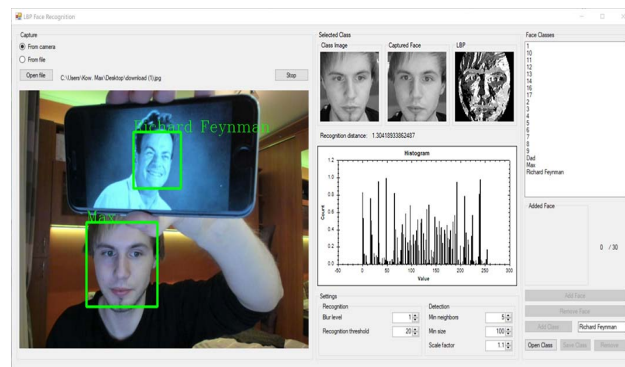


Fig. 6. Program interface

A. Testing the speed of Local Binary Patterns operators

The first dataset used for testing is the imaging database of the Cambridge University laboratory [17]. It contains images of 40 people, 10 images each. Lighting on these images does not change, but there are variations in the position of the face when shooting. Fig. 7 illustrates images of one of the faces of this database. As a training and test sample, 5 images per face were used. The test and training sample did not have any intersections, however, to obtain results on a larger data set, the test and training samples were swapped and the testing was repeated. Before processing, the images are scaled to 128x128 pixels. As a result, 400 images were classified. The second set of data is the image database of the Yale University laboratory [18]. This database contains images of 38 people, 65 images each, including various variations of lighting. Of these, 10 images per person were selected for testing. An example of images from the second database is in Fig. 7.



Fig. 7. Images from the database of Cambridge University (a) and Yale University (b)

The recognition efficiency for each of the three LBP operators when testing on the data sets is presented in Table 1 and Table 2.

TABLE I. FIRST DATASET

Blocks' division \ Method	1x1	2x2	3x3	4x4	5x5	6x6	7x7	8x8
LBP	82,5%	91%	94%	95,5%	94%	93,25%	92,5%	89,3%
Uniform LBP	81%	93,8%	97%	94,5%	92%	92%	92%	89,5%
CS-LBP	67,8%	92,3%	95%	94,8%	94,3%	93,3%	90,3%	90,3%

TABLE II. SECOND DATASET

Method	Blocks' division							
	1x1	2x2	3x3	4x4	5x5	6x6	7x7	8x8
LBP	41.8%	71.6%	88.2%	91.8%	92.6%	93.2%	95.8%	96.1%
Uniform LBP	41.3%	75.26%	91.8%	91.8%	92.9%	92.1%	95%	94.2%
CS-LBP	20.3%	61.3%	84.5%	89.2%	89.2%	91.6%	92.6%	93.9%

CSLBP in a 4x4 partition, which is used in the developed system, is optimal for the ratio of accuracy and cost of memory when partitioning the image into sub-regions. This partition provides a consistently high percentage of correct classifications with not very large memory costs as shown in Table 3.

When the image is divided into a small number of blocks, the center-symmetric LBP is inferior to other local binary patterns. However, when using a larger number of sub-regions, its classification accuracy index lags behind other LBPs by an average of no more than 3%. When testing the same on the first set of data, CSLBP completely outperforms other LBPs on some partitions.

TABLE III. DATASETS' AVERAGE ACCURACY

Method	Blocks' division							
	1x1	2x2	3x3	4x4	5x5	6x6	7x7	8x8
LBP	41.8%	71.6%	88.2%	91.8%	92.6%	93.2%	95.8%	96.1%
Uniform LBP	41.3%	75.26%	91.8%	91.8%	92.9%	92.1%	95%	94.2%
CS-LBP	20.3%	61.3%	84.5%	89.2%	89.2%	91.6%	92.6%	93.9%

B. Testing the speed of Local Binary Patterns operators

The study was carried out for measuring various LBP operators with a different number of histograms of individuals in the database. The processing speed of the video stream was estimated by the number of frames processed per second when one person was recognized in the frame. As was expected, CSLBP operator demonstrated significantly better results, as can be seen in Table 4 (Hardware specifications: Intel Core i5 3.2 GHz, 8Gb, video resolution 640x480).

TABLE IV. THE SPEED OF LOCAL BINARY PATTERNS OPERATORS

Method	Number of histograms									
	1000	2000	3000	4000	5000	6000	7000	8000	9000	10000
LBP	23	18	15	13	12	10	9	8	8	7
Uniform LBP	30	30	30	29	28	25	23	21	20	19
CS-LBP	30	30	30	30	30	30	30	30	30	30

A comparison between current state-of-the-art face detection open source solutions such as YOLO (You Only Look Once) [19], OpenFace [20] and Fisherface (OpenCV) [21] was conducted under the same testing conditions as LBP variations performance evaluation. For the testing, Yale and Extended Yale databases were used. In its area of application and functionality boundaries, CSLBP showed itself to be on par with current neural network solutions and outperformed some of the solutions with results illustrated in Fig. 8 and Fig. 9. However, it is important to keep in

mind that its advantageous position is dictated mostly by specifics of test limitations such as front face positioning of the image and monotonous change of illumination.

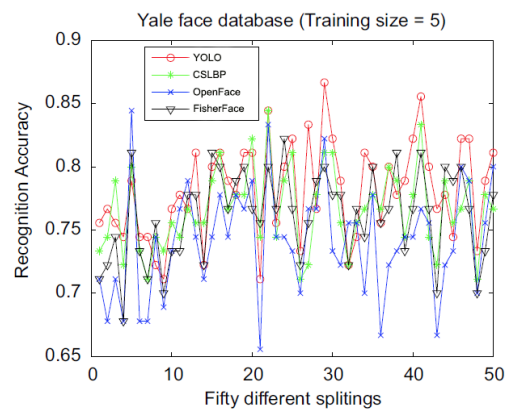


Fig. 8. Yale face database results

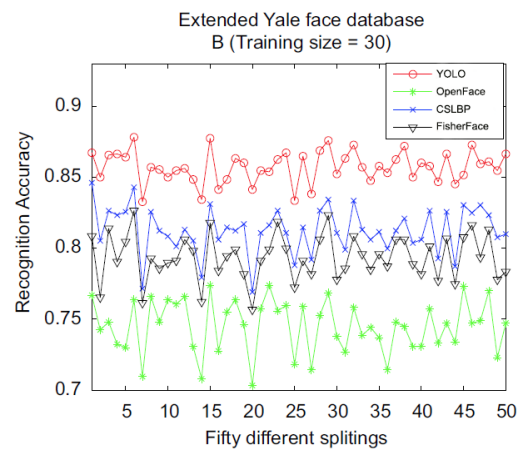


Fig. 9. Extended Yale face database results

V. CONCLUSION

The testing of the developed system confirmed results in approximately 90% of correct face recognition. The developed system can be used for solving various tasks of video analytics, and, in the first place, it can be applied and integrated in access control and identification systems (e.g. e-borders, governmental citizen checklists) considering its functionality, recognition performance, memory management and ease of use. Further improving of the system may be achieved by changing the faces classifier. It is possible to use more sophisticated classification, for example, the list of methods known as Random Forests, method of reference vectors or neural networks algorithms. It is also worth noting that the architecture of the developed application makes it easy to replace individual modules, which opens up great opportunities for further development and improvement of the system.

REFERENCES

- [1] S. Wang and W. Kelly, "Video-based Big Data Analytics in Cyberlearning," *Journal of Learning Analytics*, vol. 4, no. 2, pp. 36-46, 2017.
- [2] A. Barman and P. Dutta, "Facial expression recognition using distance and shape signature features," *Pattern Recognition Letters*, 2017.

- [3] Y. Kurylyak, I. Paliy, A. Sachenko, and V. Koval, "Improved Method of Face Detection Using Color Images," International Conference "Modern Problems of Radio Engineering, Telecommunications and Computer Science" TCSET'2006, Lviv-Slavske, Ukraine, pp. 186-188, Feb 28- Mar 4, 2006.
- [4] A. Sachenko, V. Koval, I. Paliy, and Y. Kurylyak, "Approach to Face Recognition Using Neural Networks," IEEE Third International Workshop on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications IDAACS'2005, Sofia, Bulgaria, 2005, pp. 112-115, September 5-7.
- [5] M.K Dabhi, and B.K Pancholi, "Face Detection System Based on Viola - Jones Algorithm," International Journal of Science and Research (IJSR), vol. 5, no. 4, pp. 62-64, 2016.
- [6] S. Khan, "Factors Affecting the Recognition Accuracy of Facial Expressions", MOJ Applied Bionics and Biomechanics, vol. 1, no. 4, pp. 1-2, 2017.
- [7] D. Zahorodnia, Yu. Pigovsky, P. Bykovyy, V. Krylov, B. Rusyn, and V. Koval, "Criteria to Estimate Quality of Methods Selecting Contour Inflection Points," 9th IEEE International Conference on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications (IDAACS'2017). Bucharest (Romania), pp. 969-973, September 21-23, 2017.
- [8] D. Zahorodnia, Y. Pigovsky, and P. Bykovyy, „Canny-based method of image contour segmentation,” International Journal of Computing, 15(3), pp. 200-205, 2016.
- [9] M. Pietikäinen, "Local Binary Patterns," Scholarpedia, vol. 5, no. 3, p. 9775, 2010.
- [10] Deepak Raj, Poonam Singal, and Namika Kumari, "Reducing the Computation Complexity of 2-D Gaussian Filter", International Journal of Scientific Research and Management, vol. 5, pp. 5336-5340, 2017.
- [11] L. Jeffery and G. Rhodes, "Insights into the development of face recognition mechanisms revealed by face aftereffects", British Journal of Psychology, vol. 102, no. 4, pp. 799-815, 2011.
- [12] B. Yang and S. Chen, "A comparative study on local binary pattern (LBP) based face recognition: LBP histogram versus LBP image," Neurocomputing, vol. 120, pp. 365-379, 2013.
- [13] E. Zhang, Y. Li, and J. Duan, "Moving object detection based on confidence factor and CSLBP features," The Imaging Science Journal, vol. 64, no. 5, pp. 253-261, 2016. "Visual Studio IDE, Code Editor, VSTS, & App Center", Visual Studio, 2018. [Online]. Available: <https://www.visualstudio.com/>. [Accessed: 20- Apr- 2018].
- [14] "Emgu CV", SourceForge, 2018. [Online]. Available: <https://sourceforge.net/projects/emgucv/>. [Accessed: 20- Apr- 2018].
- [15] J. Shen, W. Yang, and C. Sun, "Real-time human detection based on gentle MILBoost with variable granularity HOG-CSLBP," Neural Computing and Applications, vol. 23, no. 7-8, pp. 1937-1948, 2012.
- [16] A. Abdesselam, "Local Similarities for Boosting the Performance of Local Binary Patterns Technique," International Journal of Signal Processing Systems, pp. 510-514, 2016.
- [17] "The Database of Faces", Cl.cam.ac.uk, 2018. [Online]. Available: <http://www.cl.cam.ac.uk/research/dtg/attarchive/facedatabase.html>. [Accessed: 20- Apr- 2018].
- [18] "Face Recognition Homepage - Databases", Face-rec.org, 2018. [Online]. Available: <http://www.face-rec.org/databases/>. [Accessed: 20- Apr- 2018].
- [19] J. Redmon, "YOLO: Real-Time Object Detection," Pjreddie.com, 2018. [Online]. Available: <https://pjreddie.com/darknet/yolo/>. [Accessed: 20- Apr- 2018].
- [20] "OpenFace", Cmusatyalab.github.io, 2018. [Online]. Available: <https://cmusatyalab.github.io/openface/>. [Accessed: 20- Apr- 2018].
- [21] "OpenCV: Face Recognition with OpenCV", Docs.opencv.org, 2018. [Online]. Available: https://docs.opencv.org/3.3.1/da/d60/tutorial_face_main.html. [Accessed: 20- Apr- 2018].

Panels

Panel #1

ANDY BOSYI

Founder and Chief Executive Officer, MindCraft.ai (Lviv, Ukraine)

Topic: "Decision-Making Systems Based on Time Series"

ELENA YEGOROVA

PhD, Chief Technical Officer, Co-founder, LMX (London, United Kingdom)

Topic: "Data Processing for Behaviour Pattern Extraction: Applied Problems"

OLEG NOVOSAD

Chief Executive Officer, Severenity Senior Mobile Engineer, SoftServe Lecturer, IT Step University (Lviv, Ukraine)

Topic: "Location-based procedural generation of augmented reality portals"

Panel #2

BOHDAN KOLCHYGIN

PhD, Machine Learning Engineer, AltexSoft (Kharkiv, Ukraine)

Topic: "Under Construction: How to Debug Machine Learning Process"

DENIS FRAGKAKIS

Ph.D., Lead Data Scientist, Nested; Head of Data Science, LMX (London, United Kingdom)

Topic: "Data Processing for Reinforcement Learning: Advantages and Future"

ARTEM NIKULCHENKO

Ph.D., Chief Software Architect, Teamwork Retail (Clearwater, FL, USA)

Topic: "Using Google Cloud Platform Tools for Big Data Processing and Analyze"

Panel #3

DMYTRO LIASKOVSKYI

Software Engineering Manager, Epam (Lviv, Ukraine)

Topic: "DLab - essential toolset for analytics"

OLEKSANDR SLIPCHENKO

PhD, Team Lead, Booking.com BV (Amsterdam, Netherlands); Data Analyst, Teragence (London, United Kingdom)

Topic: "Teragence - understand the performance and customer experience of a mobile network"

ALEX GOSTEV

Product Portfolio Manager, Diligences Inc. (Kharkiv, Ukraine)

Topic: "Source Code Metrics Quantification and Analysis Personalized"

VLAD VANZHA

Lead Engineer, Exponential Inc. (Kharkiv, Ukraine)

Topic: "Machine Learning Techniques in Advertisement"

Panel #4

VOLODYMYR LYUBINETS

Software Engineer, Forethought (San Francisco, United States)

Topic: "Automated Triaging of Bugs and Tickets using Attention-based Mechanisms in RNNs"

YEHOR LYEBYEDYEV

Data Scientist, Perfectial (Lviv, Ukraine)

Topic: "Analysis of Data Streams from Social Networks"

IEVGEN GOROVYI

PhD, Chief Executive Officer, It-Jim (Kharkiv, Ukraine)

Topic: "Custom augmented reality SDK: concepts, technical solutions and deployment to mobile platforms"

SEMEN OSKERKO

Student, IT STEP University (Lviv, Ukraine)

Topic: "Hybrid Multidimensional Wavelet-Neuro-System and its Learning in Pattern Recognition for IoT Application"

Panel #5

TETIANA GLADKYKH

Ph.D., Competence Manager, Data Science Group, SoftServe (Lviv, Ukraine)

Topic: "Music Content Selection Automation"

IGOR MANZHOS

Data Science Engineer, Consultant, GlobalLogic (Kharkiv, Ukraine)

Topic: "Using Deep Convolutional Neural Networks for Medical Diagnostics"

IGOR MISHCHENKO

Data Science Engineer, Senior, GlobalLogic (Kharkiv, Ukraine)

Topic: "Medical Data Normalization with Recurrent Neural Networks "

Panel #6

OLEG VOLOSHKO

Ph.D., Head of Big Data products, Kyivstar (Kyiv, Ukraine)

Topic: "Machine Learning Algorithms in Solving Telecom Operator Issues"

IRYNA ZAYTSEVA

Project Manager, uData School (Kyiv, Ukraine)

Topic: "Ukrainian education and data science business: are there any touch points? "

IGOR VLASOV

Ph.D., Compliance Manager, Fintech United Group (Kyiv, Ukraine)

Topic: "Blockchain & Society"

Author's Index

A

Aizenberg I., 315, 392
Al-Ammouri A., 468
Al-Ammouri H., 468
Ali Rekik, 107, 342
Aliexsieiev V., 32, 94
Ambach D., 212
Ambach O., 212
Andrushchenko V., 17
Antoshchuk S., 381, 605
Artykulna N., 50
Atamanyuk V., 375
Atkinson K., 25
Azarov O., 369

B

Babichev S., 336
Babii A., 524
Balagura I., 17
Basalkevych O., 60
Basalkevych O., 60
Basystiuk O., 478
Batryn N., 554
Batyuk A., 98, 151, 356
Beglytsia V., 70
Berezska K., 554
Berezsky O., 554
Berko A., 32
Beznosyk O., 407
Bhushan Sh., 381
Bidyuk P., 70
Bilozetsev I., 514
Bodyanskiy Ye., 3, 7, 113, 327, 473, 519
Boiko Olh., 503
Boiko Ol., 519
Boiko T., 271
Borovenskyi O., 503
Borzov Yu., 157, 187, 305
Boyko N., 478
Boyun V., 498
Brazhnykova Ye., 3
Briukhovetskyi O., 227
Bulakh V., 198
Burak N., 157
Burov Ye., 128

C

Charhad M., 107
Chervoniak Ye., 236
Chornous G., 397
Chuiko G., 119

Chumachenko D., 415
Chupryna A., 509
Churyumov G., 183
Chyrun L., 139
Chyrun Lu., 139
Číž D., 528

D

Darnapuk Ye., 119
Degtiarova A., 468
Deineko A., 7, 171
Demchuk A., 128
Deriuga I., 218
Didyk O., 187
Dolotov A., 327
Dolynyuk T., 554
Dominik A., 558
Dorosh V., 102
Doroshenko A., 231
Dosyn D., 145
Dumin O., 434
Durnyak B., 584
Dvornik O., 119
Dyvak M., 444

F

Farzad Movahedi Sobhani, 286
Fedoronchak T., 574
Filipenko O., 13
Fišer J., 411
Furgala Yu., 595

G

Garkavtsev M., 75
Geche F., 151, 356
Gerasin O., 44
Gerganov M., 605
Gladkykh T., 599
Glybovets M., 207
Golovko V., 102
Golovko V., 430
Golub S., 223
Gordon B., 171
Gorokhovatskyi O., 459, 464
Gorokhovatskyi V., 464
Gorovyi Ie., 236, 534
Gostev A., 75
Gozhyj A., 70
Grinberg G., 193
Grubnyk R., 599

H

Havrysh B., 584
Hnot T., 599
Holovatch Yu., 84, 241
Horpenko D., 56
Hu Zh., 402
Hurtik P., 528

I

Ignaciuk P., 424
Izonin I., 386

K

Kaláb O., 528
Kalinina I., 70
Kalnichenko O., 346
Kalychak Yu., 563
Kashifuddin Qazi, 315
Kashpruk N., 331
Kenna R., 241
Khairova N., 21
Khaliq Z., 392
Kharchenko K., 407
Khavalko V., 438
Khaya A., 166
Khlamov S., 227
Khrutba A., 346
Khymytsia N., 223
Kirichenko L., 198
Kis Ia., 139
Klachek P., 310
Kliuvak O., 483
Klochan A., 468
Klyujnyk I., 580
Klyuvak A., 483
Kočárek P., 528
Koman B., 134
Komar M., 102
Kondratenko N., 38
Kondratenko Yu., 38, 44
Kondratiuk S., 420
Kopaliani D., 519
Korjagin S., 310
Korobchynskyi M., 336, 494
Korobov A., 503
Kotsovsky V., 356
Koval V., 609
Kovalchuk A., 542
Kovalchuk M., 609
Kovylin Ye., 322
Kozina Yu., 56
Krak Iu., 420
Kravets P., 123
Kriukova G., 207
Krupelnitsky L., 369

Krylov V., 605
Kulishova N., 473
Kutucu H., 386
Kuznetsov A., 514
Kynash Yu., 162

L

LLamonova N., 75
Lande D., 17
Levkovych M., 375
Lewoniewski W., 21
Li G., 25
Lieberman I., 310
Litovchenko O., 3
Liubyma Iu., 346
Ljaskovska S., 157, 177
Lotoshynska N., 542
Lozynska O., 145
Lozynskyy A., 251
Lyebyedyev Ye., 276
Lysa N., 538
Lytvyn V., 128, 145
Lytvynenko V., 336, 411
Lyubchik L., 193
Lyubinets V., 271

M

Makhortykh M., 276
Maksymiv O., 455
Malets I., 177, 259, 558
Malets R., 259
Malyar M., 65
Malysheva D., 473
Manakova N., 166
Mariliv A., 494
Martsyshyn R., 538
Martyn Ye., 177
Martynenko S., 503
Mashkov V., 411
Mashtalir S., 545, 549
Mashtalir V., 545
Melnyk A., 444
Melnyk R., 563
Mesbaholdin Salami, 286
Mieshkov S., 494
Mikheev I., 202
Mikhnova O., 549
Minialo V., 251
Miyushkovych Y., 538
Mochulsky Yu., 595
Mohammad Sadegh Ghazizadeh, 286
Mokrytska O., 375
Morklianyk B., 455
Morozov V., 50, 346
Moskalenko A., 503
Moskalenko V., 503

Mryglod O., 241
Mulesa O., 151
Mulesa P., 3
Musiolek D., 528
Myronova N., 574

N

Nazarenko S., 79
Nazarkevych H., 580
Nazarkevych M., 580
Nechyporenko A., 524
Nevlydov I., 13
Nicholas D., 271
Nikolskyi I., 397
Nissen Masmoudi, 342
Noga Yu., 162

O

Oborska O., 145
Oliynyk I., 444
Omelchuk A., 247
Osadchyi V., 218
Oskerko S., 305
Ostakhov V., 50

P

Pabyrivska N., 171, 361
Pabyrivskyy V., 361
Pal G., 25
Pal R., 381
Palchikov V., 84
Partyka S., 183
Pashynska N., 79
Pavlyshenko B., 255
Peleshko D., 305, 113, 455
Peleshko M., 352
Peredrii O., 459, 464
Perova I., 3
Petrasova S., 21
Pitsun O., 554
Plakhtii V., 434
Pliss I., 171, 519
Pochanin G., 434
Pohorelov A., 227
Polezhai V., 509
Polishchuk V., 65
Polyakova M., 605
Polyvoda O., 247
Ponomaryova G., 13
Popovych V., 558
Povshuk O., 162
Prishchenko O., 434
Prokopenko D., 166
Prydatko O., 177, 187, 558
Puchala D., 88

Pukas A., 444
Putrenko V., 79
Putyatin Ye., 464

R

Radivilova T., 198
Radyvonenko O., 218
Rak T., 455
Rakytyanska H., 369
Rashkevych Yu., 113
Rassomakhin S., 514
Riepin V., 7
Riznyk O., 162
Roenko A., 236
Romanov V., 407
Romanyshyn I., 251
Romanyshyn Yu., 488, 568
Rudakova H., 247
Rusyn B., 251, 595
Rybalchenko S., 281
Rydel M., 331

S

Sachenko A., 102, 605, 609
Sadek M., 107
Sakhon A., 509
Savanevych V., 227
Savka N., 554
Semenets V., 524
Semenov B., 430
Serhiienko R., 514
Setlak G., 327
Shafronenko A., 327
Shakhovska N., 478
Sharapov D., 534
Sharkadi M., 65
Shcherbakova G., 605
Shelevytska V., 430
Shelevytsky I., 430
Sherstiuk V., 590
Shevchenko M., 534
Shlokin V., 514
Shynkarenko H., 259
Shyrokorad D., 434
Sikora L., 538
Sipko O., 265
Sirenko F., 236
Skorokhoda O., 438
Skrynkovskyy R., 134, 483
Škvor J., 336
Slonov M., 494
Smelyakov K., 509
Smotr O., 157, 187
Snytyuk V., 265
Sokol I., 590
Sokolovskyy Ya., 375

Solotvinskyy I., 187
Steshenko G., 346
Stokfiszewski K., 88
Stolbovyi M., 545, 549
Suprun O., 265
Szymanski Z., 70

T

Tarasiuk P., 448
Tatarinova Yu., 364
Tesluyk T., 438
Tkachenko R., 386
Tkachov V., 183
Tkachuk R., 538
Tokarev V., 183
Tomis M., 528
Tsmots I., 438
Tsymbal Yu., 438
Turuta O., 524
Tverdokhlib Ye., 574
Tymchenko O., 584
Tymchenko O. Jr., 584
Tyshchenko O., 402

V

Vahin P., 259
Vergeles A., 166
Verhun V., 98
Vitynskyi P., 386
Vlasenko A., 352
Vlasenko N., 352
Voityshyn V., 98
Volkova M., 13
Volkova N., 56
Volkova V., 218

Volkovsky O., 322
Voloshchuk V., 151
Voloshyn Ol., 65
Voloshyn Or., 455
Voloshyn V., 305
Voronenko M., 336, 411
Vovk V., 534
Vynokurova O., 113, 305, 352
Vysotska V., 128, 139, 145

W

Walaszek-Babiszewska A., 331
Wieczorek L., 424
Wieloch K., 88

Y

Yakobchuk P., 102
Yatsymirskyy M., 88, 448
Yelmanov S., 488, 568
Yeremenko D., 509
Yerokhin A., 524
Yuzevych V., 134

Z

Zadorozhnii Ye., 574
Zahorodnia D., 609
Zaporozhets Yu., 44
Zavgorodnia O., 202
Zavgorodnii I., 3
Zayika O., 171
Zharikova M., 590
Zhernova P., 7, 171
Zozulia V., 534
Zrigui M., 107
Zyma O., 202

Chief Editors

Olena Vynokurova, Dmytro Peleshko

Editorial Board

V. Vysotska, Yu. Borzov

Printing by Publishing House of
Lviv Polytechnic National University
2 Kolessa str., Lviv, Ukraine, 79013, vlp@vlp.com.ua

softserve

GlobalLogic



Perfectial
EMPOWER YOUR IDEAS



КИЇВСТАР

СХІДНИЦЬКА 118



altexsoft
software r&d engineering



LMX



Hey
Machine Learning

teragence™
YOUR NETWORK YOUR EXPERIENCE



PM Business Solution



Lviv City
Council

