Data Stream Mining & Processing

Proceedings of
IEEE Third International Conference on
Data Stream Mining & Processing

August 21-25, 2020
Lviv, Ukraine

IEEE

Ukrainian Catholic University

Manhattan College

DSMP

Proceedings of the

2020 IEEE Third 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

IEEE Ukraine Section IM/CIS Societies Joint Chapter

Ukrainian Catholic University

Manhattan College

Kharkiv National University of Radio Electronics

Lviv, Ukraine
August 21-25, 2020

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 ©2020 by IEEE.

Additional copies may be ordered from:

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

DSMP'2020 Organizing Committee
E-mail: dsmp.conference@gmail.com

IEEE Catalog Number: CFP20J13-USB

ISBN: 978-1-7281-3213-6

DSMP'2020 Conference Committee

Honorary Chairpersons

Yuriy Rashkevych, Ukraine

Yevgeniy Bodyanskiy, Ukraine

Igor Aizenberg, USA

General Chairs

Dmytro Peleshko, Ukraine
Olena Vynokurova, Ukraine
Yaroslav Prytula, Ukraine

Technical Program Committee Chair

Dmytro Peleshko, Ukraine

Publication and Finance Chair

Olena Vynokurova, Ukraine

Local Organizing Committee Chair

Tatiana Sviridova, Ukraine

Technical Program Committee

Ankur Agrawal, USA
Svitlana Antoshchuk, Ukraine
Sergii Babichev, Czech Republic
Jayaram Balasubramaniam, India
Oleksandr Baiev, Ukraine
Oleg Berezkiy, Ukraine
Petro Bidyuk, Ukraine
Sergii Bogomolov, Australian
Vilalii Boyun, Ukraine
Gennadii Churyumov, Ukraine
Ibraim Didmanidze, Georgia
Kai Du, China
Mykola Dyvak, Ukraine
Oleksandr Dumin, Ukraine
Andrey Fisunenko, Ukraine
Mohammed Gabsi, France
Eduard Petlenkov, Estonia
Shao Cheng Qu, China
Taras Panchenko, Ukraine
Nataliya Pankratova, Ukraine
Bohdan Pavlyshenko, Ukraine
Ievgen Pichkalov, Ukraine
Kashifuddin Qazi, USA
Olga Radyvonenko, Ukraine
Ali Rekik, Tunisia
Yurii Romanyshyn, Ukraine
Bohdan Rusyn, Ukraine
Anatolii Sachenko, Ukraine
Galina Setlak, Poland
Nataliya Sharonova, Ukraine
Igor Shelevytsky, Ukraine
Aleksandr Slipchenko, Netherlands

Mounir Gabsi, Tunisia
Oleksandr Gozhiy, Ukraine)
Rostyslav Gryniv, Ukraine
Volodymyr Hnatushenko, Ukraine
Wen Bin Hu, China
Zhengbing Hu, China
Kareem Kamal A. Ghany, Egypt
Bekir Karlik, Albania
Ghédira Khaled, Tunisia
Vyacheslav Kharchenko, Ukraine
Frank Klawonn, Germany
Illya Kokshenev, Brazil
Viktor Krylov, Ukraine
Yurii Krak, Ukraine
Yurii Kondratenko, Ukraine
Chui Wei Lu, China
Volodymyr Lytvynenko, Ukraine
Leonid Lyubchik, Ukraine
Yaroslav Lubinets, Ukraine
Mykola Malyar, Ukraine
Krassimir Markov, Bulgaria
Viktor Mashkov, Czech Republic
Volodymyr Mashtalir, Ukraine
Sergii Mashtalir, Ukraine
Andrian Nakonechn, Ukraine
Andrzej Smolarz, Poland
Vitalii Snytyuk, Ukraine
Oleksandr Sokolov, Poland
Yaroslav Sokolovsky, Ukraine
Makram Souii, Tunisia
Volodymyr Stepashko, Ukraine
Martin Štěpnička, Czech Republic
Sergii Subbotin, Ukraine
Jun Su, China
Zdislav Szymanski, Poland
Moncef Temani, Tunisia
Oleksii Tyshchenko, Czech Republic
Roman Tkachenko, Ukraine
Ivan Tsmots, Ukraine
Kristina Vassiljeva, Estonia
Valentyna Volkova, Ukraine
Roman Vorobel, Ukraine
Sergii Vorobyov, Finland
Waldemar Wójcik, Poland
J.Q. Wu, China
Felix Yanovsky, Ukraine
Myhailo Yatsymirskyy, Poland
Zhi Wei Ye, China
Elena Yegorova, United Kingdom
Yevhenii Yakishyn, China

Partners

Gold partner

GeoGuard
www.geoguard.com

Partners

Lviv City Council
http://city-adm.lviv.ua/

Lviv Convetion Bureau
http://www.lvivconvention.com.ua/en/

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'2020), which is held in Lviv, UKRAINE, 21-25 August, 2020. 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, IEEE Ukraine Section IM/CIS Societies Joint Chapter, Ukrainian Catholic University, Manhattan College and Kharkiv National University of Radio Electronics.

Agenda of the DSMP'2020 is very rich. This year we have nominated a 88 number of accepted papers coming from about 12 countries which makes DSMP a truly international high impact conference. Major highlights of DSMP'2020 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 Igor Aizenberg

General List of Topics

Table of Contents

Topic #1

Hybrid Systems of Computational Intelligence

IEEE Third International Conference on Data Stream Mining & Processing
August 21-25, 2020, Lviv, Ukraine

Hybrid Machine Learning System for Solving Fraud Detection Tasks

Olena Vynokurova
GeoGuard
Kharkiv, Ukraine
vynokurova@gmail.com

Dmytro Peleshko
GeoGuard
Lviv\Kharkiv, Ukraine
dpeleshko@gmail.com

Oleksandr Bondarenko
GeoGuard
Kharkiv, Ukraine
alexander@geoguard.com

Vadim Ilyasov
GeoGuard
Kharkiv, Ukraine
vadim@geoguard.com

Vladislav Serzhantov
GeoGuard
Kharkiv, Ukraine
vladislav@geoguard.com

Marta Peleshko
Lviv State University of Life Safety
Lviv, Ukraine
marta.peleshko@gmail.com

***Abstract* — In parallel with technological development the problem of fraud detection is becoming more and more important. Increasing number of electronic transactions in various business environments, on the one hand, and software and technology development, on the other hand, lead to an active increase in electronic crime. In the paper the hybrid system of machine learning for solving tasks of anomalies detection has been proposed. This hybrid system consists of two subsystems – anomalies detection subsystem (based on unsupervised learning) and the interpretation subsystem of anomaly type (based on supervised system). The advantage of proposed hybrid system is the high-speed data processing when the data are fed in real time. The effectiveness of the proposed approach was confirmed during the solution of the detecting anomalies problem based on real data streams.**

Keywords— fraud detection, anomaly detection, hybrid system, isolation forest, random forest, transactions, machine learning

I. INTRODUCTION

In parallel with technological development the problem of fraud detection is becoming more and more important. Increasing number of electronic transactions in various business environments, on the one hand, and software and technology development, on the other hand, lead to an active increase in electronic crime. Authentication methods are no longer the only way to protect against fraud. Early detection of fraud is one of the main ways to prevent fraud. Blocking anomalous electronic transactions in some cases is almost the main way to avoid fraud. However, the development of mathematical methods for detecting fraud stimulates the skilled development of ways for concealing fraud. This leads to the fact that the practical algorithms of fraud detection are no longer universal.

In many cases, in order to increase the accuracy of early identification of anomalous electronic transactions, it is necessary to develop specialized software solutions. The essence of specialization is to use models that are adapted to the specifics of the company's business activities. For example, most of the available scientific papers on detecting fraud-related anomalies are related to credit cards.

This means that fraud detection methods that focus on the specific nature of input data that contain information from electronic transactions related to credit cards. And the resulting classification models are tightly tied to this business domain. This specialization is not quite a disadvantage, as it allows easy enough scaling of systems or expansion of types of detection when new types of fraud appear.

II. RELATED WORK

Using set of rules is one of first approach for developing fraud detection systems. This approach has been developed in the form of knowledge base system. The most well-known mechanisms for their implementation are expert systems [1].

Using a predefined set of rules simplifies the software development of fraud detection systems, but in general such systems have a number of disadvantages:

- the development of rules depends on the quality of the examination of the business environment. This determines the direct dependence of the effectiveness of the set of rules on the qualifications of expert analysts who create these rules [2].
- expanding the system of rules is costly. New experts are needed to expand the set of rules. Therefore, the appearance of new types of fraud leads to increased costs for modification of software systems.
- with a significant increase in the set of rules, the speed of the system can significantly decrease. This problem is getting more intense when large feature vectors are used.
- in the case, when the rules use a threshold, it is very difficult to achieve the adaptability of these values to environmental conditions.

Another defining characteristic of rule-based systems is the size of the rule base [1]. Small size databases occur primarily in cases where the input data vectors have a small dimension. Therefore, software solutions based on such databases are characterized by high speed. But the accuracy of these systems will again depend on experts.

In terms of support, small databases are much easier to administer. This is another advantage of small databases.

However, modern operational processes manipulate large-scale vectors. And this fact leads to a significant increase in the database and reduce the advantages of using rules for fraud detection tasks.

Other methods for solving a fraud detection tasks are statistical methods. The group of statistical methods includes methods that are based on elements of probability theory, mathematical statistics and data collected over a period of time.

Using statistical methods is one of the modern main directions of development of fraud detection methods. On the one hand, high accuracy of anomaly detection is obtained. On the other hand, the use of various inaccurate estimation parameters greatly reduces the flexibility of these methods and adaptability to changes in input data. For example, many methods require setting thresholds. Other methods require information on the statistical distribution of input data, etc. [3].

Today, all static methods of fraud detection can be divided into two categories: supervised and unsupervised methods. Both categories of these methods are united by the use of historical data (record of observations from the past) for effective fraud detection. The depth of this story for each method of different categories may be different. One of the main problems of supervised methods is the need to have sets of labeled features at the input.

This is not always possible and therefore contributes to the development of unsupervised methods. In the [4] authors have used a combination of unsupervised and supervised methods based on a self-organizing map and a neural network. Another example of a combination is hybrid methods from [5]. In [6] the classification of various hybrid methods is presented. This classification describes a variety of combined uses of unsupervised and supervised methods. In addition, a significant number of comparative experiments were conducted to assess the effectiveness of their use.

In point of view of the development and operation of software systems for fraud detection, machine learning methods have three main advantages:

1. An increase of the electronic transactions number usually leads to an increase in the accuracy of fraud detection models.

2. The dimensions of modern data sets make it impossible to analyze them without automation. Machine learning methods significantly simplify and increase the processing speed of large data sets.

3. The use of machine learning methods makes it possible to detect hidden dependencies. This is important to improve the accuracy of the systems and to increase the resistance of the system to the emergence of new types of fraud.

As the analysis of scientific results shows that among the most popular methods for Fraud Detection are Logistic regression, Random Forests and Support Vector Machines [7, 8]. It should be noted that [7] shows the efficiency of classification of anomalies using SVM. And in [8] the effectiveness of anomaly detection using Random Forests.

III. Anomaly Detection Hybrid Architecture for Fraud Detection Problems

For solving task of fraud detection the pipline for real-time anomality detection is proposed.

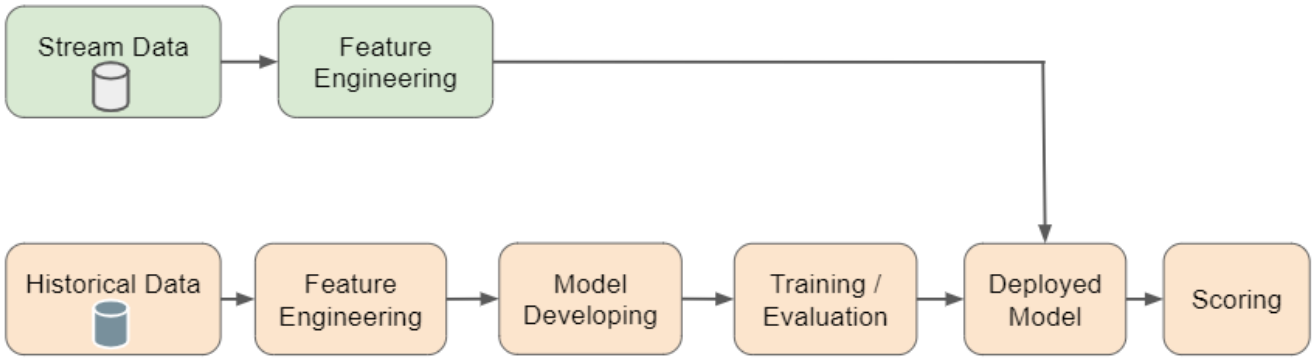


Fig. 1. Architecting an Anomaly Detection Pipeline

This pipeline consists of two flows: training/testing flow, which trains and tests the developed model and retrains it, evaluation flow that processing data stream from server in real time. The data are collected using no-sql DB Elasticsearch and each transaction is stored in json file. Depending on the solution type of transaction (Android, IOS, gdk (Windows and MAC), Solus (html), Plugin (Windows and MAC)), the different type of information about transaction can be stored in json file.

For training/testing flow we have collected the historical data for the dataset, which must be balanced for all type of transactions and their combining. After that, the data are fed to the stage of feature embedding, which is the most time-consuming, at this stage is cleaning, normalizing and embedding data. At this stage, the final training and testing dataset for the developed model is formed.

The next stage is the development of an anomaly detector model and a system for interpreting anomalies type. Then the model is trained and tested and after that we get a ready model for use on the evaluation flow.

The evaluation flow consists of capturing the data stream from the server and forwarding it to the feature engineering stage. Based on the developed embedding methods on the

training / testing flow, a dataset is obtained which is fed to the deployed model. After that the calculation of transaction scoring and making decision are performed.

Feature Engineering stage is the most time-consuming. It consists of the selecting field from database, the correlation analysis, the cleaning and preprocessing data and features embedding. Based on correlation analysis we select 64 fields for building dataset. Furthermore, we have to fill gaps in data. Since each transaction may have different filled fields in the connection with solution type. It is necessary to fill in all gaps based on machine learning methods or expert analysis of each field. It is important because quality of filling gaps affects accuracy of anomaly detection. And, also, we need to normalize and code numeric type data. The next stage of feature engineering is Categorical Data Embedding. Among the analyzed fields, 70% of fields are categorical variables. Different fields require different encoding, or the combination and encoding of several features together. We used: Label Encoding, One Hot Encoding, Embedding Vector, Binary Encoding, Hashing, Crosstab, Frequency Encoding and some modified methods.

The current version of the anomaly detector model is based on data from 64 fields of database. These fields describe the id of users and devices; the information about geolocation of the users and devices, which perform this transaction the history of transactions; the information about connection type (gsm, gps, wifi, ip); the information about running process and software on the devices that can be rooted; spoofed information about users and their location and etc.

Based on these fields and its combination 41 features is developed

$$x(k) = \{x_1(k), x_2(k), \ldots, x_n(k)\} \quad (1)$$

where $x(k)$ is transaction, $x_i(k)$, $i = 1 \ldots n$ are features (in our case n=41), k is real time.

This set of features forms the training and testing dataset.

On the case study stage, it was determined that the solution of the problem of fraud detection involves the need not only to determine anomalous transactions, but also to interpret why the transaction was failed. All existing methods of detection anomalies solve only the first problem. Thus, it is necessary to develop a hybrid model that could solve the problem of fraud detection and interpretation of the transaction anomaly type.

Proposed approach in the paper involves the use of two sequential models as entities to solve the problem of fraud detection. The first model is a binary classifier that solves the 'fraud' or 'non-fraud' problem. The next model is a multi-class classifier that defines the 'fraud' type. The general architecture is shown in fig. 2 and a more detailed architecture of the deployed model are shown in Fig. 3.

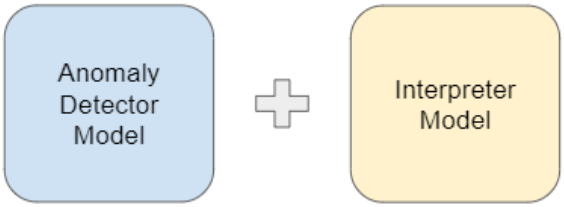


Fig. 2. General architecture of model

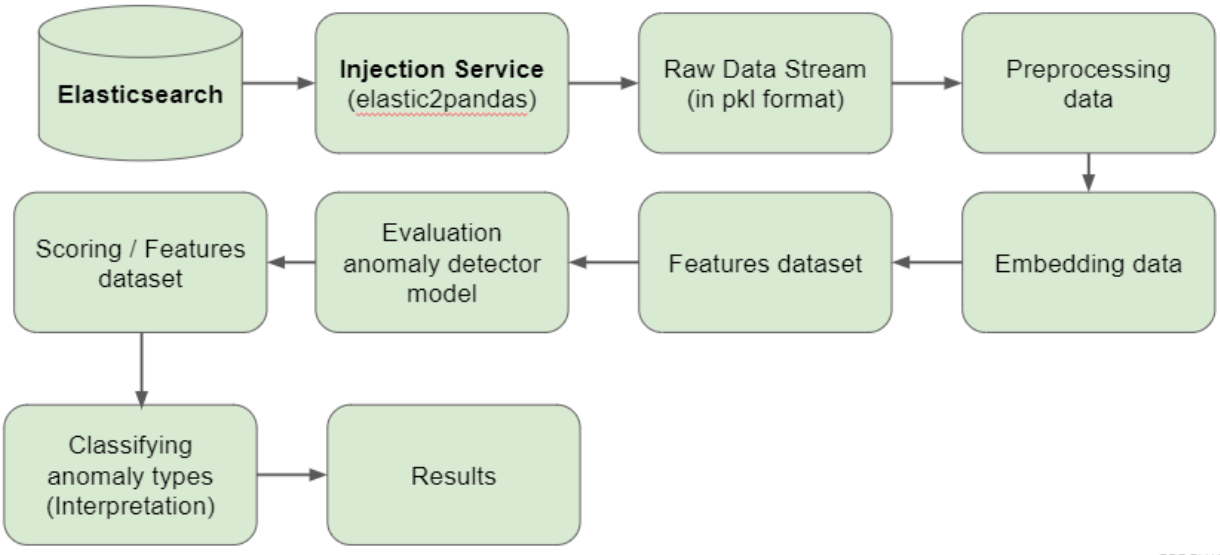


Fig. 3. Detailed architecture of deployed model

Nowadays there are a lot of methods for anomaly detection, the Robust Covariance method [9], One-class SVM method [10], Local Outlier Factor method [11], KNN method [12] and Isolation Forest [13] have been investigated, and for

developing model of anomaly detector we have selected Isolation Forest. Among the different anomaly detection algorithms, Isolation Forest is one with unique capabilities. It is a model free algorithm that is computationally efficient, can easily be adapted for use with parallel computing paradigms, and has been proven to be highly effective in detecting anomalies. In our case, it gave best accuracy among others.

Thus, the vector $x(k) = \{x_1(k), x_2(k), \ldots, x_n(k)\}$ from the constructed dataset is fed to generate an isolation tree. x is recursively separated by randomly selecting a feature and a random value of this feature between min(x_q) and max(x_q) the values of the selected feature and so on until the tree is constructed. Thus, we get an isolation tree which is a proper binary tree, where each node in the tree has exactly zero or two daughter nodes.

The task of detecting anomalies is to provide a rating of transactions that reflects the degree of their anomaly. Thus, one way to detect an anomalous transaction is to sort the transactions according to their length or anomaly scores; and anomalies are transactions that will be at the top of the list. The path length and anomaly estimate are determined by the algorithm proposed in [13].

In the case of Isolation Forest, anomaly score is defined as:

$$s(x,l) = 2^{-\frac{E(p(x))}{m(l)}} \tag{2}$$

where $p(x)$ is the path length of observation x, $m(l)$ is the average path length of unsuccessful search in a Binary Search Tree and l is the number of external nodes. More on the anomaly score and its components can be read in [13].

Each observation is given an anomaly score and the following decision can be made on its basis:

- a score close to 1 indicates anomalous transactions;
- score much smaller than 0.5 indicates normal transactions;
- if all scores are close to 0.5 then the entire sample does not seem to have clearly distinct anomalous transactions

The random forest method [14] was chosen to develop an interpreter model of the fraud detection hybrid system. The interpreter model provides an explanation to the provider-companies, why the transaction was determined as abnormal. Because the system under development may have situations where several types of anomalies can be present in a single transaction, the use of random forest-based methods is a priority.

For each decision tree, Scikit-learn calculates a nodes importance using Gini Importance, assuming only two child nodes (binary tree):

$$n_j^{im} = w_j U_j - w_j^r U_j^r - w_j^l U_j^l \tag{3}$$

where n_j^{im} is the importance of node j, w_j is weighted number of samples reaching node j, U_j is the impurity value of node j, $\bullet_j^r$ is child node from right split on node j, $\bullet_j^l$ is child node from left split on node j.

The importance for each feature on a decision tree is obtained in the form:

$$f_i^{im} = \frac{\sum_j n_j^{im}}{\sum_{k \in all\ nodes} n_k^{im}} \tag{4}$$

where f_i^{im} is the importance of feature i, n_j^{im} is the importance of node j.

After that these features importance are normalized using expression

$$\overline{f}_i^{im} = \frac{f_i^{im}}{\sum_{j \in all\ features} f_j^{im}}. \tag{5}$$

The final feature importance, at the Random Forest level, is its average over all the trees. The sum of the feature's importance value on each tree is calculated and divided by the total number of trees:

$$RF_i^{im} = \frac{\sum_{j \in all\ trees} \overline{f}_{ij}^{im}}{Tr} \tag{5}$$

where RF_i^{im} is the importance of feature i calculated from all trees in the Random Forest model, $\overline{f}_{ij}^{im}$ is the normalized feature importance for i in tree j, Tr is total number of trees.

Anomalous transactions and borderline transactions that have been detected by the anomaly detector model are fed to the interpreter model. Cross validation has used for tuning hyperparameters of interpreter model. Therefore, we have a cascade of classifiers which in general presents the proposed anomaly detector model.

IV. EXPERIMENTS

The proposed hybrid system was developed to solve the problem of detecting anomalies in the geolocation of users during transactions (GPS spoofing, Wi-Fi spoofing, location jumping, etc.). Experimental studies were conducted on the DB of GeoGuard company.

The specific feature of the system is that the decision cannot be made at the level of an anomaly, the type of anomaly must be explained to justify the decision.

The input vector consisted of 41 features based on information collected in the no-sql Elasticsearch database in json form for each transaction. Transaction information depends on the type of user's operating system (Android, IOS, MacOs, Windows) and consists of fields describing user geolocation when executing transaction from gsm, gps, wi-fi sources types and fields describing user's device. Frequency of transaction appearance in the environment is 800 transactions per 1 minute.

Training dataset consists of about 90 000 samples, number of trees was 300 trees for anomaly detection model and 500 trees for interpreter model.

Table I shows the results of detection and accuracy of interpretation of proposed hybrid machine learning system.

TABLE I. THE RESULTS OF ANOMALIES DETECTION AND THEIR INTERPRETATION

Solution	**Accuracy, %**		
	Training Detector	***Testing Detector***	***Classification & Interpreter***
all solution	91	90	90
ios	91	90	95
android	92	90	95
plugin (Windows+ MacOs)	99	96	99
gdk (Windows+ MacOs)	97	96	98

The results show that the anomaly detection on each individual solution is more accurate than when all solutions are combined into one dataset. The specificity of the system is the need to balance the dataset, in which all known anomalies on each operating system must be presented. It is also a feature of the system that multiple anomalies may be present in a transaction at the same time, which complicates the process of interpretation.

V. CONCLUSION

In the paper a hybrid system of machine learning for solving problems of anomaly detection is proposed. Such hybrid system consists of two subsystems - subsystem of anomaly detection and subsystem of anomaly type interpretation (classification), which are based on a cascade of decision trees with supervised and unsupervised learning. The advantage of the hybrid system is the speed of processing the data that are fed in real time. The effectiveness of the proposed approach has been confirmed in solving the practical problem of detecting anomalies in user geolocation during transactions execution (GPS spoofing, Wi-Fi spoofing, location jumping, etc.).

REFERENCES

[1] N.F. Ryman-Tubb, P. Krause, and W. Garn. “How Artificial Intelligence and machine learning research impacts payment card fraud detection: A survey and industry benchmark,” Engineering Applications of Artificial Intelligence, no. 76, pp. 130–157, 2018.

[2] R. P. Dazeley, “To The Knowledge Frontier and Beyond: A Hybrid System for Incremental Contextual- Learning and Prudence Analysis,” PhD thesis, University of Tasmania, 2006 http://https://eprints.utas.edu.au/8173

[3] A. Patcha, and J. M. Park, “An overview of anomaly detection techniques: Existing solutions and latest technological trends,” Computer Networks, vol. 51(12), pp. 3448–3470, 2007.

[4] J. T. Quah, and M. Sriganesh “Real-time credit card fraud detection using computational intelligence,” Expert Systems with Applications, vol. 35 (4), pp. 1721–1732, 2008.

[5] Y. Moreau, E. Lerouge, H. Verrelst, C. Stormann, P. Burge, and K. U. Leuven, “A hybrid system for fraud detection in mobile communications,” Neural Networks, pp. 447–454, April 1999.

[6] C. Phua, V. Lee, K. Smith, and R. Gayler, “A Comprehensive Survey of Data Mining-based Fraud Detection Research.” DOI: 10.1016/j.chb.2012.01.002. Arxiv:1009.6119., 2010.

[7] S. Bhattacharyya, S. Jha, K. Tharakunnel, and J. C. Westland. “Data mining for credit card fraud: A comparative study,” Decision Support Systems, vol. 50 (3), pp. 602–613, 2011.

[8] D. Meyer, F. Leisch, K. Hornik. “The support vector machine under test,” Neurocomputing, vol. 55 (12), pp. 169 –186, 2003.

[9] P. J. Rousseeuw and M. Hubert, “Anomaly Detection by Robust Statistics” arXiv:1707.09752v2 [stat.ML] 14 Oct 2017

[10] R. Zhang, Sh. Zhang, S. Muthuraman, and J. Jiang. “One Class Support Vector Machine for Anomaly Detection in the Communication Network Performance Data,” In: 5th WSEAS Int. Conference on Applied Electromagnetics, Wireless and Optical Communications, 2007.

[11] M. M. Breunig, H.-P. Kriegel, R. T. Ng, J. Sander, “LOF: Identifying Density-Based Local Outliers” in Proc. ACM SIGMOD 2000 Int. Conf. On Management of Data, Dalles, TX, 2000

[12] Y. Djenouri, A. Belhadi, J. C. Lin and A. Cano, "Adapted K-Nearest Neighbors for Detecting Anomalies on Spatio–Temporal Traffic Flow," in IEEE Access, vol. 7, pp. 10015-10027, 2019. doi: 10.1109/ACCESS.2019.2891933

[13] T.L. Fei, M. T. Kai, Z. Zhi-Hua, “Isolation Forest,” in Proceedings of the 2008 Eighth IEEE International Conference on Data Mining, pp. 413–422, December 2008, https://doi.org/10.1109/ICDM.2008.17

[14] L. Breima. “Random Forests. Machine Learning”, v. 45 (1), pp. 5–32, 2010, doi:10.1023/A:1010933404324

IEEE Third International Conference on Data Stream Mining & Processing
August 21-25, 2020, Lviv, Ukraine

Speech Signal Structuring Method for Biometric Personality Identification

Eugene Fedorov
Department of Robotics and Specialized Computer Systems
Cherkasy State Technological University
Cherkasy, Ukraine
fedorovee75@ukr.net

Tetyana Utkina
Department of Robotics and Specialized Computer Systems
Cherkasy State Technological University
Cherkasy, Ukraine
t.utkina@chdtu.edu.ua

Olga Nechyporenko
Department of Robotics and Specialized Computer Systems
Cherkasy State Technological University
Cherkasy, Ukraine
olne@ukr.net

Yaroslav Korpan
Department of Robotics and Specialized Computer Systems
Cherkasy State Technological University
Cherkasy, Ukraine
populusdocti@gmail.com

Abstract. *This paper proposes a method for structuring a speech signal. For this, segmentation method, methods for determining the fundamental tone of the vocal segment and determining on its basis the boundaries of the quasiperiodic oscillations of the vocal segment, the geometric transformation of quasiperiodic oscillations of the vocal segment were suggested. The proposed segmentation of the speech signal uses statistical estimation of short-term energies, which allows the use of an adaptive threshold, thus increasing the vocal segments determination accuracy. The proposed definition of fundamental tone of the vocal segment uses bandpass filtering and statistical estimation of local extremum, which reduces computational complexity, and also reduces noise dependency and allows the use of an adaptive threshold, thus increasing the accuracy of determining the fundamental tone and the boundaries of quasiperiodic oscillations of the vocal segment. The proposed geometric transformation of quasiperiodic oscillations of the vocal segment allows you to transform quasiperiodic oscillations to a single amplitude-time window, which allows you to form frames of the vocal segment, taking into account its structure.*

Keywords: *biometric identification of a person, speech signal, segmentation structuring, calculating the fundamental tone, dividing segments into quasiperiodic oscillations, geometric transformation of quasiperiodic oscillations.*

I. INTRODUCTION

Automated biometric identification of a person implies making decisions based on acoustic and visual information, which improves the recognition quality of the studied person [1-3]. Unlike the traditional approach, computer-assisted biometric identification speeds up and improves the accuracy of the recognition process, which is especially crucial in a limited time conditions.

Methods based on the analysis of acoustic information form a special class of biometric identification of a person [4-8].

Well-known methods of voice biometric identification of a person, such as dynamic programming [9, 10], vector quantization [11, 12], artificial neural networks [13, 14] and decision tree [15], Gaussian mixture models (GMM) [16-19], or their combination [20], when analyzing the signal, divide the signal into frames (sections of equal length) without taking into account its structure, which reduces the effectiveness of biometric identification.

Speech signal structuring is based on the segmentation and partitioning of segments of vocal sounds into quasiperiodic oscillations based on the fundamental tone.

The following features are usually used for segmentation [21]:

- the number of transitions through zero;
- energy.

Said methods do not use an adaptive threshold, thus reducing the segmentation accuracy.

Methods that are used to calculate the fundamental tone are based on the analysis of the following representations of the signal [21, 22]:

- amplitude-time;
- spectral (amplitude-frequency);
- cepstral (amplitude-random);
- wavelet spectral (amplitude-frequency-time).

These methods have one or more of the following disadvantages:

- have high computational complexity;
- depend on the noise level, which reduces the accuracy of determining the fundamental tone;
- do not use an adaptive threshold, which reduces the accuracy of determining the fundamental tone.

Thereby, it is of current interest to create a method for speech signal structuring, which will eliminate these drawbacks.

The aim of the work is to increase the efficiency of biometric identification of a person using preliminary signal structuring.

To achieve this goal it is necessary to solve the following problems:

- vocal speech signal segments determination based on short-term statistical evaluation of energies;
- fundamental tone of the vocal segment determination based on bandpass filtering and statistical estimation of local extremum;
- determination of the boundaries of quasiperiodic oscillations of the vocal segment based on the fundamental tone;
- geometric transformation of quasiperiodic oscillations of the vocal segment to a single amplitude-time window;
- the speech signal structuring quality criteria determination.

II. Vocal Speech Signal Segments Determination Based on Short-Term Statistical Evaluation of Energies

The paper proposes a method for determining vocal segments of a speech signal based on statistical estimation of short-term energies, which includes the following steps:

1. Define the speech signal with one vocal sound $y(n)$, $n \in \overline{1, N^f}$. Set the number of quantization levels of a speech signal L (for an 8-bit sound sample $L = 256$). Specify the length of the frame N, on which the momentary energy is calculated, $N = 2^b + 1$, wherein an integer parameter b is selected from the inequality $b - 1 < log_2\left(\frac{f_d}{f_{min}}\right) < b$, f_d – a speech signal sampling rate in Hz, $8000 \le f_d \le 22050$, f_{min} – the minimum frequency of the person's fundamental tone in Hz, $f_{min} = 50$. Set a parameter for the adaptive threshold β, $0 < \beta < 1$.

2. Calculate short-term energies

$$E(n) = \sum_{m=-N/2}^{N/2} y^2(m+n),\ n \in \overline{N/2+1,\ N^f - N/2 - 1}.$$

3. Calculate the mathematical expectation of short-term energies

$$\mu = \frac{1}{N^f - N - 1} \sum_{n=N/2+1}^{N^f - N/2 - 1} E(n).$$

4. Calculate the standard deviation of short-term energies

$$\sigma = \sqrt{\frac{1}{N^f - N - 1} \sum_{n=N/2+1}^{N^f - N/2 - 1} E^2(n) - \mu^2}.$$

5. Calculate the adaptive threshold

$$T = \mu - \beta\sigma.$$

6. Determine the left and right boundaries of the vocal segment

6.1. Set the reference number $n = 1$.

6.2. If $E(n) < T \wedge E(n+1) \ge T$, then $N^l = n + 1$ go to step 6.4.

6.3. If $E(n) \ge T \wedge E(n+1) < T$, then $N^r = n$, complete.

6.4. If $n < N^f - N - 1$, then go to the next sample, i.e. $n = n + 1$, go to step 6.2., otherwise $N^r = n$, complete.

As a result, the left and right boundaries of the vocal segment are determined.

III. Fundamental Tone of the Vocal Segment Determination Based on Bandpass Filtering and Statistical Estimation of Local Extremum

The paper proposes a method for determining the fundamental tone of the vocal segment based on bandpass filtering and statistical estimation of local extremum, which includes the following steps:

1. Define a speech signal with vocal sound $y(n)$, $n \in \overline{1, N^f}$. Set filter parameter α, $0 < \alpha < 1$. Set the left and right boundaries N^l, N^r of the vocal segment in the signal $y(n)$. Set the minimum frequency of the person's fundamental tone in Hz, $f1$, $f1 = 85$. Set the maximum frequency of the person's fundamental tone in Hz, $f2$, $f2 = 255$. Set the number of frames $I = \left[N^f / N\right]$ of length N. $N = 2^b$, wherein an integer parameter b is selected from the inequality $b - 1 < log_2\left(\frac{f_d}{f_{min}}\right) < b$, f_d – a speech signal sampling rate in Hz, $8000 \le f_d \le 22050$, f_{min} – the minimum frequency of the person's fundamental tone in Hz, $f_{min} = 50$, [ ] – getting the integer part of the number.

2. Break the speech signal into frames

$$s_i(n) = y((i-1) * N + n + 1),\ n \in \overline{0, N-1},\ i \in \overline{1, I}.$$

3. Balance frames spectrums, that have a steep decline in the high frequency region, using LPF

$$\breve{s}_i(n) = s_i(n+1) - \alpha s_i(n),\ n \in \overline{0, N-1},\ i \in \overline{1, I}.$$

4. Calculate suspended frames spectrums, using a direct discrete Fourier transform and the Hamming window

$$\hat{s}_i(n) = \breve{s}_i(n)\, w(n),\ w(n) = 0.54 + 0.46\, cos\frac{2\pi n}{N},$$
$$n \in \overline{0, N-1},\ i \in \overline{1, I},$$
$$\hat{S}_i(k) = \sum_{n=0}^{N-1} \hat{s}_i(n)\, e^{-j(2\pi/N)kn},\ k \in \overline{0, N-1},\ i \in \overline{1, I},$$

where $w(n)$ – Hamming window.

$$\widetilde{S}_i(k)=\begin{cases}\widehat{S}_i(k), & f1*N/f_d \le k \le f2*N/f \\ 0, & f1*N/f_d > k \vee k > f2*N/f\end{cases}$$
$$k \in \overline{0,N-1},\ i \in \overline{1,I}.$$

6. Calculate the inverse discrete Fourier transform of the filtered frames

$$\widetilde{s}_i(n)=Re\left(\frac{1}{N}\sum_{n=0}^{N-1}\widetilde{S}_i(k)\,e^{j(2\pi/N)kn}\right),\ n \in \overline{0,N-1},\ i \in \overline{1,I}.$$

7. Combine the filtered frames into a speech signal

$$\widetilde{y}\left((i-1)*N+n+1\right)=\widetilde{s}_i(n),\ n \in \overline{0,N-1},\ i \in \overline{1,I}.$$

8. Calculate the position of local extremum in the filtered vocal segment.

8.1. Set the reference number $n=N^l+1$. Set the number of local extremum $Q=0$.

8.2. If $\widetilde{y}(n)>\widetilde{y}(n-1)\wedge\widetilde{y}(n)>\widetilde{y}(n+1)\wedge\widetilde{y}(n)>0$, then fixate the position of the local extrema, i.e. $e_{Q+1}=n$, increase the number of local extremum, i.e. $Q=Q+1$.

8.3. If $n<N^r-1$, then go to the next sample, i.e. $n=n+1$, go to step 8.2.

9. Calculate the distances between local extremum in the filtered vocal segment

$$\Delta_m=e_{m+1}-e_m,\ m \in \overline{1,Q-1}.$$

10. Calculate the mathematical expectations of the distances

$$\mu=\frac{1}{Q-1}\sum_{m=1}^{Q-1}\Delta_m.$$

11. Calculate the standard deviation of the distances

$$\sigma=\sqrt{\frac{1}{Q-1}\sum_{m=1}^{Q-1}(\Delta_m)^2-\mu^2}.$$

12. Exclude random outliers from the distances

12.1. Set distance number $m=1$. Set the number of new distances $\widetilde{Q}=0$.

12.2. If $\mu-\sigma \le \Delta_m \le \mu+\sigma$, then fixate a new distance, i.e. $\widetilde{\Delta}_{\widetilde{Q}+1}=\Delta_m$, increase the number of new distances, i.e. $\widetilde{Q}=\widetilde{Q}+1$.

12.3. If $m<Q-1$, then go to the next distance, i.e. $m=m+1$, go to step 12.2.

13. Calculate the length of the fundamental tone as the mathematical expectation of new distances

$$N^{FT}=\frac{1}{\widetilde{Q}}\sum_{m=1}^{\widetilde{Q}}\widetilde{\Delta}_m.$$

As a result, the length of the fundamental tone of the vocal segment is determined.

IV. Determination of Quasiperiodic Oscillations of the Vocal Segment

The author's method for determining the boundaries of quasiperiodic oscillations of the vocal segment include the following steps:

1. Define the speech signal $y(n)$, $n \in \overline{1,N^f}$. Set the left and right boundaries of the vocal segment N^l, N^r in the signal $y(n)$. Set the length of the fundamental tone of the vocal segment N^{FT}. Set parameter γ for determining the boundaries of quasiperiodic oscillations of the vocal segment, $0<\gamma<1$.

2. Initialize variables for determining the boundaries of quasiperiodic oscillations of the vocal segment as

$$N_0^{max}=arg\max_n y(n),$$
$$n \in \left\{N^l-\gamma\cdot N_0^{FT},\ldots,N^l+\gamma\cdot N_0^{FT}\right\},\ N_0^{FT}=N^{FT}.$$

3. Set the quasiperiodic vocal segment oscillations number $I=1$.

4. Determine the boundaries of quasiperiodic oscillation of a segment of vocal speech sound in the form of

$$N_I^{min}=N_{I-1}^{max},\ N_I^{max}=arg\max_n y(n),$$
$$n \in \left\{N_I^{min}+(1-\alpha)\cdot N_{I-1}^{FT},\ldots,N_I^{min}+(1-\alpha)\cdot N_{I-1}^{FT}\right\},$$
$$N_I^{FT}=N_I^{max}-N_I^{min}.$$

5. If $N_I^{max} \le N^r$, then increase the number of quasiperiodic oscillations, i.e. $I=I+1$, go to step 4.

As a result, the set of boundaries of the quasiperiodic oscillations of the segment is formed.

V. Geometric Transformation of Quasiperiodic Oscillations of the Vocal Segment to a Single Amplitude-Time Window

Method for geometric transformation of quasiperiodic oscillations of the vocal segment to a single amplitude-time window proposed in this work consists of the following steps:

1. Define the speech signal $y(n)$, $n \in \overline{1, N^f}$. Set the number of quantization levels of a speech signal L (for an 8-bit sound sample $L = 256$). Define the set of boundaries of the quasiperiodic oscillations of the vocal segment $\left\{\left(N_i^{min}, N_i^{max}\right)\right\}$. Set the length of the amplitude-time window N, $N = 2^b$, where the integer parameter b is selected from the inequality $b-1 < log_2\left(\frac{f_d}{f_{min}}\right) < b$, f_d – a speech signal sampling frequency in Hz, $8000 \le f_d \le 22050$, f_{min} – the minimum frequency of the person's fundamental tone in Hz, $f_{min} = 50$.

2. Determine the minimum and maximum values of the quasiperiodic oscillations of the vocal segment as

$$A_i^{min} = \min_n y(n),\ n \in \left\{N_i^{min}, \dots, N_i^{max}\right\},$$
$$A_i^{max} = \max_n y(n),\ n \in \left\{N_i^{min}, \dots, N_i^{max}\right\}.$$

3. Determine the finite set of discrete samples shifted in time and amplitude, described as finite family of integer bounded finite discrete functions $X^s = \left\{x_i^s \mid i \in \{1, \dots, I\}\right\}$, as

$$x_i^s(n) = \begin{cases} y\left(n + N_i^{min} - 1\right) - A_i^{min}, & n \in \{1, \dots, N_i + 1\} \\ 0, & n \notin \{1, \dots, N_i + 1\} \end{cases},$$
$$N_i = N_i^{max} - N_i^{min},\ A_i = A_i^{max} - A_i^{min}.$$

4. Determine a finite set of continuous samples obtained as a result of linear interpolation and described by a finite family of real-valued bounded finite continuous functions $\Psi = \{\psi_i \mid i \in \{1, \dots, I\}\}$, in the form of

$$\forall t \in \left[\breve{T}^{min}, \breve{T}^{max}\right]$$
$$\psi_i(t) = \sum_{n=1}^{N_i} \chi_{(t_n, t_{n+1})}(t)\left(x_i^s(n) + \frac{x_i^s(n+1) - x_i^s(n)}{\Delta t}(t - t_n)\right) + \sum_{n=1}^{N_i+1} \chi_{\{t_n\}}(t) x_i^s(n),$$
$$\forall t \notin \left[\breve{T}^{min}, \breve{T}^{max}\right]\ \psi_i(t) = 0,\ \breve{T}^{min} = \Delta t,\ \breve{T}^{max} = 2^b \Delta t,$$
$$t_n = n\,\Delta t,\ \chi_B(t) = \begin{cases} 1, & t \in B \\ 0, & t \notin B \end{cases},$$

where Δt is the time quantization step of the speech signal, $\chi_B(t)$ is the indicator function.

5. Define a finite set of shifted and time-scaled continuous samples described by a finite family of real-valued bounded finite continuous functions $\Psi^s = \left\{\psi_i^s \mid i \in \{1, \dots, I\}\right\}$, in the form of

$$\forall t \in \left[\breve{T}^{min}, \breve{T}^{max}\right]\ \psi_i^s(t) = \psi_i\left(T_i \frac{t - \breve{T}^{min}}{\breve{T}^{max} - \breve{T}^{min}}\right),$$
$$\forall t \notin \left[\breve{T}^{min}, \breve{T}^{max}\right]\ \psi_i^s(t) = 0,\ \breve{T}^{min} = \Delta t,\ \breve{T}^{max} = 2^b \Delta t.$$

6. Determine the finite set of shifted and amplitude-scaled continuous samples described by a finite family of real-valued bounded finite continuous functions $\Psi^{ss} = \left\{\psi_i^{ss} \mid i \in \{1, \dots, I\}\right\}$, in the form of

$$\forall t \in \left[\breve{T}^{min}, \breve{T}^{max}\right]\ \psi_i^{ss}(t) = \breve{A}^{min} + \frac{\breve{A}^{max} - \breve{A}^{min}}{\widetilde{A}_i^{max} - \widetilde{A}_i^{min}} \psi_i^s(t),$$
$$\forall t \notin \left[\breve{T}^{min}, \breve{T}^{max}\right]\ \psi_i^{ss}(t) = 0,$$
$$\widetilde{A}_i^{max} = \max_t \psi_i^s(t),\ t \in \left[\breve{T}^{min}, \breve{T}^{max}\right],$$
$$\widetilde{A}_i^{min} = \min_t \psi_i^s(t),\ t \in \left[\breve{T}^{min}, \breve{T}^{max}\right],\ \breve{A}^{min} = 1,\ \breve{A}^{max} = L.$$

7. Determine a finite set of discrete samples obtained from continuous ones using time discretization and described by a finite family of integer bounded finite discrete functions $S = \{s_i \mid i \in \{1, \dots, I\}\}$ as

$$s_i(n) = round\left(\psi_i^{ss}(n\,\Delta t)\right)\ n \in \left\{\breve{N}^{min}, \dots, \breve{N}^{max}\right\},$$
$$\breve{N}^{min} = \breve{T}^{min} / \Delta t,\ \breve{N}^{max} = \breve{T}^{max} / \Delta t,$$

where *round ()* – function that rounds the number to the nearest integer.

As a result, the set of samples is formed, all of which are in a single amplitude-time window.

VI. The Speech Signal Structuring Quality Criteria Determination

This paper formulates the speech signal structuring quality criteria as selection of such parameters β and γ, that deliver a minimum of standard error

$$F = \frac{1}{2I} \sum_{i=1}^{I} \left(\widetilde{N}_i^{min} - N_i^{min}\right)^2 + \left(\widetilde{N}_i^{max} - N_i^{max}\right)^2 \to \max_{\beta,\gamma}, \quad (1)$$

where $\widetilde{N}_i^{min}, \widetilde{N}_i^{max}$ – boundaries of quasiperiodic oscillations of the vocal segment, defined by an expert, N_i^{min}, N_i^{max} – calculated boundaries of quasiperiodic oscillations of the vocal segment.

VII. Numerical Study of the Speech Signal Structuring Method

For speech signals containing vocal sounds, the sampling frequency is established at $f_d = 8$ kHz and the number of quantization levels is $L = 256$. Frame length is $N = 256$.

As a result of a numerical study of the speech signal structuring method with parameters $\beta = 0.4$, $\gamma = 0.1$ with

people's speech signals from the TIMIT database according to criteria (1), the standard error of 0.02 was obtained.

Figure 1-4 shows an example of structuring the initial speech signal (Fig. 1) with the segmentation of the speech signal (Fig. 2), determining the boundaries of the quasiperiodic oscillations of the vocal segment (Fig. 3) and the geometric transformation of the quasiperiodic oscillations of the vocal segment to a single amplitude-time window (Fig. 4).

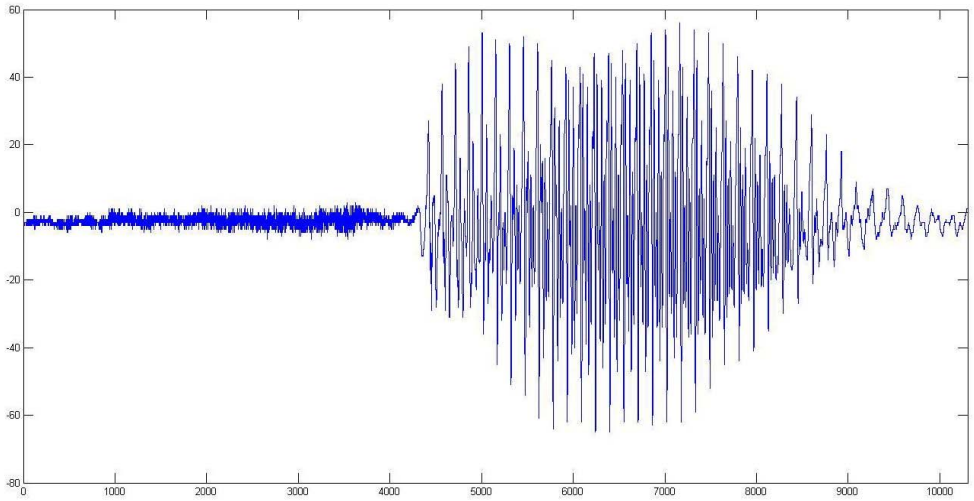


Fig. 1. The initial speech signal (8-bit, 22050 Hz, length 10304)

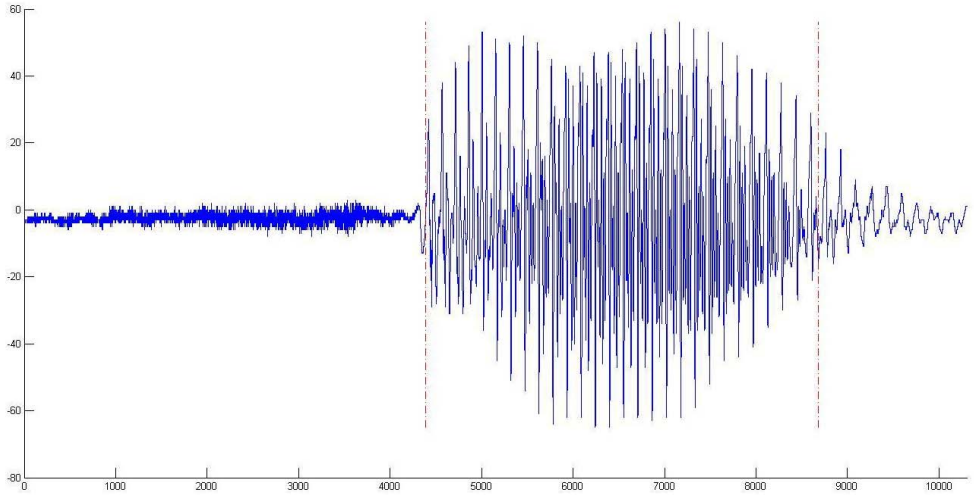


Fig. 2. Speech signal after segmentation, which distinguishes the boundaries of the vocal segment (parameter $\beta = 0.4$)

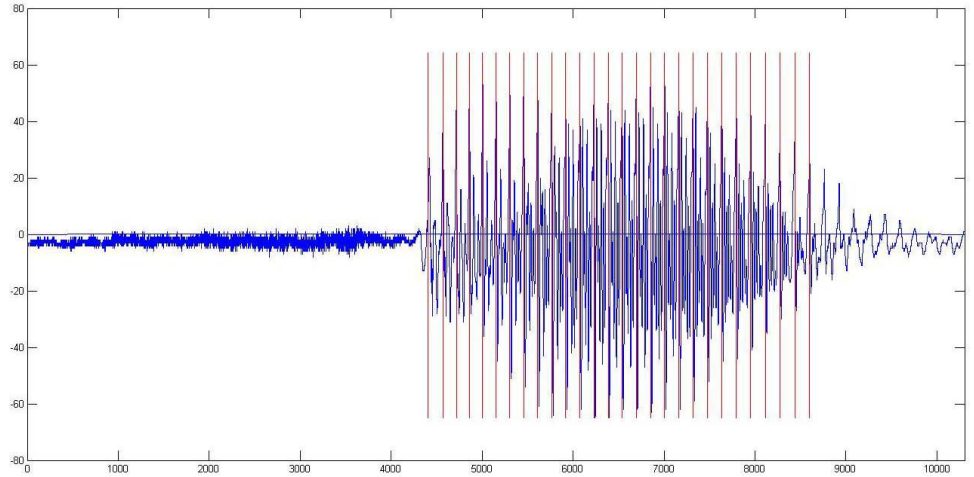


Fig. 3. The speech signal after marking the boundaries of the sound oscillations of the vocal segment (parameter $\gamma = 0.1$)

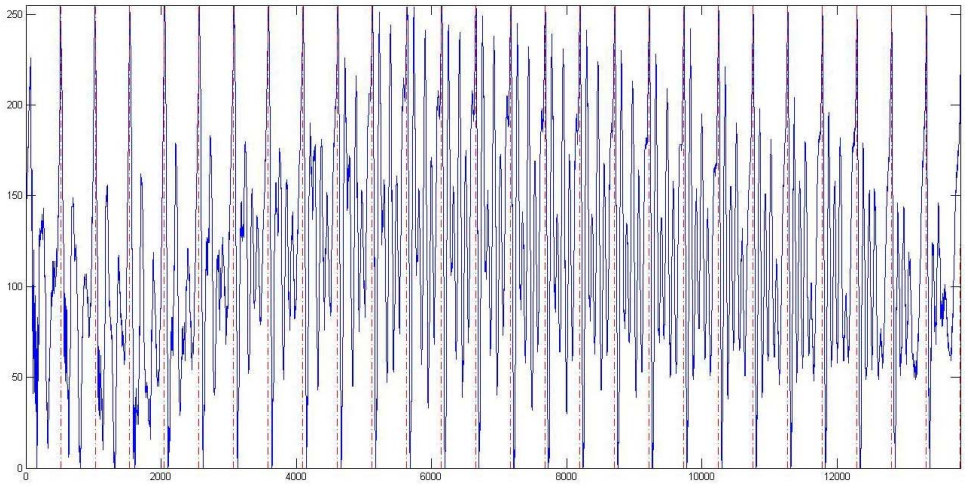


Fig. 4. Vocal sound after geometric transformations of sound oscillations to a single amplitude-time window

CONCLUSIONS

1. To solve the problem of improving the quality of biometric identification by voice, speech signal preprocessing methods, such as segmentation and fundamental tone calculation, were investigated.

2. A speech signal structuring method was proposed, which includes: determination of vocal segments of a speech signal based on statistical estimation of short-term energies, which allows the use of an adaptive threshold, which increases the accuracy of determining vocal segments; determination of the fundamental tone of the vocal segment based on bandpass filtering and statistical estimation of local extremum, which reduces computational complexity, and also reduces the noise dependence and allows the use of an adaptive threshold, which increases the accuracy of determining the fundamental tone and the boundaries of quasiperiodic oscillations of the vocal segment; geometric transformation of sound oscillations of the vocal segment, which allows the transformation of quasiperiodic oscillations to a single amplitude-time window, allowing the vocal segment frames to be formed based on their structure.

3. A numerical study of the speech signal structuring method was carried out, which made it possible to evaluate the proposed method.

4. The proposed structuring method allows to set and solve the tasks of the speech signal preprocessing, which is used to analyze and store biometric information.

REFERENCES

[1] R. M. Bolle, J. H. Connell, S. Pankanti, N. K. Ratha, and A. W. Senior, Guide to Biometrics. New York: Springer, 2004, 364 p.

[2] A. K. Jain, P. Flynn, and A. A. Ross, Handbook of Biometrics. New York, NY: Springer, 2008, 556 p.

[3] T. Dunstone, and N. Yager, Biometric System and Data Analysis Design, Evaluation, and Data Mining. New York, NY: Springer, 2009, 267 p.

[4] N. Singh, R. A. Khan, and R. Shree, "Applications of Speaker Recognition", Procedia Engineering, vol. 38, pp. 3122-3126, 2012.

[5] Q. Li, Speaker Authentication. Heidelberg: Springer, 2012, 237 p.

[6] J. Keshet, and S. Bengio, Automatic Speech and Speaker Recognition: Large Margin and Kernel Methods. Chichester, West Sussex: John Wiley & Sons, 2009, 253 p.

[7] T. Herbig, F. Gerl, and W. Minker, Self-Learning Speaker Identification a System for Enhanced Speech Recognition. Heidelberg: Springer, 2011, 172 p.

[8] J. P. Campbell, "Speaker Recognition: a Tutorial", in Proc. of the IEEE, vol. 85, pp. 1437-1462, 1997.

[9] R. Togneri, and D. Pullela, "An Overview of Speaker Identification: Accuracy and Robustness Issues", IEEE Circuits and Systems Magazine, vol. 11, pp. 23-61, 2011.

[10] H. Beigi, Fundamentals of Speaker Recognition. New York, NY: Springer, 2011, 942 p.

[11] D. A. Reynolds, "An Overview of Automatic Speaker Recognition Technology", in Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), vol. 4, pp. 4072-4075, 2002.

[12] T. Kinnunen, and H. Li, "An Overview of Text-Independent Speaker Recognition: from Features to Supervectors", Elsevier Speech Communication, vol. 52, pp. 12-40, 2010.

[13] D. A. Reynolds, and R. C. Rose, "Robust Text-Independent Speaker Identification Using Gaussian Mixture Speaker Models", IEEE Transactions on Speech and Audio Processing, vol. 3, pp. 72-83, 1995.

[14] F.-Z. Zeng, and H. Zhou, "Speaker Recognition Based on a Novel Hybrid Algorithm", Procedia Engineering, vol. 61, pp. 220-226, 2013.

[15] C. Jeyalakshmi, V. Krishnamurthi, and A. Revathi, "Speech Recognition of Deaf and Hard of Hearing People Using Hybrid Neural Network", Mechanical and Electronic Engineering (ICMEE), Kyoto, pp. 83-87, 2010.

[16] P. K. Nayana, D. Mathew, and A. Thomas, "Comparison of Text Independent Speaker Identification Systems Using GMM and i-Vector Methods", Procedia Computer Science, vol. 115, pp. 47-54, 2017.

[17] V. Chauhan, Sh. Dwivedi, P. Karale, and S. M. Potdar, "Speech to Text Converter Using Gaussian Mixture Model (GMM)", International Research Journal of Engineering and Technology (IRJET), vol. 3, pp. 160-164, 2016.

[18] D. A. Reynolds, "Automatic Speaker Recognition Using Gaussian Mixture Speaker Models", IEEE Transactions on Speech and Audio Processing, vol. 3, pp. 1738-1752, 1995.

[19] E. Fedorov, V. Lukashenko, T. Utkina, A. Lukashenko, and K. Rudakov, "Method for Parametric Identification of Gaussian Mixture Model Based on Clonal Selection Algorithm", CEUR Workshop Proceedings, vol. 2353, pp. 41-55, 2019.

[20] E. E. Fedorov, and V. J. Larin, "Combination of PNN Network and DTW Method for Identification of Reserved Words, Used in Aviation during Radio Negotiation", Radioelectronics and Communications Systems, vol. 57, no. 8, pp. 362-368, 2014.

[21] L. R. Rabiner, and B. H. Jang, Fundamentals of Speech Recognition. Englewood Cliffs, NJ: Prentice Hall PTR, 507 p., 1993.

[22] J. D. Markel, and A. H. Gray, Linear Prediction of Speech. Berlin: Springer Verlag, 288 p., 1976.

IEEE Third International Conference on Data Stream Mining & Processing
August 21-25, 2020, Lviv, Ukraine

KLN: A Deep Neural Network Architecture for Keypoint Localization

Petr Hurtik
Institute for Research and Applications of Fuzzy Modeling
Centre of Excellence IT4Innovations, University of Ostrava
Ostrava, Czech Republic
petr.hurtik@osu.cz

Oleksii K. Tyshchenko
Institute for Research and Applications of Fuzzy Modeling
Centre of Excellence IT4Innovations, University of Ostrava
Ostrava, Czech Republic
oleksii.tyshchenko@osu.cz

***Abstract*—Localization of keypoints on pixel-level precision is an essential step for stitching panoramic images because the keypoints are matching, and their locations are used for computing stitching transformation. We recall the main standard computer vision techniques for keypoint localization and focus on the precise localization. We design a neural network architecture containing an encoder, a latent representation handler, and a decoder, where the encoder is motivated by SIFT. In contrast to domain-agnostic neural network architectures, the developed encoder reflects the scale-space construction as well as the difference of Gaussians estimation used in SIFT. In the benchmark, we show that our architecture has a higher number of keypoints localized with pixel-level precision than other standard and neural network-based approaches.**

***Index Terms*—keypoint detection, neural networks, deep learning, panoramic images, SIFT**

I. Problem statement

In computer graphics, feature point (keypoint) localization and description belong among the essential methods used for various purposes such as object recognition [1], object tracking [2], or for stitching panoramic images [3]. Since the probable applicability of these tools will deal with huge amounts of data and possibly be processed on mobile devices with restricted computational sources, there is a thriving necessity for local descriptors to be quick to match, rapid to compute and memory efficient. The keypoint descriptor is a scheme for obtaining feature vectors that represent local regions (patches) of an illustration. Besides that, these feature vectors are supposed to be both scale-invariant and invariant to rotation, illumination, and translation. For the most part largely, the mentioned descriptors are beneficial for pairing objects and matching patches between pictures. Considering object recognition/tracking, it is important to create descripting vectors of keypoints to be easily matched with high precision while the precision of keypoints' localization is irrelevant. On the other hand, when a nice panoramic image has to be produced, the pixel-level precision of keypoint localization is essential. The reason is that the corresponding keypoint coordinates are used for computing homography (it is usually enough to have at least three keypoints to define the homography matrix) that transforms one image into the form of the second one. Errors in keypoint localization lead to a non-seamless output, as is shown in Figure 1. **In this study, we address the problem of the keypoint localization with the emphasis on the production of the localization with pixel-level of precision.**

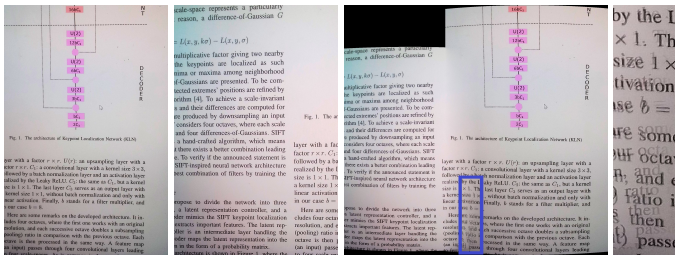


Fig. 1. The illustration shows why the pixel-level precision of keypoint localization is crucial. The two most left images are two images taken by a camera. The next image is output composed by Autostitch (software utilizing SIFT), where the keypoints were matched correctly, but because of not ideal localization precision, it includes artifacts. The marked blue part is magnified in the most right image. See the best zoomed-in in an electronic version.

Our contributions are as follows:

- We propose KLN, a novel neural network architecture that uses encoder motivated by SIFT keypoint localization scheme.
- We produced a dataset of 29 thousand images with corresponding labels and proposed an evaluation criterion.
- We trained NN-based approaches to detect keypoints and realized a stability comparison with KLN and classic algorithms.

II. Current state

The actual state can be divided into two groups: classical methods that include SIFT, SURF, ORB, etc., and methods based on neural networks. The former algorithm of the first group is SIFT [1] (Scale-Invariant Feature Transform), which detects corresponding points in every particular picture. The method holds a number of benefits like the fact that specific features can be matched to a broad set of targets; multiple traits can be produced for even tiny objects; since features are local, they are quite robust to obstruction. On the other hand, it is quite slow and computationally intensive; it does not provide real-time performance.

The real-time performance has been proposed by the SURF scheme (Speeded Up Robust Features) [4], where the main

concern of the SURF strategy is its high-speed calculation of operators applying box filters for tracking and object recognition. The additional speed-up was introduced by FAST (Features from accelerated segment test) [5] that may be utilized for the extraction of characteristic details (corners) and later adapted for objects' tracking and mapping in various problems of computer vision. The trade-off for the speed is that the method does not identify corners on computer-generated illustrations to be ideally adjusted to axes. In [6], binary strings are used as an effective keypoint descriptor (BRIEF). It is extremely fast for both design and comparison. BRIEF transforms image patches into a binary feature vector in such a way that collectively they can draw an object. Nevertheless, BRIEF works with an illustration at the pixel level, and consequently, it is highly noise-sensitive. Binary Robust Invariant Scalable Keypoints (BRISK) [7] should be discussed in the framework of dealing with the rotation invariance, noise, and scale transformation. It builds the trait descriptor of a local illustration through the grayscale connection of arbitrary point couples in the vicinity of a local icon and gets a binary characteristic descriptor. Contrasted to the conventional analogs, a comparable pace of BRISK is quicker, and the storage memory is lower, but the BRISK's robustness is diminished. The last algorithm to be mentioned in this context is ORB (Oriented FAST and Rotated BRIEF) [8]. It starts with the detection of specific areas (keypoints, which are highly-distinctive locations) in an icon. ORB functions as well as SIFT in the assignment of feature discovery (and is more reliable than SURF) while being nearly two orders of magnitude faster. ORB grounds on the FAST keypoint detector and the BRIEF descriptor.

It is quite understandable that a wave of highly significant outcomes by Deep Convolutional Neural Networks has made a crucial influence on the domain of Computer Vision. The deep structures described in [9]–[11] were introduced for obtaining feature descriptors. At the same time, attention tools also represent a significant position and usually train salient features in an unsupervised fashion. Spatial Transformer Networks (STN) [12] are exercised for the identification of an arranged quantity of likely patch matches. The network tries to discover and compare patches concurrently, with particularly weak-supervision of match/no-match tags on a picture level. A different path to produce area proposals is Region Proposal Networks (RPN) [13] that detect ranges of the image space that include objects. RPNs are frequently learned in an entirely supervised fashion.

Typical common strategies handle the position coordinates of every keypoint for training a pattern to identify keypoints. The main lack of this procedure is scale-dependence, which implies that the model is perceptive to the inputs' scale and cannot ensure running on various scale inputs. For this reason, a method suggested by [14] to defeat the claimed effect. In this paradigm, a confidence map is designed for specific keypoints and later employed as the ground truth for learning a fully convolutional neural network. This strategy helps make the trained model scale-invariant.

III. KLN: A NEURAL NETWORK ARCHITECTURE FOR LOCALIZING KEYPOINTS

As we state in Section *Current state*, there exist neural network architectures able to make the pixel-level precision. They work on the general principle that decreases spatial resolution while feature resolution is increased. That also leads to neural networks with a huge number of parameters to be learned; for example, U-Net in our benchmark has 14M parameters distributed in the convolutional layers. That is opposite to standard approaches, where, e.g., SIFT needs only several convolutional filters. Our motivation is to help neural network training by an appropriate architecture construction that will reflect the SIFT principle.

For SIFT, David Lowe supposes [1] scale-space L defined as

$$L(x, y, \sigma) = G(x, y, \sigma) * I(x, y),$$

where I stands for an image function, $*$ represents convolution and G is a Gaussian kernel with an appropriate σ. Roughly speaking, a particular scale-space represents a particularly blurred image. For that reason, a difference-of-Gaussian (DOG) G is determined as

$$D(x, y, \sigma) = L(x, y, k\sigma) - L(x, y, \sigma),$$

where k is a constant multiplicative factor giving two nearby scale-spaces. Finally, the keypoints are localized as such points, where local minima or maxima among neighborhood points and DOG are presented. To be complete, let us note the detected extremes' positions are refined by the Brown&Lowe algorithm [15]. To achieve a scale-invariant algorithm, scale-spaces and their differences are computed for various octaves that are produced by downsampling an input image. In detail, SIFT considers four octaves, where each scale has five scale-spaces and four DOGs. SIFT works well, but it is a hand-crafted algorithm, which means that it is a chance that there exists a better combination leading to better performance. To verify if the announced statement is valid, we create the SIFT-inspired neural network architecture and search for the best combination of filters by training the network.

A. The architecture

In detail, we propose to divide the network into three parts: an encoder, a latent representation controller, and a decoder. The encoder mimics the SIFT keypoint localization functionality and extracts important features. The latent representation controller is an intermediate layer handling the features. The decoder maps the latent representation into the original resolution, which has a form of a probability matrix.

The proposed architecture is shown in Figure 2, where the meaning of the abbreviation is as follows. $P(r)$: a pooling layer with a factor $r \times r$. $U(r)$: an upsampling layer with a factor $r \times r$. C_1: a convolutional layer with a kernel size 3×3, followed by a batch normalization layer and an activation layer realized by the Leaky ReLU. C_2: the same as C_1, but the kernel size is 1×1. The last layer C_3 serves as an output layer with a kernel size 1×1, without batch normalization and only

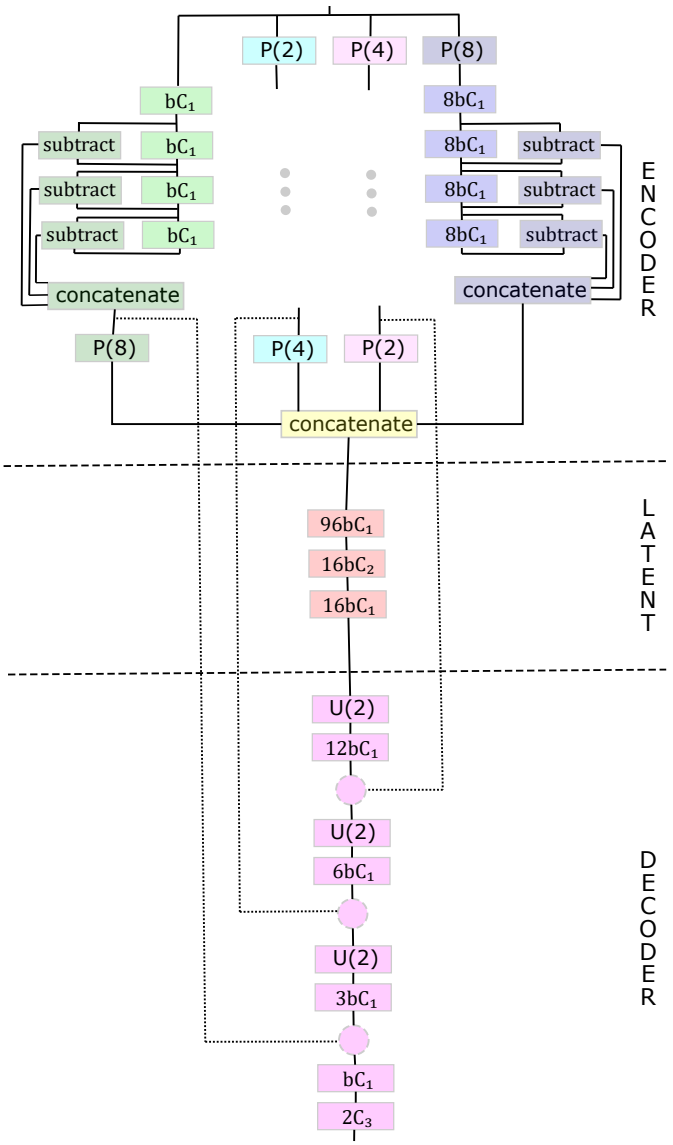


Fig. 2. The architecture of Keypoint Localization Network (KLN)

with linear activation. Finally, b stands for a filter multiplier, and in our case $b = 8$. The reference implementation is available through our GitLab[1].

Here are some remarks on the developed architecture. It includes four octaves, where the first one works with an original resolution, and each successive octave doubles a subsampling (pooling) ratio in comparison with the previous octave. Each octave is then processed in the same way. A feature map (an input) passes through four convolutional layers leading to four scale-spaces. As is common in the neural networks, for smaller spatial dimensions, i.e., a higher octave, more convolutional filters are used. Then, three subtractions are computed to reflect the SIFT's difference of Gaussians, and the outputs are concatenated. Finally, the concatenated output is downsampled in order to hold $r'r'' = 8$ for $P'(r')$ and $P''(r'')$, where $P'(r')$ stands for the pooling performed at the beginning of the block and $P''(r'')$ implies the pooling performed at the end of the block. The condition ensures that all the blocks have the same resolution and, therefore, can be concatenated into a tensor with dimensions of $c \times 3b(8+4+2+1) \times s/8 \times s/8$, where c is a batch size and s represents a size of a side of an input image.

In the latent representation control part, first of all, we use a higher number of filters which serve as an expansion layer, followed by a lower number of filters with a kernel size 1×1 which works as a compression layer forcing the neural network to focus only on the most important features. Finally, C_1 convolution follows and gives the final latent representation. Let us note that the latent representation control part is performed for feature maps with the small resolution ($s/8 \times s/8$), which allows us to have there a high number of the convolution filters, i.e., trainable parameters. It is possible without a bigger impact on computation speed because processing a small-resolution feature map is not so time-consuming.

The decoder phase follows the SOTA strategy [16]: a couple given by upsampling and convolutional layer is followed by a residual connection [17] until the original resolution is reached. The residual connection is realized by add operation. To match the two tensors to be added, we adjust the number of convolutional filters in the skip connection.

A possible general disadvantage of the neural networks is their non-interpretability. In Figure 3, we show how our neural network is functioning. It is evident, the network learns itself such filters, which are directional-sensitive to edges and which lead again to directional edges after subtraction. Let us recall, SIFT uses the Gaussian kernels in the form of DOG that is an omnidirectional edge detector [18]. That is reasonable enough when only one convolution per the scale space is performed. Here, the directional edge detector implemented by the network is a natural extension as it performs several convolutions per the scale-space presented by the convolutional layer.

B. Modularity of the architecture

The essential feature of the modern neural network architectures is modularity, i.e., the ability to establish a model with a various number of parameters. That can be observed, e.g., for the well-known VGG, ResNet, or EfficientNet. The variability is significant because it allows a user to use GPUs with various RAM for training, to create an NN capacity adequate to a dataset, or for deploying an NN to diverse powerful hardware, including mobile devices.

For the proposed architecture, there are three ways, how the number of parameters can be controlled.

- A coefficient b controls a number of the convolutional filters in all the layers.
- A number of the octaves that is a direct parallel to the number of pooling layers in Resnet etc. It generally holds that the bigger the image resolution is, the bigger the final pooling ratio (number of octaves) can be used. A change in a number of the octaves also affects a number of the upscaling blocks in the encoder.

[1]https://gitlab.com/PetrPH/kln

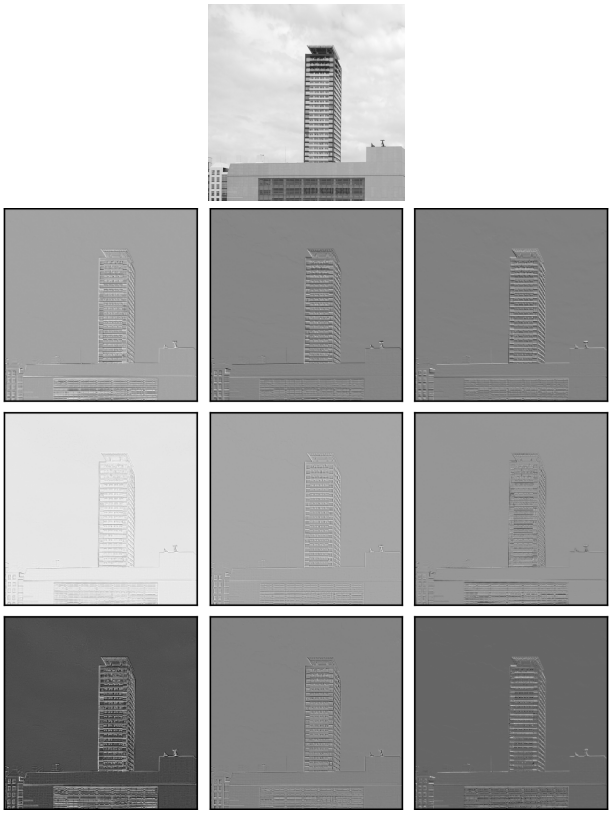


Fig. 3. From top: an input image; three feature maps produced by the first convolutional layer; by the second convolutional layer; produced by the subtract layer, where the two previous feature maps are inputs.

- A number of the convolutional layers that produce differences in the encoder. Here, only light adjustment is reasonable. The big increase may lead to the gradient vanishing and bigger sensitivity to initialization [19].

The reasonable combinations of the three ways and their impact on the number of parameters are proposed in Table I. The architectures KLN6 and KLN7 are too big to be trained on our hardware, but we mention them as they can be helpful from double-descent phenomena [20] point of view and can lead to even higher performance.

TABLE I
ARCHITECTURE MODULARITY

Version	Octaves	Differences	b	# Params
KLN1	2	3	2	50,160
KLN2	3	3	4	381,086
KLN3	4	3	8	3,055,098
KLN4	5	3	8	6,434,426
KLN5	5	5	8	10,794,234
KLN6	5	5	16	43,138,546
KLN7	5	5	32	172,477,410

IV. BENCHMARK

A. Dataset

To have a diverse dataset, we have involved three sources: private photos, images from the COCO dataset[2], and artificially generated images. From them, we have created 512×512px non-overlapping crops. In total, we collected 29388 crops and split them randomly into a training set (28000 crops) and a testing set (1388 crops). Due to the fact that we need labels for our supervised end-to-end training, we created the ground truth labels with the help of SIFT. The example of the dataset images from with their labels is shown in Figure 4.

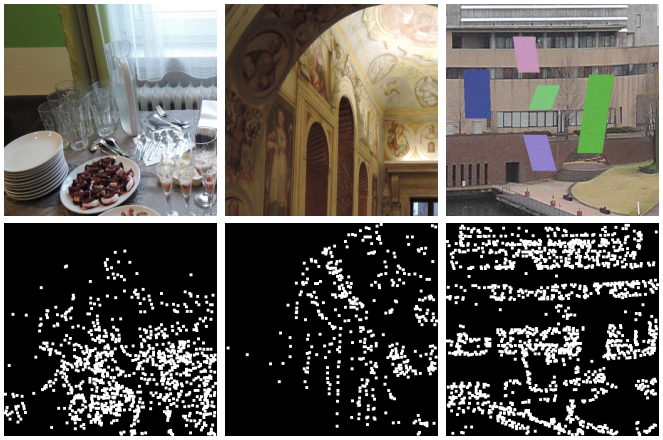


Fig. 4. Illustration of the images from the dataset (top) with their ground truth (bottom). Note: the ground truth images were dilatated in order to increase the visibility.

B. Used algorithms and their settings

To put our work into the SOTA context, we include into the benchmark three classical keypoint detection algorithms and three neural-network ones. Namely, we examined SIFT [1], ORB [8], and BRISK [7] as the standard ones and U-Net [16], LinkNet [21], and FPN [22] as the NN-based ones. In order to avoid possible bugs in source code, we use reference implementations for these methods. In detail, we use OpenCV implementations[3] for the standard methods and the Segmentation Model library[4] for the NN-based ones. We set the three neural networks (U-Net, LinkNet, FPN) from the segmentation models library to use Resnet18 [17] as a backbone (the encoder) in order to obtain a comparable number of parameters to our Keypoint Localization Network.

All the neural networks are trained using the same setting. Since the input resolution is quite large, we were forced to set the batch size to four; a bigger batch cannot be placed to the video memory of our graphic card (RTX 2080). We set a number of epochs to 30, as the dataset is reasonably big, and the small-batch size implies a lot of gradient updates per epoch. The optimizer is Adadelta [23] with $lr = 1.0$

[2]http://cocodataset.org/#home
[3]https://docs.opencv.org/3.4/d5/d51/group__features2d__main.html
[4]https://github.com/qubvel/segmentation_models

and the dynamic learning rate reduction on a plateau. Since the ratio of keypoints/non-keypoints is highly imbalanced, we use the binary focal loss [24]. The training time varies from approximately 8 hours (LinkNet) up to 17 hours (FPN). For completeness, the number of parameters of the three SOTA NNs is as follows: U-Net 14,340,715; LinkNet 11,521,835; FPN 13,816,523.

In order to avoid overfitting, online data augmentation by means of the Albumentations library[5] is realized. We use standard techniques: flips, transpose, rotations, intensity changes, and blurrings.

C. Evaluation criterion

The output of all the algorithms is $f : S \times S \to \{0, 1\}$, where $S = 512$ according to the dataset images. Then, 1 denotes the presence of a keypoint and vice versa. Further, let f be the keypoint detection output for an original image and f' be an output of the keypoint detection for a modified (rotated, blurred, etc.) image. Then, we propose an accuracy coefficient A to be defined as

$$A(f, f') = 1 - \frac{\sum_{x,y \in f} |f(x, y) - f'(x, y)|}{\sum_{x,y \in f} \min(1, f(x, y) + f'(x, y))}.$$

It is evident that the accuracy coefficient is ill-defined when no keypoint is detected. Because our dataset does not include fully homogeneous images, so we expect the presence of the keypoints in every image, we set A to zero for such cases.

D. Results

In Graphs 5–7, we show results for the particular transformations and in Graph 8 the overall performance, which is average of the three transformations. The graphs of transformations are computed as the mean of the following number of measurements. Flips: three (90, 180, 270 degrees), intensity change: four (0.7, 0.9, 1.1, 1.3), blur: four (Gaussian with kernels 3×3, 5×5, 7×7, 9×9).

Based on the overall performance, we can observe the NN-based approaches have significantly higher stability over the standard ones. Considering the standard methods, ORB yields the best performance. It excels for blurs, where it achieved similar performance to NN-based models. On the other hand, it suffers from flips, where only 5% of keypoints were detected with pixel-level precision. Also, it has the worst standard deviation from the standard methods ($\sigma = 0.28$), and from that point of view, SIFT with a slightly worse overall performance but a highly better std. dev. ($\sigma = 0.20$) may be a better choice.

Focusing on the NN-based detectors, KLN1 and KLN5 yield the best overall performance followed by KLN3 that reached the same result as U-Net. The lowest std. dev. has KLN3 ($\sigma = 0.151$) followed by KLN1 ($\sigma = 0.159$). The worst std. dev. has U-Net ($\sigma = 0.168$), which is still better than the best std. dev. of the standard methods.

Summarizing the results, **we can state that our proposed architecture strongly overperform stability of pixel-level keypoint localization of standard methods and slightly overperform the NN-based approaches while our KLN1 architecture needs 285× lesser parameters than U-Net**. Such a small model is important for mobile devices where memory is limited. From the computation time point of view, our solution takes from 54ms (KLN1), through 60ms (KLN3), to 78ms (KLN5) per single 512×512px image using an RTX2080Ti graphic card. That is on a par with SIFT that needs 66ms for the same image. Our computation can also be significantly speeded-up using batch processing.

[5]https://github.com/albumentations-team/albumentations

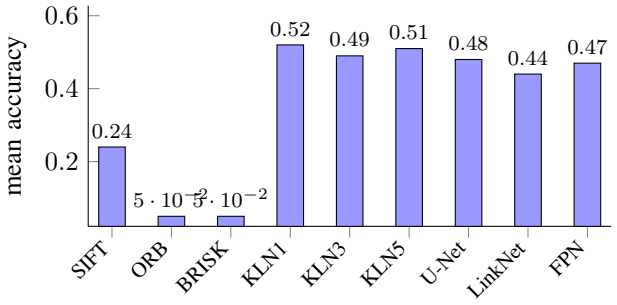


Fig. 5. Performance for 90° flips

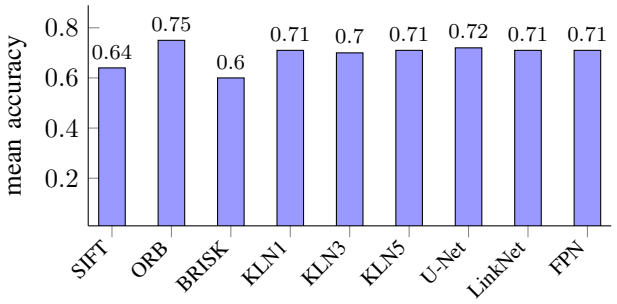


Fig. 6. Performance for intensity changes

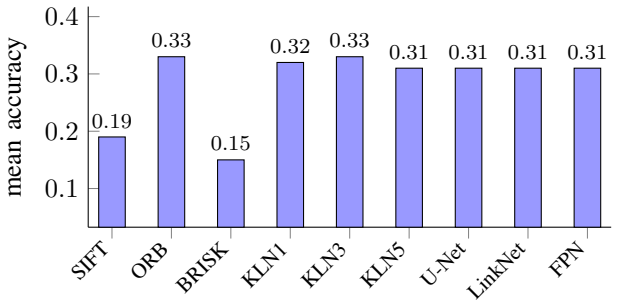


Fig. 7. Performance for blurs

V. Summary

We have tackled the problem of pixel-level precision keypoint localization that is necessary for producing seamless panoramic images. We have examined the SIFT keypoint localization procedure, and have used the part of establishing

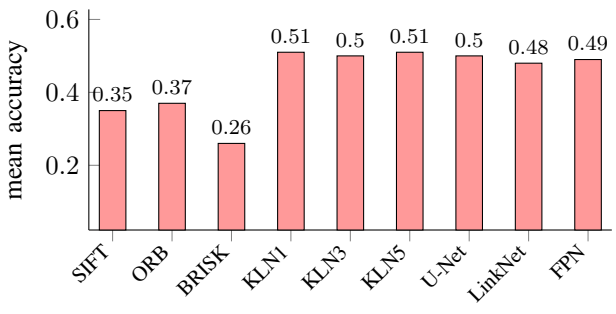


Fig. 8. Overall performance

scale-spaces and computing difference of Gaussians to design our own neural network encoder. The encoder is followed by a latent representation controller where the extracted features are combined together and finally by an encoder that transforms the abstract representation into the output one. The output representation gives us a probability matrix of being a keypoint. In the benchmark phase, we have involved both classical algorithms (SIFT, ORB, BRISK) and NN-based ones (U-Net, LinkNet, FPN) to be compared with our proposed KLN. The comparison relies on measuring accuracy that a certain pixel marked as a keypoint by a particular algorithm remains keypoint after a distortion and vice verse. We have tested three kinds of distortions: flips, intensity changes, and blurrings with the result that KLN yields the best stability with the lowest standard deviation among all the tested algorithms. Moreover, KLN has a significantly lower number of parameters than the other NN-based approaches. That shows that design problem-specific NN architecture may be highly beneficial in comparison with domain-agnostic architectures. Future work should be focused on mining information from the trained network in order to establish description vectors needed for matching keypoints among images.

ACKNOWLEDGMENT

The work is supported by ERDF/ESF "Centre for the development of Artificial Intelligence Methods for the Automotive Industry of the region" (No. CZ.02.1.01/0.0/0.0/17_049/0008414)

For more supplementary materials and an overview of our lab's work, please visit http://graphicwg.irafm.osu.cz/storage/pr/links.html.

REFERENCES

[1] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International journal of computer vision*, vol. 60, no. 2, pp. 91–110, 2004.

[2] G. Nebehay and R. Pflugfelder, "Consensus-based matching and tracking of keypoints for object tracking," in *IEEE Winter Conference on Applications of Computer Vision*. IEEE, 2014, pp. 862–869.

[3] M. Brown and D. G. Lowe, "Automatic panoramic image stitching using invariant features," *International journal of computer vision*, vol. 74, no. 1, pp. 59–73, 2007.

[4] H. Bay, T. Tuytelaars, and L. Van Gool, "Surf: Speeded up robust features," in *European conference on computer vision*. Springer, 2006, pp. 404–417.

[5] E. Rosten and T. Drummond, "Machine learning for high-speed corner detection," in *European conference on computer vision*. Springer, 2006, pp. 430–443.

[6] M. Calonder, V. Lepetit, C. Strecha, and P. Fua, "Brief: Binary robust independent elementary features," in *European conference on computer vision*. Springer, 2010, pp. 778–792.

[7] S. Leutenegger, M. Chli, and R. Y. Siegwart, "Brisk: Binary robust invariant scalable keypoints," in *2011 International conference on computer vision*. Ieee, 2011, pp. 2548–2555.

[8] E. Rublee, V. Rabaud, K. Konolige, and G. Bradski, "Orb: An efficient alternative to sift or surf," in *2011 International conference on computer vision*. Ieee, 2011, pp. 2564–2571.

[9] S. Zagoruyko and N. Komodakis, "Learning to compare image patches via convolutional neural networks," *arXiv preprint arXiv:1504.03641*, 2015.

[10] X. Han, T. Leung, Y. Jia, R. Sukthankar, and A. C. Berg, "Matchnet: Unifying feature and metric learning for patch-based matching," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2015, pp. 3279–3286.

[11] E. Simo-Serra, E. Trulls, L. Ferraz, I. Kokkinos, P. Fua, and F. Moreno-Noguer, "Discriminative learning of deep convolutional feature point descriptors," in *2015 IEEE International Conference on Computer Vision (ICCV)*. IEEE, 2015, pp. 118–126.

[12] M. Jaderberg, K. Simonyan, A. Zisserman, and K. Kavukcuoglu, "Spatial transformer networks," *arXiv preprint arXiv:1506.02025*, 2015.

[13] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," *arXiv preprint arXiv:1311.2524*, 2014.

[14] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh, "Realtime multi-person 2d pose estimation using part affinity fields," *arXiv preprint arXiv:1611.08050*, 2017.

[15] M. Brown and D. G. Lowe, "Invariant features from interest point groups." in *BMVC*, vol. 4, 2002.

[16] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*. Springer, 2015, pp. 234–241.

[17] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[18] L. Assirati, N. R. d. Silva, L. Berton, A. d. A. Lopes, and O. M. Bruno, "Performing edge detection by difference of gaussians using q-gaussian kernels," in *Journal of Physics: Conference Series*, vol. 490, no. 1. IOP Publishing, 2014, p. 012020.

[19] B. Hanin, "Which neural net architectures give rise to exploding and vanishing gradients?" in *Advances in Neural Information Processing Systems*, 2018, pp. 582–591.

[20] P. Nakkiran, G. Kaplun, Y. Bansal, T. Yang, B. Barak, and I. Sutskever, "Deep double descent: Where bigger models and more data hurt," *arXiv preprint arXiv:1912.02292*, 2019.

[21] A. Chaurasia and E. Culurciello, "Linknet: Exploiting encoder representations for efficient semantic segmentation," in *2017 IEEE Visual Communications and Image Processing (VCIP)*. IEEE, 2017, pp. 1–4.

[22] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 2117–2125.

[23] M. D. Zeiler, "Adadelta: an adaptive learning rate method," *arXiv preprint arXiv:1212.5701*, 2012.

[24] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2980–2988.

IEEE Third International Conference on Data Stream Mining & Processing
August 21-25, 2020, Lviv, Ukraine

Determining the Eligibility of Candidates for a Vacancy Using Artificial Neural Networks

Maksym Lupei
Department of Information Management Systems and Technologies,
Uzhhorod National University,
Uzhhorod, Ukraine,
maxim.lupey@gmail.com

Alexander Mitsa
Department of Information Management Systems and Technologies,
Uzhhorod National University,
Uzhhorod, Ukraine,
alex.mitsa@gmail.com

Igor Povkhan
Department of Sortware Systems,
Uzhhorod National University,
Uzhhorod, Ukraine,
igor.povkhan@uzhnu.edu.ua

Vasyl Sharkan
Department of Journalism,
Uzhhorod National University,
Uzhhorod, Ukraine,
vasyl.sharkan@uzhnu.edu.ua

***Abstract*—With the development of technology, new methods are being created to solve various tasks. Methods that use artificial neural networks (ANN) have become part of our routine and are very rapidly rooted in the Internet and computers. This work uses modern achievements in the ANN and methods of their training to solve the problem of correlations between objects, namely, vacancies and candidate information. The task of recruiting staff to expand or upgrade their staff has been and still is today. With the development of the Internet, this task has made new sense, because many resumes are scattered across different web services. So the question is: How do employers review and process information on all job applicants? This is what artificial neural networks will help, because with their help this process can be completed several times faster.**

Methods that use artificial neural networks (ANN) have become part of our routine and are very rapidly rooted in the Internet and computers. This work uses modern achievements in the ANN and methods of their training to solve the problem of correlations between the objects, namely a vacancy and information about a candidate. A vast array of data is processed and relations between the specific components are discovered. The suggested approach can be used not only for this specific task but also for a whole class of similar tasks like classification and pattern recognition

All these tasks can be solved with different ANN parameters which should be configured for every particular set of data. There are many ways to improves. From input text normalization and different vectorization algorithms to ANN topology, parameters, and training algorithms. Considered ANN and its parameters with chosen training algorithm works with high accuracy level only for tasks described in this paper. for every other task, there is a need to experimentally and theoretically proven other parameters and topology to reach high accuracy.

Keywords—Recruiting, Text analysis, Multilayer Perceptrons, Backpropagation.

I. INTRODUCTION

Nowadays, ANN (Artificial neural networks) only begins taking its place in recruitment and personnel management.

Other services that solve similar tasks to what are given in this article are costly and not always flexible. Information on the approaches used in such services is often secured.

The approach described in this article can be abstracted from a specific task and used to find correlations between any data, depending on how the network is trained and what input and output data are selected [1]. In this research, the best architectures [2] and training methods [3] will be determined to achieve maximum results.

This method can be used in online recruitment expert systems to make decisions in active recruitment or process candidates' applications. This system uses and processes publicly available candidates' text data [4] on LinkedIn and internal data of a recruitment agency. All data are anonymized and comply with personal data protection requirements.

In the world of globalization and information technology, people increasingly change their jobs and employers continually look for new people to implement their projects. Under such conditions, the automated search and recruitment systems as well as recruitment expert systems gain their popularity.

II. PROBLEM STATEMENT

The task of determining the best candidate for the job is to select the candidates who have the highest performance for certain parameters. To the input, we submit a dataset, which contains data about the skills we need, each of which has a weight. We identify the keywords from the resume and check if they are vacancies. On the way out we have to get the candidate who is most suitable for the vacancy. The goal is to minimize the learning criterion, which will produce high results.

There are different parameters, which can help to improve the accuracy of candidates eligibility for particular vacancy. On this paper, it will be started from candidate key skills, on next research, there will be added other parameters like candidate's social activity and other resume parameters. For example candidate most likely will be suitable for vacancy and accept job offer if he is actively searching for job. Also other aspects are valuable for recruiter decision, as work experience and hobbies. On next papers, it will be

interesting to add such parameters and see correlation between then and vacancy eligibility.

III. REVIEW OF THE LITERATURE

The problem of speeding up the recruitment problem is covered in many papers. One such article is Data mining to improve personnel selection and enhance human capital: A case study in the high-tech industry by Chen-Fu Chien and Li-Fei Chen [5]. Analyzing the article we can conclude. In order to find the best candidate, there are several parameters that can be divided into the following groups: demographics, channel - where you learned about your job, education and work experience. The research concluded that employees who got into the company on an internal channel (that is, they learned about work from their friends who already work there) are more likely to work better than those who got on another channel. Education and work experience also play an important role. Employees who had a college degree, as well as more work experience, performed more effectively than others. In the future, the author suggests increasing the amount of input. In [6], the authors have developed a system that uses a group of tests to minimize conflicts in the group being created.

In [7] developed a computer system that informs employers about the availability of one or more suitable candidates for the position. In addition to information from the resume, it allows you to collect information from other sources. Each evaluation feature requires a task and appropriate weights.

In [8] information was collected through a survey. The information obtained was used to assess the suitability of candidates for selection for various assignments. To do this, authors used fuzzy logic, which was based on a neural network approach. Another hybrid combination of expert systems and artificial intelligence was considered in [9]. In this work, the neural network learns to recognize information patterns in the employee's database, and a component of the expert system combines these results with analysis based on criteria that should give a result - to recommend or not to recommend. In [10], authors describe the implementation of fuzzy expert system (FES) tool for selection of qualified job applicants to minimize the rigour and subjectivity associated with the candidate selection process.

There are different papers and publications on this area, most valuable and recent results were reviewed and compared to current paper results, which made an experimental proof, that this paper results are highly valuable and interesting.

IV. MATERIALS AND METHODS

Let us consider the theoretical aspects of the approach described in this article. In this work, the multilayered perceptron (MLP) [11] is used, and the training algorithm with a teacher, which implements the function $f(\cdot): R^m \to R^1$ by training on a data set [12], where m is the dimension of the input data, and 1 is the dimension of the output. Having a set of input data:

$$X = (x_1, x_2, ..., x_m)^T \quad (1)$$

and output y, a nonlinear approximator of functions for classification or regression can be created. Such regression differs from the logistic, by the fact that there may be one or more nonlinear layers, called the hidden layers, between the input and output layers. Figure 1 shows 3 - layered MLP with the scalar output.

The input layer consists of a set of receptors $x_1, x_2, ..., x_m$, which accepts the input data. Each neuron in a hidden layer converts the value from the previous layer with a weighted linear summation:

$$w_1 x_1 + w_2 x_2 + ... + w_m x_m$$

with non-linear activation function:

$$f(\cdot): R^1 \to R^1$$

as hyperbolic tangent function. At the output, the value of the output signal y appears. As the training algorithm, we use the algorithm of the backpropagation [11]. Because we have a fixed training set:

$$\left\{\left(x^{(1)}, y^{(1)}\right), \left(x^{(2)}, y^{(2)}\right), ..., \left(x^{(M)}, y^{(M)}\right)\right\}$$

consisting of the M sets of input data, we can train our neural network using gradient descent. For this only training example (x, y) let us define the cost function for this particular example:

$$J(W, b; x, y) = \frac{1}{2}\left\| h_{W,b}(x) - y \right\|^2,$$

where W is synaptic weights, b is the displacement vector.

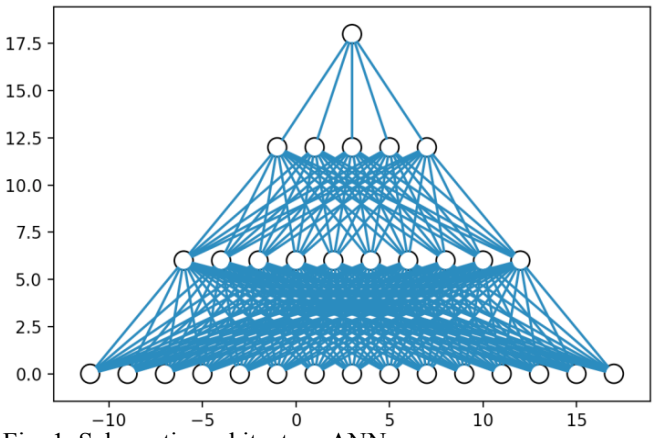


Fig. 1. Schematic architecture ANN

Taking into account the set of training examples, we define the general function of costs:

$$J(W, b) = \frac{1}{m}\sum_{i=1}^{m} J\left(W, b; x^{(i)}, y^{(i)}\right) +$$

$$+ \frac{\lambda}{2}\sum_{l=1}^{n_l - 1}\sum_{i=1}^{s_l}\sum_{j=1}^{s_{l+1}}\left(W_{ji}^{(l)}\right)^2 =$$

$$= \frac{1}{m}\sum_{i=1}^{m}\left(\frac{1}{2}\left\| h_{W,b}\left(x^{(i)}\right) - y^{(i)}\right\|^2\right) +$$

$$+\frac{\lambda}{2}\sum_{l=1}^{n_l-1}\sum_{i=1}^{s_l}\sum_{j=1}^{s_{l+1}}\left(W_{ji}^{(l)}\right)^2,$$

where λ — regularizing momentum term.

The first term in the definition $J(W,b)$ is the average square error indicator. The second term is a regularizing member that tends to reduce the weight and helps to prevent excessive training. Note that usually the decay weight does not apply to the offset terms $b_i^{(l)}$, as reflected in our definition for $J(W,b)$. The use of the regularizing member to the displacement elements does not significantly differentiate this criterion from the traditional quadratic one. Coefficient λ controls the relative importance of both members in the cost function.

The training criterion (1) can be used for both classification and regression. For classification, the training signal can take two values of 0 or 1 (in the case of using a sigmoidal activation function) or -1 and +1 in case of activation function in the form of a hyperbolic tangent. It is clear that the input data must first be scaled in the interval (0, 1) in case of a sigmoid or in the interval (-1, 1) in case of hyperbolic tangent.

The general purpose of training is minimization $J(W,b)$ as the function of W and in this case, before starting the training, all the parameters $W_{ij}^{(l)}$ and $b_i^{(l)}$ must be initialized in the form of a random variable close to zero (say, according to $Normal\left(0,\varepsilon^2\right)$ distribution for some small ε, let us say 0,01). As an algorithm of training, the optimization algorithm can be used, for example, a recursive gradient descent, because $J(W,b)$ is generally a non-convex function.

One iteration of the gradient descent can be written as follows:

$$W_{ij}^{(l)} = W_{ij}^{(l)} - \alpha\frac{d}{dW_{ij}^{(l)}}J(W,b),$$

$$b_i^{(l)} = b_i^{(l)} - \alpha\frac{d}{db_i^{(l)}}J(W,b),$$

where α— learning rate parameter.

In order to write down the backpropagation algorithm [5], it is necessary to enter all partial derivatives of the cost function using the synaptic weights in the form:

$$\frac{d}{dW_{ij}^{(l)}}J(W,b) =$$

$$=\frac{1}{m}\sum_{i=1}^{m}\frac{d}{dW_{ij}^{(l)}}J\left(W,b;x^{(i)},y^{(i)}\right)+\lambda W_{ij}^{(l)}$$

$$\frac{d}{db_i^{(l)}}J(W,b)=\frac{1}{m}\sum_{i=1}^{m}\frac{d}{db_i^{(l)}}J\left(W,b;x^{(i)},y^{(y)}\right).$$

Introducing into consideration δ-errors for i-th neuron of n_l layer in the form (2):

$$\delta_i^{(n_l)} = \frac{d}{dz_i^{(n_l)}}\frac{1}{2}\left\|y-h_{W,b}(x)\right\|^2 = -\left(y_i - a_i^{n_l}\right)\cdot f'\left(z_i^{(n_l)}\right) \quad (2)$$

or what is the same (3):

$$\delta_i^{(n_l)} = \left(\sum_{j=1}^{s_{l+1}} W_{ji}^{(l)}\delta_j^{(l+1)}\right)f'\left(z_i^{(l)}\right) \quad (3)$$

and calculating partial derivatives (4):

$$\frac{d}{dW_{ij}^{(l)}}J(W,b;x,y) = a_j^{(l)}\delta_i^{(l+1)},$$

$$\frac{d}{db_i^{(l)}}J(W,b;x.y) = \delta_i^{(l+1)} \quad (4)$$

we can finally write procedure of parameters tuning:

$$f'\left(z_i^{(l)}\right) = a_i^{(l)}\left(1-a_i^{(l)}\right).$$

We also note that the quality of the learning process is substantially influenced by a well-founded choice of parameters of the algorithm and which are usually chosen on a case-by-case basis, based on empirical considerations [13].

V. EXPERIMENTS

The problem solution of the optimal choice of candidates with the help of the neural network is provided as a result of the implementation of the sequence of iterations:

1. As for example, export data with the key candidate's skills and the fact of agreement to be interviewed for the vacancy of a JavaScript developer.
2. Vectorize the input information (translate the text into numeric binary arrays).
3. Perform the normalization of input vectors.
4. Train ANN.
5. Analyze the results.

TABLE I. DESCRIPTION OF THE EXECUTION ENVIRONMENT

Programming language	**Python**
Programming libraries	Numpy, Matplotlib, Scikit-Learn, Pandas.
IDE	Pycharm Edu

TABLE II. ANN PARAMETERS

The structure of the neurons and layers of ANN	[1000 : 100 : 5 : 2 : 1]
Random state	1
Solver	lbfgs
Alpha	1e-5
Converter	vectoriseText
Activation function	tanh
Max. iterations	200

VI. RESULTS

After this computational experiment is performed, we can see how using the database of the recruitment agency, a recommendation mechanism can be created to simplify the recruiter's work. Table 2 shows the effectiveness of ANN with a different number of layers and input data. The analysis shows that for a sample of 4,000 candidates with a division between education and testing - 2 (2-fold cross-validation) the most effective is the structure of the neurons of the ANN [1000:100:5:2:1] in the 2-dimensional networks [12]. The total length of keywords is 300, the activation function is the hyperbolic tangent, the maximum number of iterations is 200, the calculated average accuracy is 0.9257, and the error f1 0.8126. It is proved experimentally that such network parameters at this particular sample are the most effective compared to smaller dimensional neural networks and their other parameters specified in Table II-III.

TABLE III. COMPARISON OF RESULTS

Exp. #	k	The structure of ANN		The error of the results	
		The structure of the neurons of ANN	Max. itera-tions	accuracy	f1 score
1	2	[1000:100:5:2:1]	300	**0.9245**	**0.8100**
2	2	[1000:10:5:2:1]	200	**0.9213**	**0.8022**
3	2	[1000:10:5:3:2:1]	200	**0.9162**	**0.7947**
4	2	[400:100:5:2:1]	200	**0.9023**	**0.7550**
5	2	[10000:100:5:2:1]	200	**0.9251**	**0.8082**
6	2	[100:100:5:2:1]	200	**0.8915**	**0.7350**
7	2	[2000:100:5:2:1]	200	**0.9212**	**0.8034**
8	2	[1000:100:10:5:2:1]	200	**0.9111**	**0.7777**
9	2	[1000:15:5:2:1]	200	**0.9233**	**0.8060**
10	2	[1000:100:5:2:1]	200	**0.9257**	**0.8126**

VII. DISCUSSION

The results of this research can be applied in practice to large companies that often have vacancies. The developed method of determination gives high indicators of accuracy, which positively influences the correct selection of candidates. Compared to other methods, this method also takes into account certain characteristics of the candidates (usually between 10 and 20) and results from them. Further research will help to increase the accuracy and correctness of the selection of candidates suitable for this position.

VIII. CONCLUSION

Typically, a recruiter when selecting a candidate is guided by a set of candidate's keywords as the priority indicator to determine the level of suitability for a particular job. Automation of this work along with the correct assessment of the considered parameters helps to improve the efficiency of the recruitment agency significantly. The conducted computational experiment shows that the developed approach and software product can be used for real recruitment systems (Table II). The next step is to increase the number of inputs that may correlate with the candidate's suitability for a particular vacancy, which should significantly improve the output results. It should be mentioned that the application of more complicated activation functions (e.g. proposed in [14, 15]) instead of hyperbolic tangent can slightly improve the performance of the ANN.

REFERENCES

[1] Y. Bodyanskiy, "Computational Intelligence Techniques for Data Analysis" in Leipziger Informatik-Tage, 2005, pp. 15-36.

[2] M. Lupei, A. Mitsa, V. Repariuk, V. Sharkan, "Identification of authorship of Ukrainian-language texts of journalistic style using neural networks," Eastern-European Journal of Enterprise Technologies, vol. 1 (2 (103)), pp. 30–36.

[3] V. Kotsovsky, F. Geche, and A. Batyuk, "Finite generalization of the offline spectral learning," in Proceedings of the 2018 IEEE Second International Conference on Data Stream Mining & Processing, DSMP 2018. Lviv, Ukraine, 21–25 August 2018, pp. 356–360.

[4] D. Sarkar, "Text Summarization," Text Analytics with Python, Apress, Berkeley, CA, 2016, pp. 217-263.

[5] C. F. Chien, L. F. Chen, "Data mining to improve personnel selection and enhance human capital: A case study in high-technology industry," Expert Systems with applications, 2008, vol. 34(1), pp. 280-290.

[6] M. Nussbaum et al. "Decision support system for conflict diagnosis in personnel selection." Information & Management 36.1 (1999): 55-62.

[7] D. Hardtke et al. "Identifying candidates for job openings using a scoring function based on features in resumes and job descriptions." U.S. Patent Application No. 13/662,312.

[8] N. Goonawardene et al. "A neural network based model for project risk and talent management." International Symposium on Neural Networks. Springer, Berlin, Heidelberg, 2010.

[9] Labate, Fran, and Larry Medsker. "Employee skills analysis using a hybrid neural network and expert system," in Proceedings IEEE International Conference on Developing and Managing Intelligent System Projects. IEEE, 1993.

[10] Daramola, J. Olawande, Olufunke O. Oladipupo, and A. G. Musa. "A fuzzy expert system (FES) tool for online personnel recruitments." International Journal of Business Information Systems 6.4 (2010): 444-462.

[11] S. Haykin, Neural networks and learning machines. Third Upper Saddle River, NJ: Pearson Education, 2009.

[12] Y.Bodyanskiy et al. “Deep 2D-Neural Network and its Fast Learning,” in Proceedings of the 2018 IEEE Second International Conference on Data Stream Mining & Processing, DSMP 2018. Lviv, Ukraine, 21–25 August 2018, pp. 519-523.

[13] Y. Rashkevych, D. Peleshko, and M. Pasyeka, “Optimization search process in database of learning system,” in Proceedings of the 2nd IEEE International Workshop on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications, IDAACS 2003. Lviv, Ukraine, 8-10 Sept. 2003, pp. 358-361.

[14] V. Kotsovsky, F. Geche, and A. Batyuk, “On the computational complexity of learning bithreshold neural units and networks,” in: Advances in Intelligent Systems and Computing, vol 1020., V. Lytvynenko. et al., Eds., Heidelberg: Springer, 2020, pp. 189–202.

[15] V. Kotsovsky, F. Geche, and A. Batyuk, “Artificial complex neurons with half-plane-like and angle-like activation function,” in Proceedings of the International Conference on Computer Sciences and Information Technologies, CSIT 2015. Lviv, Ukraine, 14–17 September 2015, pp. 57–59.

IEEE Third International Conference on Data Stream Mining & Processing
August 21-25, 2020, Lviv, Ukraine

Training Neural Network Over Encrypted Data

Vojtech Molek
Institute for Research and Applications of Fuzzy Modeling
Centre of Excellence IT4Innovations, University of Ostrava
Ostrava, Czech Republic
vojtech.molek@osu.cz

Petr Hurtik
Institute for Research and Applications of Fuzzy Modeling
Centre of Excellence IT4Innovations, University of Ostrava
Ostrava, Czech Republic
petr.hurtik@osu.cz

***Abstract*—We are answering the question whenever systems with convolutional neural network classifier trained over plain and encrypted data keep the ordering according to accuracy. Our motivation is need for designing convolutional neural network classifiers when data in their plain form are not accessible because of private company policy or sensitive data gathered by police. We propose to use a combination of fully connected autoencoder together with a convolutional neural network classifier. The autoencoder transforms the data info form that allows the convolutional classifier to be trained. We present three experiments that show the ordering of systems over plain and encrypted data. The results show that the systems indeed keep the ordering, and thus a NN designer can select appropriate architecture over encrypted data and later let data owner train or fine-tune the system/CNN classifier on the plain data.**

***Index Terms*—image permutation, image data encryption, neural network, image classification**

I. Problem statement

In image processing, the task of image classification, i.e., assigning a particular label to an image with the usage of the image attributes/features, became a well-solved task due to deep neural networks [1]. Currently, as the result of the continual development of efficient neural network architectures such as ResNet [2], SENet [3], or EfficientNet [4], the performance of the artificial systems surpass the abilities of a human. That is helpful in industry, especially in optical quality control systems, where machines replace humans in highly repetitive processes [5]. The advantages are evident: increased processing rate up to hundreds of classifications per second [6], improved accuracy of classification, and replacing subjective evaluation with objective one.

The problem that remains is data that are necessary to design appropriate neural network architecture and which are used for its training. If we consider a standard use case scenario where an industrial partner collects data and transfer it to a NN-developer, the industrial partner risks a potential data leak. Some data are even forbidden to transfer due to company policies or the law, e.g., sensitive data such as captured faces of people. The solution is to encrypt the data and design the network on top of it. Then, after the network is designed/pre-trained, it can be safely given to the data provider to fine-tune the network on the original data. However, several conditions have to be met to create such a process successfully:

Fig. 1. The illustration shows plain images from the CIFAR-10 dataset in the top row. The bottom row contains the images from the top row, where all values have been permuted. The bottom row presents the encrypted data that are uninterpretable for a human.

- Encrypted data have to be uninterpretable by the NN designers. For the illustration of how the encrypted data should look like, see Figure 1.
- The encryption scheme is known only to the data provider.
- The neural network performance has to be the same for both encrypted/plain data.
- The system should have two stages – the first stage transforms encrypted data into a suitable form of the second stage, and the second classifies them. The former stage should be removable for the usage on o the plain data.

As it is evident, there is no assumption of reaching a certain level of accuracy on the encrypted data. That is because we expect the encryption process to decrease the NN accuracy significantly. That is not an obstacle as we assume maximum accuracy after removal of the stage transforming encrypted data. Such a scheme allows a NN-designer to choose the proper architecture, improve it, and transfer to a data provider for fine-tuning on plain data.

Our contributions are as follows:

- We describe various image permutation schemes and their suitability for our system.
- We propose a system of two networks trained together for handling the encrypted (permuted) data.
- We experimentally prove that ordering of our system according to accuracy, is identical on plain and encrypted data.

978-1-7281-3214-3/20/$31.00 ©2020 European Union

II. Current state

The motivation for our work came from permuted MNIST [7]. Permutation here creates a dataset which contains image samples that would appear to human as noise, but in fact, there is no noise, the pixels are just permuted. The permutation scheme remains the same throughout the dataset, so a fully-connected NN would be able to learn such permutation partially. Initially, the permuted MNIST was used for continual training [8] of an NN. The research of permuted MNIST has been further extended into permuted CIFAR-10, where the efficiency of fully connected (FC) versus convolution architectures was examined [9]. The finding is that convolutional architectures suffer from permutations as the local 2D spatial data structures are broken and cannot be reasonably reconstructed using the convolutional filters. The FC architectures omit spatial information, so their performance remains unchanged. That is limiting as convolutional-based architectures are generally more useful as they reach higher accuracies than FC ones. The benefits of using a permutation as a data encryption scheme are: the encryption process is swift, the output is uninterpretable for a human, and the global information (e.g., global histogram) remains the same.

The different way, how to encrypt data is to use *classical* encryption schemes known from cryptography. For example, CryptoNets [10] encrypt images using homographic encryption (HE), where each pixel of an image is encoded by five polynomials. The disadvantage is that the global information of an encrypted image is not equal to the global information of plain images. So, tuning a network and increasing its performance over encrypted data will not necessarily lead to better performance over plain data. Another use of classical algorithms (*AES* [11] and *DES* [12]), in combination with extreme learning machine (ELM) is described by Wang et al. [13]. Their results show that DES encryption leads to better testing accuracy than AES, and multi-layer ELM is on par with classical CNN. A novel way to encryption itself shows Tanaka [14]. In his work, Tanaka uses CNN to downsample input images in $M \times M$ blocks and then upsample again to receive the encrypted representation of the original image. These are used for training the classifier (CIFAR-10/100). Sirichotedumrong [15] uses XOR to flip the color values of pixels. The pixels are chosen according to generated secret keys. Moreover, the color channels can be permuted, as stated by authors. Finally, images are augmented with shits and flips. Such images are then preprocessed by *adaptation network* (CNN with 1×1 filters) and classified. The results are reported on CIFAR-10, using ResNet-18.

EPIC [16] demonstrates an elegant, simple, cloud-based solution to encrypted image classification using transfer learning. The image owner uses one of the pre-trained CNNs to obtain generic feature representation and send it to the server for the classification. The server classifies feature representation and sends results back. Since feature representation cannot be used by the server to obtain the original image, privacy is maintained. A similar motivation of defense against model-inversion in the healthcare environment presents Wu [17]. The author introduces a new variant of stochastic gradient descent that preserves patient privacy. Their optimizer is tested with several traditional and modern CNN on 1216 patients. The more traditional way, using random forests with two levels of encryption, was introduced in [18]. Compress sensing encryption with random matrices for each forest is used to encrypt feature vectors on the first level. After classification with forests, the label information is again encrypted (the second level) and send back to the user. Therefore, the classifier has no access to the full result of classification. The authors mention a speed difference between the proposed approach and HE encryption, $\approx 10^5$ speedup.

Above mentioned researches have a similar topic of privacy and encryption; however, none of them quite tackle our use case. Our pivotal motivation is to answer the question if neural networks trained on encrypted and plain data keep the ordering with respect to testing accuracy. The order ensures that satisfactory results obtained on plain data will be reached on the encrypted data as well.

III. Working with secured data

In this section, we discuss different ways to permute data, used scheme, and its parts. We briefly describe a combination of a fully connected autoencoder with CNN architecture and the potential weaknesses. Finally, we formulate a hypothesis that the ordering of networks trained on secured (permuted) data with respect to accuracy stay the same on the plain data.

A. Permuting the data

In the literature, we can find three methods, how image data can be permuted:

1) Coordinates of all pixels across all channels are permuted.
2) An image is processed channel-by-channel, and all pixels inside a single channel are permuted.
3) An image is divided into non-overlapping blocks that fully coverage of the image. Then, pixels inside a block are permuted separately by using the first or second method.

The first method uses the whole domain, so the combinatorial complexity is the highest, and therefore, the security is highest too. The second method preserves the most dominant colors, so if an image is mainly green before the permutation, it will be mainly green after the permutation too. This color-preservation may be considered an information leak. In the special case of the grayscale image, the first and second methods are equal. The motivation for the third method is to preserve local spatial relations partially, so a fully convolutional network can adapt to such conditions. On the other hand, data encrypted by block permutation can be visually interpretable for a human. Because the first method is the safest, we will use it further in this work. The visual demonstration of the mentioned approaches is shown in Figure 2.

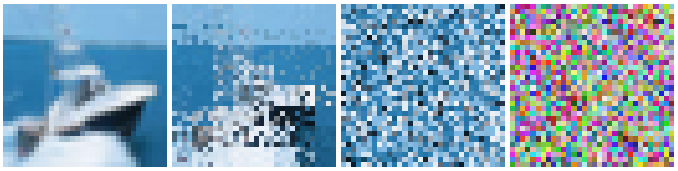


Fig. 2. A comparison of various permutation approaches for image encryption. From left: original image. Permuted values inside non-overlapping blocks. Permuted values inside a particular channel. All values permuted.

B. Neural network architecture

We propose to use a two-stage system where the first stage transforms the encrypted data, and the second one performs classification. The difference between transformation and decryption term is that we suppose the data are confidential, and their full decryption and visualization can be illegal. By transforming the data from the given space into a new space, we keep the non-interpretability for humans but allow the convolutional filters in the second stage to keep the necessary locality for their functionality.

Based on the fact that fully connected layers are able to handle the permutation [9], we propose to use fully connected autoencoder-based architecture in the first stage and a SOTA CNN in the second stage. Let us remark; an autoencoder [19] is a special architecture that takes input data, reduces their dimensionality gradually until a latent representation is reached and then, projects the latent representation back into original form. An example of an autoencoder is shown in Figure 3. The advantage of an autoencoder is that it can be trained without labels. More precisely, the labels are given by the input data itself, and the autoencoder is trained to produce the same output data as are on the input, so it is forced to extract the most important/descriptive features. For visualization of the scheme, see Figure 3. In our application, the 2D data are reshaped into a 1D structure before the encoder and then transformed back after the decoder.

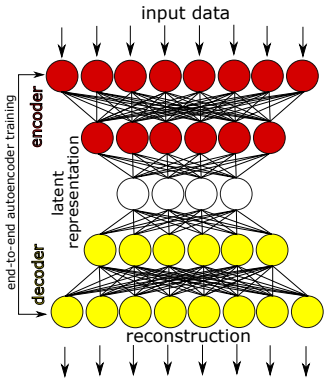


Fig. 3. Generic scheme of an autoencoder. Input data are fed in and encoded into a latent representation and subsequently the latent representation is decoded back to original data.

The weak point of our proposed scheme is a restriction for data augmentation. Generally, online augmentation *distorts* input data, so it is able to produce an almost infinite amount of sample variations. That helps neural networks to postpone the overfit and significantly increase achieved accuracy [20]. The data augmentations techniques for images can be divided into two logical groups: spatial augmentations and intensity augmentations. Into the first group belong image flipping, rescaling, shifting, or rotating. The second group consists of linear intensity change, channel shift, gamma modification, or blurs. The fully connected autoencoder is not able to handle the spatial augmentations while using permuted data. For example, the proposed scheme with ResNet50 classifier and spatial augmentations techniques yields an accuracy of 21.40% on CIFAR-10 dataset with secured images (i.e., it is almost not able to be trained); without the augmentation, the accuracy is 57.39%. Because our aim is to create a solution that will preserve ordering by accuracy for various neural networks trained over encrypted/plain data and disregarding the overfit, we will omit data augmentation in the following benchmark. Notwithstanding, this weak point shall be aimed at in future work.

IV. Benchmark

In this section, we show and discuss the results of training using three different use-cases. The use-cases cover different training setup and encryption of data. We show that the ordering, according to accuracy, is mostly preserved except for a single outlier. We briefly describe the used datasets and used hyperparameters.

A. Dataset

We use the well-known dataset, CIFAR-10 [21], which has been selected because of our experiments are realized in many epochs (see the following section), and the small image resolution helps us to keep a reasonable training time. CIFAR-10 includes 50000 training and 10000 testing images, which are colored, have a resolution of 32×32px, and capture various natural objects. It has ten classes where every single image belongs precisely to one of the classes.

B. Algorithms settings

The architecture of the first stage, i.e., fully connected autoencoder, is as follows: $3072 \rightarrow 1024 \rightarrow 512 \rightarrow 256 \rightarrow 512 \rightarrow 1024 \rightarrow 3072$, where the number expresses the number of neurons. 3072 was selected to be equal to the original resolution: $32{\cdot}32{\cdot}3$. The first and the last layers have a linear activation function, and the hidden layers use ReLU. Furthermore, because fully connected neural networks suffer from overfitting, each layer is followed by Dropout [22], where a specific fraction (0.3 in our case) of randomly selected neurons is turned off.

We test five SOTA neural networks used in the second stage: MobileNet [23], ResNet50 [2], MobileNetv2 [24], DenseNet [25], and InceptionResnetv2 [26]. All of them have the same setting. Namely, 60 epochs divided into 40 epochs with a full learning rate followed by 20 epochs with the learning rate reduced by factor 0.1, with AdaDelta [27] optimizer. Each training is repeated five times, and the highest

test accuracy is selected. In total, we processed 4500 epochs (5 NNs · 60 epochs · 5 repetitions · 3 data type – encrypted, plain and plain without first stage). The batch size is set to 50. For training, categorical cross entropy loss is used. For all convolutional classification models, we use their reference Keras implementation[1]. Both training and testing data are normalized and have subtracted mean.

C. Results

The main goal is to test the hypothesis that the behavior of a neural network in our system is not affected by the encryption of the data. In order to validate the claim, the three following cases have to be benchmarked:

1) Train both FC autoencoder and a convolutional classifier on the encrypted data. This scenario reveals if it is even possible to train a selected network over the encrypted data.
2) Train both FC autoencoder and a convolutional classifier on the plain data. Here, an ideal result should be equal to the first case from the accuracy point of view. That means that the two-stage system can perform identically on original and secured data.
3) Train only the convolutional classifier on plain data. Here, accuracy should significantly increase. The important point is that the ranking of various neural networks remains the same as in the first two cases. It means that if a network 'A' was more accurate than network 'B' for cases 1 and 2, it should also be more accurate in this case.

The results for the CIFAR-10 dataset are shown in Graphs 4–6. In the first graph, reflecting case 1, we can observe that the ordering is from the lowest accuracy to the highest as follows: MobileNet, ResNet50, MobileNetv2, DenseNet, and InceptionResnetv2. Here, we can also compare our results with the current state. In [9], authors reached the accuracy of 57.3% with CNN and 59.3% with FCN on permuted CIFAR-10, so their results are on par with ours. The difference is they use channel-permutations, i.e., each channel is permuted separately. In our case, we realize permutations through all channels, which makes the problem significantly more difficult for a NN. That is also the reason why they are able to use CNN – for example, images from class *ship* contain sea, so they are mostly blue. If only pixels are permuted, the image will be again blue, and it is trivial for a CNN to classify. In our case, the color distribution is not preserved and, therefore, stand-alone CNN (without FCN autoencoder) is able to reach only approx 10%, i.e., spatial patterns and color distribution is broken and cannot be trained. Therefore, we can say that our data are more strongly encrypted. Finally, they do not perform three cases as we propose, so the behavior of an NN on plain data is unknown.

The graph in Figure 5 shows the scenario when the two-stage system is trained on plain CIFAR-10 data. Here, two essential findings can be observed. The first finding is that

[1]https://keras.io/applications/

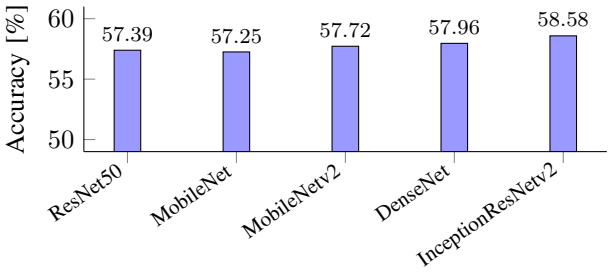


Fig. 4. CIFAR-10 dataset, performance on encrypted data by the two-staged networks.

the ordering of neural networks according to their accuracy is identical to the previous case, i.e., when the networks are trained on encrypted data. The second finding is that the accuracy of the neural networks is almost identical to the previous setting. Namely, the mean absolute error is 0.19, which is caused due to the negligible oscillations in results because of the random factor of the training process. That low mean absolute error means that the used FCN autoencoder extracts features that are invariant to permutation and are suitable for the convolutional network that follows.

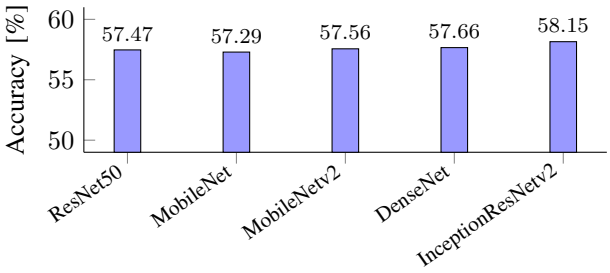


Fig. 5. CIFAR-10 dataset, performance on plain data by the two staged networks.

The last graph in Figure 6 shows the third case and reflects the use-case when a researcher creates/selects NN architecture with the usage of encrypted data and transfers it to an industrial partner. The partner then fine-tunes the network with the removed first stage (FC autoencoder) on plain data. As for the previous case, it holds that the order by accuracy has to be preserved. According to the graph, the condition is satisfied for all but DenseNet that is an outlier. Note, the reached accuracies are lower than those achieved by SOTA [2], which is caused by the fact that we do not use augmentation techniques (see Section III).

V. Summary

In this study, we have tackled the problem of training neural networks over encrypted image data for an image classification task. The problem is significant because there exist cases where the data cannot be transferred to a NN-designer due to the threat of data leak, or law forbidding it. The current state usually uses homographic encryption that is bandwidth-heavy and, therefore, not suitable for fast inferences. We

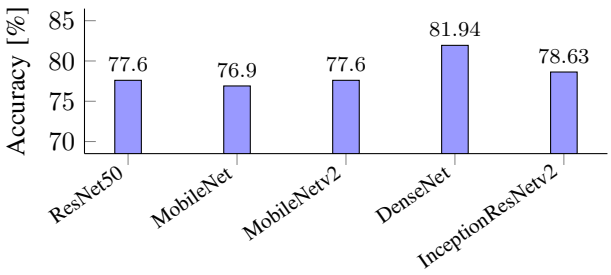


Fig. 6. CIFAR-10 dataset, performance on original data taken by networks without the first stage.

propose to use permuted data, which are uninterpretable for a human. Because the local spatial relations are not preserved in a permuted image, standard convolutional networks for classification cannot be used as they are not cannot to be trained. To solve the training issue, we propose a system that consists of two networks: fully connected autoencoder followed by a SOTA convolutional neural network; the two networks are trained together. The FC autoencoder transforms the permuted data into a suitable form for the CNN classifier. In the benchmark realized over the CIFAR-10 dataset, we have shown firstly that the proposed system is trainable over the encrypted data. Secondly, we have demonstrated that the accuracy is identical to accuracy reached by training over the plain data. Finally, we have proven the ordering of the neural networks according to the accuracy is preserved even for the case when the autoencoder is removed and the classifier trained over the plain data. Such a scheme allows a NN-designer to select the architecture that is ideal for the task with the usage of encrypted data, transfer the solution to the customer, and the customer does the final fine-tuning.

ACKNOWLEDGMENT

The work is supported by ERDF/ESF "Centre for the development of Artificial Intelligence Methods for the Automotive Industry of the region" (No. CZ.02.1.01/0.0/0.0/17_049/0008414)

For more supplementary materials and an overview of our lab's work, please visit http://graphicwg.irafm.osu.cz/storage/pr/links.html.

REFERENCES

[1] Q. Xie, E. Hovy, M.-T. Luong, and Q. V. Le, "Self-training with noisy student improves imagenet classification," *arXiv preprint arXiv:1911.04252*, 2019.

[2] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[3] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7132–7141.

[4] M. Tan and Q. V. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," *arXiv preprint arXiv:1905.11946*, 2019.

[5] P. Hurtik, M. Burda, and I. Perfilieva, "An image recognition approach to classification of jewelry stone defects," in *2013 Joint IFSA World Congress and NAFIPS Annual Meeting (IFSA/NAFIPS)*. IEEE, 2013, pp. 727–732.

[6] S. Bianco, R. Cadene, L. Celona, and P. Napoletano, "Benchmark analysis of representative deep neural network architectures," *IEEE Access*, vol. 6, pp. 64 270–64 277, 2018.

[7] R. Kemker, M. McClure, A. Abitino, T. L. Hayes, and C. Kanan, "Measuring catastrophic forgetting in neural networks," in *Thirty-second AAAI conference on artificial intelligence*, 2018.

[8] D. Lopez-Paz and M. Ranzato, "Gradient episodic memory for continual learning," in *Advances in Neural Information Processing Systems*, 2017, pp. 6467–6476.

[9] C. Ivan, "Convolutional neural networks on randomized data," *arXiv preprint arXiv:1907.10935*, 2019.

[10] R. Gilad-Bachrach, N. Dowlin, K. Laine, K. Lauter, M. Naehrig, and J. Wernsing, "Cryptonets: Applying neural networks to encrypted data with high throughput and accuracy," in *International Conference on Machine Learning*, 2016, pp. 201–210.

[11] J. Nechvatal, E. Barker, L. Bassham, W. Burr, M. Dworkin, J. Foti, and E. Roback, "Report on the development of the advanced encryption standard (aes)," *Journal of Research of the National Institute of Standards and Technology*, vol. 106, no. 3, p. 511, 2001.

[12] Des, "Data encryption standard," in *In FIPS PUB 46, Federal Information Processing Standards Publication*, 1977, pp. 46–2.

[13] W. Wang, C.-M. Vong, Y. Yang, and P.-K. Wong, "Encrypted image classification based on multilayer extreme learning machine," *Multidimensional Systems and Signal Processing*, vol. 28, no. 3, pp. 851–865, 2017.

[14] M. Tanaka, "Learnable image encryption," in *2018 IEEE International Conference on Consumer Electronics-Taiwan (ICCE-TW)*. IEEE, 2018, pp. 1–2.

[15] W. Sirichotedumrong, T. Maekawa, Y. Kinoshita, and H. Kiya, "Privacy-preserving deep neural networks with pixel-based image encryption considering data augmentation in the encrypted domain," in *2019 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2019, pp. 674–678.

[16] E. Makri, D. Rotaru, N. P. Smart, and F. Vercauteren, "Epic: Efficient private image classification (or: Learning from the masters)," in *Cryptographers' Track at the RSA Conference*. Springer, 2019, pp. 473–492.

[17] B. Wu, S. Zhao, G. Sun, X. Zhang, Z. Su, C. Zeng, and Z. Liu, "P3sgd: Patient privacy preserving sgd for regularizing deep cnns in pathological image classification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 2099–2108.

[18] M. W. Fakhr, "Multiple encrypted random forests using compressed sensing for private classification," in *2018 International Conference on Innovation and Intelligence for Informatics, Computing, and Technologies (3ICT)*. IEEE, 2018, pp. 1–7.

[19] G. E. Hinton and R. R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," *Science*, vol. 313, no. 5786, pp. 504–507, 2006.

[20] E. D. Cubuk, B. Zoph, D. Mane, V. Vasudevan, and Q. V. Le, "Autoaugment: Learning augmentation strategies from data," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2019, pp. 113–123.

[21] A. Krizhevsky, G. Hinton *et al.*, "Learning multiple layers of features from tiny images," 2009.

[22] 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.

[23] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "Mobilenets: Efficient convolutional neural networks for mobile vision applications," *arXiv preprint arXiv:1704.04861*, 2017.

[24] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "Mobilenetv2: Inverted residuals and linear bottlenecks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 4510–4520.

[25] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 4700–4708.

[26] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. A. Alemi, "Inception-v4, inception-resnet and the impact of residual connections on learning," in *Thirty-first AAAI conference on artificial intelligence*, 2017.

[27] M. D. Zeiler, "Adadelta: an adaptive learning rate method," *arXiv preprint arXiv:1212.5701*, 2012.

IEEE Third International Conference on Data Stream Mining & Processing
August 21-25, 2020, Lviv, Ukraine

Forecasting Nonlinear Nonstationary Processes in Machine Learning Task

Peter Bidyuk
Department of Mathematical Methods of System Analysis,
National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute",
Kyiv, Ukraine
pbidyuke_00@ukr.net

Aleksandr Gozhyj
Department of Intelligent Information Systems,
Petro Mohyla Black Sea National University,
Nikolaev, Ukraine
alex.gozhyj@gmail.com

Irina Kalinina
Department of Intelligent Information Systems,
Petro Mohyla Black Sea National University,
Nikolaev, Ukraine
irina.kalinina1612@gmail.com

Victoria Vysotska
Department of Information Systems And Networks
National University "Lvivska politechnika", Lviv, Ukraine
Victoria.A.Vysotska@lpnu.ua

Mikhail Vasilev
Department of Intelligent Information Systems,
Petro Mohyla Black Sea National University,
Nikolaev, Ukraine
mishavas15@gmail.com

Romanna Malets
Programming Department,
Ivan Franko Lviv National University,
Lviv, Ukraine
romashkaandko@gmail.com

Abstract— **The article discusses the features of the solving the forecasting problems using machine learning techniques. The issues of accounting and correctly processing non-linear non-stationary processes in the problems of modeling and forecasting time series in various areas are considered. The analysis of stages and methods for solving machine learning problems. As an example, consider the problem of predicting currency pairs based on historical data. A comparative analysis of the normalization methods in data clustering is given. For the six currency pairs a short-term forecast is proposed.**

Keywords — nonlinear nonstationary processes, data preprocessing, model selection, short term forecasting, financial processes

I. INTRODUCTION

The main problem solved by many applied sciences, such as economics, finance, ecology, etc., is to obtain correct predictions touching upon the behavior of complex systems based on data on their past behavior. However, many problems arising in practical applications cannot be solved by known methods and respective algorithms. This happens due to the fact that the mechanisms for generating the initial data are not exactly known or there is no enough statistical data to build the predictive model. At the same time, one has to comprehensively study the initial sequence of the source data and try to make forecasts by constantly improving this scheme of processing the source data in the process of estimating the forecasts. The approach in which past data or examples are used to formulate and improve the forecasting scheme is called the Machine Learning method. Machine learning is an extremely broad and dynamically developing field of research that is using a huge number of theoretical and practical methods. Most models and forecasting methods in machine learning problems use the methods of probability and estimation theory as well as mathematical statistics [1-4]. As a part of this approach, we consider the problems of classification and regression analysis. The learning process consists of choosing the classification or regression function from a predetermined wide class of possible functions.

The article discusses a method for solving the problem of forecasting non-linear non-stationary data (processes) using machine learning methods. To construct the forecasts as non-linear and time-dependent data are taken historical data of exchange quotations.

II. PROBLEM STATEMENT

The purpose of this study is to develop techniques for constructing forecasts for nonlinear non-stationary financial processes based on machine learning techniques and technologies.

III. FEATURES OF NONLINEAR AND NONSTATIONARY DATA

In various fields, such as finance, ecology, social systems, management related to data processing, first of all, it is necessary to determine and classify what information and with what type of process we will have to work with and on this basis to determine data processing methods [5- 9, 11, 12]. However, usually there are known features of the process, such as linearity / nonlinearity and stationarity / nonstationarity, which can be quickly determined and classified in order to select the appropriate modeling and forecasting methods [5,10]. In Fig. 1 shows the classification of processes, from which it is possible to determine and classify the structure of mathematical models that are used to describe dynamic processes in solving the forecasting problem..

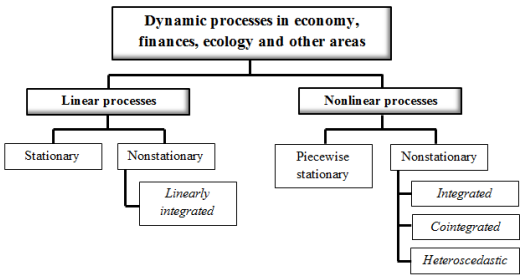


Fig. 1. Classification of the processes being studied

Linear processes can be stationary without a trend and unsteady when they include a component in the form of a first-order trend. If the variance (or covariance) of a stochastic linear process depends on time, then it is classified as heteroskedastic and requires the use of nonlinear models to describe the variance of the process and the process itself [11,13,15]. There is also a wide range of non-linear processes (types of non-linearity), although for the analysis of economic processes only some of them that are more common in economics and finance will be important. These processes can be classified as non-linear in parameters and non-linear in variables. The first type is more complex in terms of modeling and parameter estimation and usually requires more effort and time to develop their models. For example, logistic regression. Some non-linear processes may exhibit linear behavior in their stable or long-term mode of operation. This function allows you to linearly describe the process near the operating point. This type of nonlinearity is very common in economic problems and is used to solve them using machine learning methods. Typically, complex processes include integrated processes (IP) that contain a trend of a second or higher order, as well as co-integrated processes with trends of the same order and heteroskedastic processes with time-varying dispersion. A significant part of the financial processes related to the dynamics of prices for stock exchange instruments belong to this class [15, 16]. In engineering applications, such processes are studied and used in diagnostic systems, where an appropriate decision is made regarding the current state of the system. Since there are always elements of non-stationarity and non-linearity in financial processes, the subject of research will be the solution of the problem of forecasting exchange rates using machine learning methods.

IV. Features Forecasting in Problems of Machine Learning

The Machine Learning procedure for solving forecasting problems consists of the following steps: preliminary data processing stage, data normalization, classification (clustering), aggregation of data, forecasting.

Preliminary data processing. At the initial stage of data preparing, preliminary processing (preprocessing) and data filtering are necessary. Data pre-processing and filtering are tasks that must be completed before the data set can be used to train the model selected. Raw data is often distorted, there may be missing data in it. The use of such data in modeling and forecasting can lead to inaccurate and incorrect results. These tasks are part of the data processing and analysis process and are usually used in the initial review of the data set required during pre-processing. Data can be collected from various sources and processes. They may contain the following problems:

• incompleteness (data do not contain attributes, or knowledge is missing in them);
• noise (data contain erroneous records or outliers);
• independence (data conflict with each other).

High-quality of source data is necessary condition for creating high-quality forecasting models.

Data normalization. The purpose of normalization is to change the values of the numeric columns in the dataset to a common scale without distorting differences in the ranges of values. To perform machine learning, each data set needs normalization. Consider the two most popular: *min-max normalization* and *mean normalization*.

Minimax normalization, also known as minimum scaling or normalization of the minimum maximum, is the simplest method and is applied to scale the range of functions to the range of [0, 1] or [−1, 1]. The choice of target range depends on the nature of the data available. The general formula for min-max [0, 1] is given as:

$$x' = \frac{x - \min(x)}{\max(x) - \min(x)}$$

where x is original observation value, x' is the normalized value. To zoom between an arbitrary set of values [*a, b*], the formula becomes as follows:

$$x' = a + \frac{(x - \min(x))(b - a)}{\max(x) - \min(x)}$$

where *a, b* are the minimum and maximum values.

Mean based normalization:

$$x' = a + \frac{x - average(x)}{\max(x) - \min(x)}$$

where *x* is initial value, and x' is the normalized value.

Clustering. Clustering is the task of dividing a population or data points into a number of groups (clusters), so that data points in the same groups where more like other data points of the same group than other groups.

Clustering can be divided into two subgroups:

• Rigid clustering: in hard clustering, each data point either belongs to the cluster completely or not.
• Soft clustering: in soft clustering, instead of putting each data point in a separate cluster, the probability or probability that this data point will belong to these clusters is assigned.

Since clustering tasks are subjective, there are many tools that can be used to achieve this goal. Each methodology adheres to a different set of rules for determining similarity among data points. In fact, more than 100 clustering algorithms are known [14, 16-20].

Aggregation of data. Aggregation of time series can be considered as a "data reduction" process in the sense that it generalizes a set of time series. In practice, such a process is widely used in problems of clustering or classification.

In clustering procedures most algorithms repeat the 2 main steps: the assignment step and the centering or recalculation step. During the destination phase, the algorithm calculates

the distances between each observation and each centroid. Then the observations are distributed to its nearest clusters (minimum distance with the centroid). The choice of averaging methods is indeed central from the point of view of accuracy and mainly depends on the final goal (visualization, classification, clustering). Two types of averaging approaches are known: *DTW Barycenter Averaging* and *Soft-DTW*. The main advantage of these methods is that the series are averaged and therefore there is no effect on the order of using of the series.

Forecasting. Forecasting can be done on the basis of the variety of known methods including visual.

V. Example of Exchange Rate Forecasting

As an example of application of the machine learning methods consider forecasting exchange rates. As initial data, historical data of 6 currency pairs were used: AUDUSD, CHFJPY, EURUSD, GBPUSD, NZDUSD, USDJPY from 1990 to 2020 with an interval of 1 day. The data source for forecasting was the service Investing.com. Table I shows the values of descriptive statistics for the studied pairs.

TABLE I. Descriptive Statistics for the Studied Currency Pairs

	AUD USD	CHF JPY	EUR USD	GBP USD	NZD USD	USD JPY
Quantity	6890	6940	7709	6881	6693	7718
mean	0.761	91.8777	1.2065	1.5865	0.6594	110.5596
Std	0.1358	15.4668	0.1480	0.182	0.1123	15.3731
Min	0.4786	59.2300	0.8267	1.2023	0.3907	75.7800
25%	0.6778	81.7475	1.1155	1.4827	0.5896	102.5200
50%	0.7494	88.3300	1.2077	1.5790	0.6773	110.3500
75%	0.8264	104.2725	1.3144	1.6578	0.7295	119.8500
Max	1.1023	138.3200	1.5997	2.1065	0.8819	159.850

Normalization. At the data preprocessing stage, the normalization was applied. As a normalization method, the method based on the average value is used. Normalization results based on the average method are presented in Fig. 2

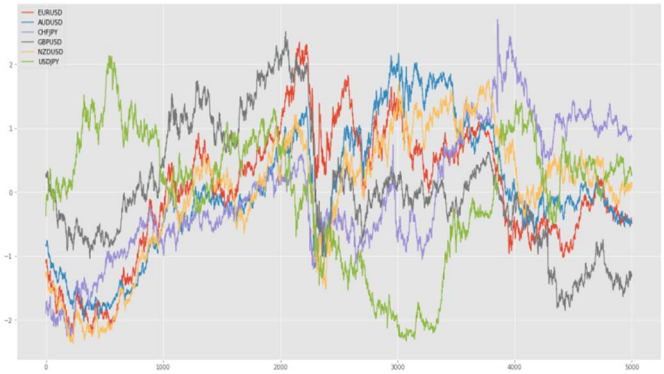


Fig. 2. Normalization results for various currency pairs

Clustering. In this example, the clustering procedure is used to perform search for price patterns and mark time series according to the similarity of patterns. For clustering the time series, the KNN method with the accuracy metric DTW is used. This approach allows one to accurately group time series into clusters by similarity even if there is a certain time shift. The training data was divided into 100 clusters due to the large amount of data and for greater accuracy in forecasting the price of currency pairs. In Fig. 3 examples of time series clustering are shown.

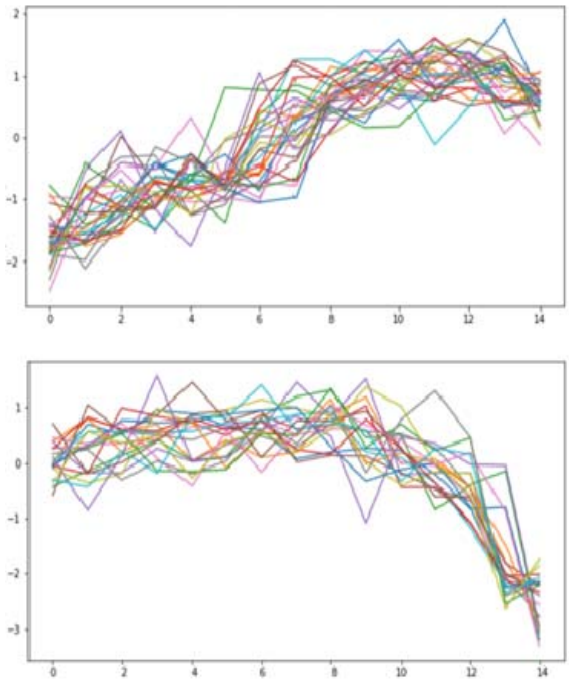


Fig. 3. Examples of clustering the time series selected for the cluster 0 and 2

Aggregation of the time series selected is performed by a search for a barycenter. Various methods were used to search for the barycenter, and the average DTW distance for the accuracy metric. The results of the study are presented in Table II, and their corresponding graphical comparison in Fig. 4.

To study the effectiveness of various machine learning methods in the task of classifying price patterns, we used the historical data of the closing prices of GBPUSD and AUDUSD currency pairs with an interval of 1 day from 1990 to 2020. The data were divided into training and test (validation) sets. The training set contains data for 7437 days, and the test set - 281. Further on, the data were combined into time series of 15 days long (3 weeks without days off) for marking DTW-KNN data and 10 days long for forecasting. All time series were normalized to mean. The result of data marking by the DTW-KNN method is shown in the form of distribution in Fig. 5.

TABLE II. Descriptive Statistics of the Studied Currency Pairs

Algorithm	Mean distance DTW
Euclidean barycenter	2.201
DTW	1.603
Soft DTW (γ = 1)	1.861
Soft DTW (γ = 0.1)	1.605
Soft DTW (γ = 0.01)	1.600

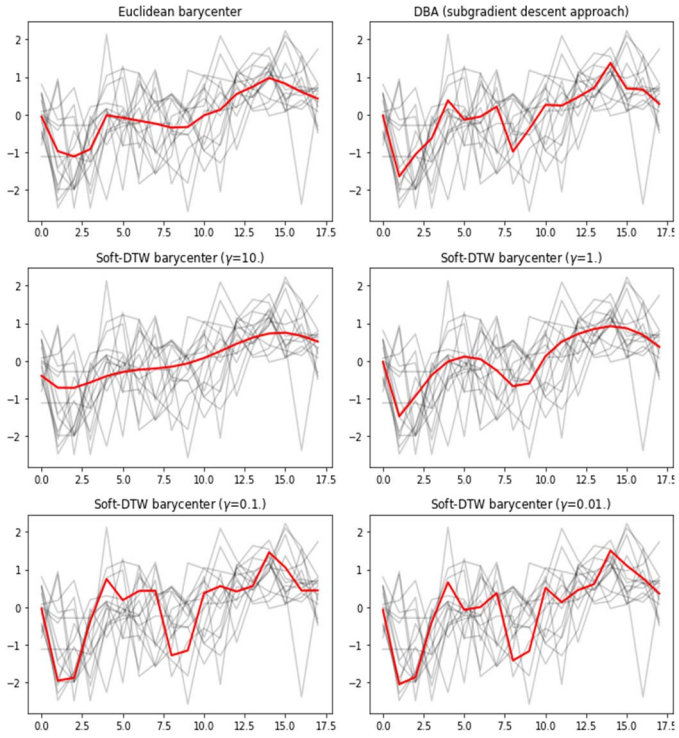


Fig. 4. Results of estimating the barycenter for different algorithms

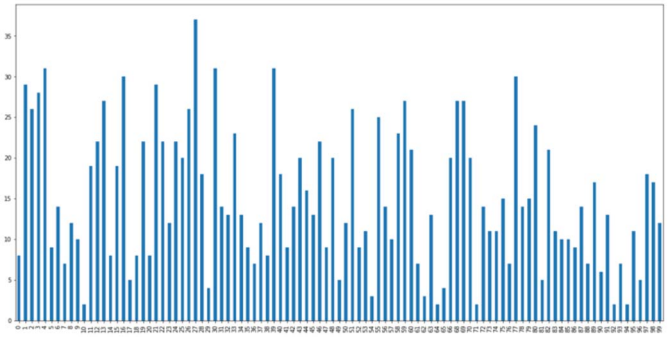


Fig. 5. Distribution of test case marks

The results of using the algorithms to solve the problem of classifying time series are shown in Table III.

TABLE III. THE RESULTS OF ALGORITHMS FOR CLASSIFICATION PROBLEMS

Algorithms	Template classification accuracy, %	Accuracy of classification price direction,%	Area ROCAUC
Decision tree	22.22	68.52	0.69
Logistic regression	25.93	74.07	0.75
Multilayer perceptron	35.19	72.2	0.73
Support vector machine (SVM)	38.89	75.93	0.76
Naive Bayesian classifier	37.03	70.37	0.71
Gradient Boosting	27.78	64.81	0.65

Forecasting. The forecasting model can be deployed into a container and launched as a micro service to create on-demand forecasts. Two types of models are used in this container: KNN models and SVM models. These models have the *sklearn API*. The models are loaded from the model database. The model database contains the configuration for KNN and SVM, including the basic settings of the models. For each currency pair a separate model was trained. The resulting information was saved in JSON format. To load KNN models, use the *sklearn API* as an extension of the from JSON method. The *joblib* Python library is used to load SVM models.

To obtain the forecast, SVM and price patterns (barycenters) KNN are used.

Forecasting Algorithm:

1. Submit an SVM time series containing information about the past 10 days.
2. Classify the time series and get the price template number
3. Export the price template with KNN as a time series of 15 days.
4. Use the last 5 days of the price template to classify price movements.

The *Jupyter Notebook* is used to visualize the results and user input. The output is presented in textual and graphical form in the form of a graph that contains the most similar price template with a real-time price range. Figure 6 and 7 shows the results of forecasting GBPUSD and AUDUSD pairs for 5 days.

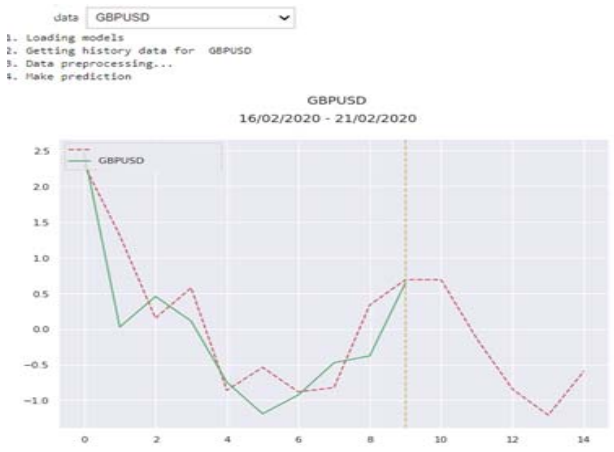


Fig. 6. GBPUSD 5 day forecasting results

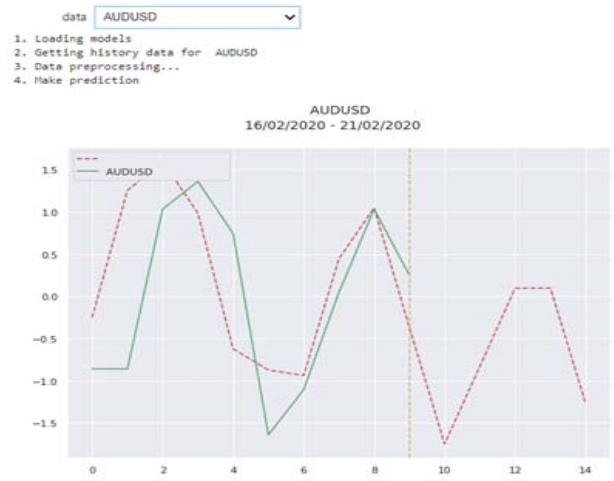


Fig. 7. AUDUSD 5 day forecasting results

As a result of the study the methods for classifying time series the SVM method was chosen because it showed better classification accuracy compared to other methods. To aggregate the KNN cluster, the Soft-DTW method was chosen. As a result of testing the system, these algorithms were based on the test historical data – the data that were not used for training models, the following forecasting results were obtained on currency pairs (Fig. 8).

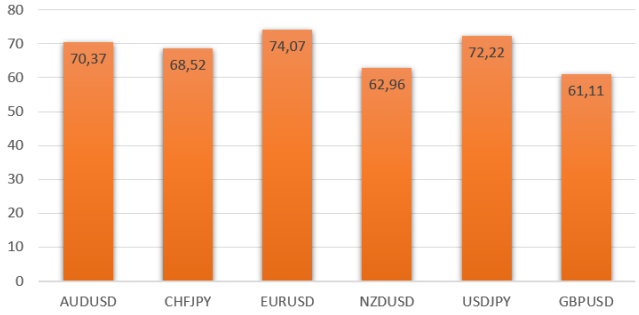


Fig. 8. Forecasting results of currency pairs based on historical data for 2019-2020.

CONCLUSION

Obtaining accurate forecasts in non-linear and non-stationary data is impossible without the use of various machine learning methods. High quality of the final forecasting result is achieved due to the proper control of the computational procedures used at all stages of data processing: data preprocessing, normalization, clustering, model structure and parameter estimation, calculation of short-term and medium-term forecasts. As an example of such data, exchange rates for 30 years were used. It is planned to further expand the developed methodology with new forecasting methods and machine learning methods.

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, T. Bollerslev, "Modelling the persistence of conditional variances", Econometric Reviews. Vol. 5, no. 1., pp. 1–50. 1986.

[3] A.C. Harvey, "Forecasting, structural time series models and the Kalman filter". Cambridge: The MIT Press, 1990. 554 p.

[4] C.E. Rasmussen, C.K.I. Williams "Gaussian processes for machine learning," Cambridge (Massachusetts), The MIT Press, 2006. 248 p.

[5] P.I. Bidyuk, V.D. Romanenko, O.L. Tymoshchuk Time series analysis. Kyiv: Polytechnika, NTUU «KPI», 2013. 600 p.

[6] G. Succarat, A. Escribano, "Automated model selection in finance: general-to-specific modeling of the mean, variance and density". Oxford Bulletin of Economics and Statistics, 2012, Vol. 74, Issue 5, pp. 716 – 735.

[7] F. Pretis, J.J. Reade, G.Succarat, "Automated general-to-specific (GETS) regression modeling and indicator saturation for outliers and structural breaks", Journal of Statistical Software, 2018, Vol. 86, Issue 3, pp. 1 – 44.

[8] P. Bidyuk, A. Gozhyj, I. Kalinina and et al: The Methods Bayesian Analysis of the Threshold Stochastic Volatility Model. Proceedings of the 2018 IEEE 2nd International Conference on Data Stream Mining and Processing, DSMP 2018, Lviv. pp. 70-75. (2018).

[9] P. Bidyuk, A. Gozhyj, I. Kalinina and V. Gozhyj, "Analysis of uncertainty types for model building and forecasting dynamic processes," Advances in Intelligent Systems and Computing II 689. Springer-Verlag. pp. 66-78, 2017.

[10] 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.

[11] T. Bollerslev, "Generalized autoregressive conditional heteroskedasticity". Journal of Econometrics. Vol. 31, no. 3. pp. 307–327. 1986.

[12] T. Bollerslev, "A conditionally heteroskedastic time series model forspeculative prices and rates of return ". The Review of Economics and Statistics.Vol. 69, no. 3. pp. 542–547. 1987.

[13] 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.

[14] 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.

[15] R.J. Hyndman, E. Wang and N. Laptev. "Large-scale unusual time series detection". In Proceedings 15th IEEE International Conference on Data Mining Workshop, ICDMW 2015, pages 1616–1619, 2016.

[16] B.D. Fulcher and N.S. Jones. "Highly comparative feature-based time-series classification". IEEE Transactions on Knowledge and Data Engineering, 26(12):3026– 3037, 2014.

[17] B.D. Fulcher, M.A. Little and N.S. Jones "Highly comparative time-series analysis: the empirical structure of time series and their methods". Journal of the Royal Society Interface, 10(83), 2013.

[18] M.H. Pesaran, A. Pick, M. Pranovich "Optimal forecasts in the presence of structural breaks", Journal of Econometrics, Vol. 177, Issue 2, 2013, pp. 134 – 152.

[19] T.J. Watsham, and K. Parramore "Quantitative Methods in Finance". London: International Thomson Business Press, 1997, 395 p.

[20] E. Xekalaki, S. Degiannakis, "ARCH Models for Financial Applications", New York: John Wiley & Sons, Ltd, Publication, 2010. – 535 p.

IEEE Third International Conference on Data Stream Mining & Processing
August 21-25, 2020, Lviv, Ukraine

Matrix Fuzzy-Probabilistic Neural Network in Image Recognition Task

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

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

Iryna Pliss
Control Systems Research Laboratory
Kharkiv National University of Radio Electronics
Kharkiv, Ukraine
iryna.pliss@nure.ua

Olha Chala
Artificial Intelligence Department
Kharkiv National University of Radio Electronics
Kharkiv, Ukraine
olha.chala@nure.ua

Anna Nortsova
Control Systems Research Laboratory
Kharkiv National University of Radio Electronics
Kharkiv, Ukraine
anna.nortsova@nure.ua

***Abstract* — The fuzzy matrix probabilistic neural network (FMPNN) designed to solving image classification tasks where images are represented in matrix form is proposed. This neuro-fuzzy system has four layers of data processing which are fed into the system in data stream form on the "sliding window". It is supposed that formed classes of data-images can be arbitrarily overlapped in feature space creating a situation of fuzziness. The tuning of FMPNN is the combination of learning procedures such as "Lazy learning", "Winner takes all", "Learning vector quantization", "Fuzzy C-means clustering". The feature of the proposed fuzzy classification - image recognition system is the high speed of data processing allowing to process data within the concept of Data Stream Mining.**

Keywords— probabilistic neural network, matrix images, fuzzy reasoning, self-organizing map, learning vector quantization.

I. INTRODUCTION

The problem of pattern recognition is quite common within the general problem of Data Mining, and there exist now a lot of approaches, methods, and algorithms for its solution. Among them, the most effective today are the ones based on computational intelligence approaches [1-3]. Currently, deep neural networks (DNNs) [4-6] are the most used to solve this problem, which provide the high quality of the received solution but are characterized by low learning speed, so their tuning can be implemented only in batch mode on a plurality of epochs. Thus, in situations where the size of the dataset which to be processed is unlimited, and the same data are fed to the recognition system in a sequential online mode, DNNs are ineffective. In this case, when the speed of the recognition system comes to the fore, DNNs (like the traditional SNNs of the three-layer perceptron type) can hardly be used.

One of the most rapid and efficient neural network that solve the problem of pattern recognition is the so-called probabilistic neural network (PNN) introduced by D. Specht [7-8] that is trained using the principle of "lazy" learning: "Neurons at data points "[9]. At the heart of these neural networks are the kernel methods [10] and, above all, the reasoning according to Bayes, the Parzen windows, and Nadaraya-Watson's estimates. As activating functions, these networks use bell-shaped ones (Gaussian, Cauchian, etc.), and learning is done by establishing centers of functions in images that arrive at the receptive (zero) layer. Due to this, it provides super-fast settings such as "Just in time models" [11].

It should be noted that although the accuracy of classification PNN is yielded to the modern DNN, their speed in some cases gives them a significant advantage.

Generally, the classical PNN solves the problems of crisp classification, that is, a priori it is assumed that the restored classes do not overlap in one another. If it is not so, instead of crisp neural networks, neuro-fuzzy systems should be used, which, in addition to forming classes itself, can also estimate the level of membership of each particular observation to each possible class. In this regard, in [12-16], fuzzy modifications of crisp PNNs were introduced that process information under various assumptions about the properties of data and classes.

It should also be noted that the input information of most known neural networks, including the PNN, comes in the form of vectors, which means that image recognition tasks must first be vectorized, which means, matrix signals must be converted into vectors. In principle, this can be accomplished by using operations such as convolution and pooling as it is done in a convolutional DNN. However, the use of these operations significantly reduces the speed of the recognition system in general. Therefore, it is expedient not to vectorize the input images, but to feed them to the receptive layer in a matrix form.

That is why we propose to introduce into consideration fuzzy probabilistic neural network for image recognition, which provides a high speed of information processing, which comes in the matrix form and can restore the observation classes arbitrarily intersected in the space of features.

II. The Architecture of Matrix Fuzzy-Probabilistic Neural Network

The proposed matrix fuzzy-probabilistic neural network (MFPNN) contains four layers of information processing: the first hidden layer of patterns, the second hidden layer, formed by elements of summation, the third hidden layer of probability distributions correction, and, finally, the fourth output layer, which determines the levels of membership of the classified images to specific classes.

Information comes to the zero (receptive) layer of the system is given by a sample of images in the form ($n_1 \times n_2$) – matrices $X = \{x(1), x(2), \ldots, x(k), \ldots, x(N)\}$, $x(k) = \{x_{i_1 i_2}(k)\}$ $\in R^{n_1 \times n_2}$, where N_1 observations belong to the first class Cl_1, N_2 - to the second class $Cl_2, ..., N_m$ to Cl_m, where $j = 1, 2, ..., m$, $k = \underbrace{1, 2, \ldots, N_1}_{Cl_1 \text{ contains } N_1 \text{ observations}}, \underbrace{N_1 + 1, \ldots, N_1 + N_2}_{Cl_2 \text{ contains } N_2 \text{ observations}}$, $..., N_1 + N_2 + N_3, \underbrace{\ldots, N}_{Cl_m \text{ contains } N_m \text{ observations}}$.

The number of radial-basis neurons [10] is defined by the amount of observations *N*, with the activation functions of this layer in the future we will denote as $\varphi(x, c_{\tau_j}, \sigma^2_{\tau_j})$ where τ_j varies in the $N_1 + N_2 + \ldots + N_j + 1, N_1 + N_2 + + \ldots N_j + 2, N_1 + N_2 + \ldots N_j, c_{\tau_j} \in R^{n_1 \times n_2}$ – matrix center of activation function, that is $\{c^j_{\tau_j i_1 i_2}\}$ – $n_1 \times n_2$ matrix, $\sigma^2_{\tau_j}$ – a parameter that specifies the "width" of the corresponding bell-shaped activation function. As an activation function can be used either matrix Gaussian

$$\varphi(x, c_{\tau_j}, \sigma^2_{\tau_j}) = \exp\left(-\frac{Tr(x - c_{\tau_j})(x - c_{\tau_j})^T}{2\sigma^2_{\tau_j}} \right) \quad (1)$$

or any other bell-shaped function, such as Cauchian

$$\varphi(x, c_{\tau_j}, \sigma^2_{\tau_j}) = \frac{1}{1 + \sigma^{-2}_{\tau_j} Tr(x - c_{\tau_j})(x - c_{\tau_j})^T} \quad (2)$$

where

$$Tr(x - c_{\tau_j})(x - c_{\tau_j})^T = d^2_F(x, c_{\tau_j}) \quad (3)$$

– matrix Frobenius metrics, which specifies the distance between the matrix centers c_{τ_j} and the input image *x*.

Network tuning usually occurs only at the level of pattern layer based on the concept of "Neurons at data points" [9], that is, according to the rule

$$c_{\tau_j} = x(k), \text{ if } x(k) \in Cl_j, \tau_j = N_1 + N_2 + \ldots + N_{j-1} + 1, \ldots, N_1 + N_2 + \ldots N_j. \quad (4)$$

When all of the radial-basis neurons of the first hidden layer of an image *x* with an unknown classification on the outputs of current layer appear *N* signals $o^{[1]}_k = \varphi(x, c_{\tau_j}, \sigma^2_{\tau_j})$ $\forall$ $j = 1, 2, \ldots, m$; $k = 1, 2, \ldots, N$. The second hidden layer consists of m elementary adders (one for each class), and on its outputs signals are formed

$$o^{[2]}_j(x) = \sum_{\tau_j = N_1 + N_2 + \ldots + N_{j-1} + 1}^{N_1 + N_2 + \ldots + N_j} o^{[1]}_{\tau_j}(x) \quad (5)$$

which are the Parzen estimates of the probability distribution density, i.e.

$$p_j(x) = \sum_{\tau_j = N_1 + N_2 + \ldots + N_{j-1} + 1}^{N_1 + N_2 + \ldots + N_j} o^{[1]}_{\tau_j}(x). \quad (6)$$

In the third hidden layer, these estimates are corrected to take into account empirical a priori probabilities

$$P_j(x) = \frac{N_j}{N} \quad (7)$$

and cost of classification's errors S_j so that

$$o^{[3]}_j(x) = o^{[2]}_j(x) N_j N^{-1} S_j \quad (8)$$

and, finally, in the output layer, the class *m* to which the image *x* belongs is determined, that is the "class-winner":

$$j^* = \arg\max_j o^{[3]}_j(x). \quad (9)$$

In principle, the described procedure does not differ from the work of the standard PNN in addition to the fact that the input of the system obtains not traditional vectors-images but matrices-images.

In real-world problems, the situation is often complicated by the fact that an image – subject to classification can simultaneously belong to several classes at the same time with different levels of membership. This situation occurs when the classes formed by the training dataset *X* are mutually overlapping, in other words, a fuzzy situation arises.

The centroids for each class are taken into consideration in the form

$$c_j = \frac{1}{N_j} \sum_{\tau_j = N_1 + N_2 + \ldots + N_{j-1} + 1}^{N_1 + N_2 + \ldots + N_j} c_{\tau_j} \quad (10)$$

and the radii of these classes –

$$r_j = \max d_F(x(\tau_j), c_j). \quad (11)$$

Then, if there is a situation for two arbitrary classes Cl_j and Cl_l

$$d_F(c_j, c_l) < r_j + r_l, \quad (12)$$

then these classes overlap.

To estimate the level of the membership of the observation x to class Cl_j can be used as an estimate, that arises in the popular method of fuzzy C-means clustering (FCM) in the form [2, 3]:

$$\mu_j(x) = \frac{d_F^{-2}(\mathrm{x},\mathrm{c}_j)}{\sum_{l=1}^{m} d_F^{-2}(\mathrm{x},\mathrm{c}_l)} = \frac{\left(Tr(\mathrm{x}-c_{\tau_j})(\mathrm{x}-c_{\tau_j})^T\right)^{-1}}{\sum_{l=1}^{m}\left(Tr(\mathrm{x}-c_{\tau_j})(\mathrm{x}-c_{\tau_j})^T\right)^{-1}} =$$
$$= \frac{1}{1+d_F^2(\mathrm{x},\mathrm{c}_j)\sum\limits_{\substack{l=1\\ l\neq j}}^{m} d_F^{-2}(\mathrm{x},\mathrm{c}_l)} = \frac{1}{1+\sigma_j^{-2}d_F^2(\mathrm{x},\mathrm{c}_j)} =$$
$$= \frac{1}{1+\sigma_j^{-2}Tr(\mathrm{x}-c_j)(\mathrm{x}-c_j)^T} \qquad (13)$$

where

$$\sigma_j^2 = \left(\sum_{\substack{l=1\\ l\neq j}}^{m} d_F^{-2}(x,c_l)\right)^{-1}. \qquad (14)$$

It's easy to see those fuzzy memberships are determined with Cauchian (2), which can also be used as activation functions of the second hidden layer (13). It should be noted that the width parameters σ_j^2 during evaluating fuzzy memberships are determined automatically using the relation (14).

Thus, in a fuzzy case in the output layer of the network is determined both the winner class and the membership level of each image x to each class Cl_j.

III. Online Matrix Fuzzy-Probabilistic Neural Network Training

In Data Stream Mining tasks, the situation when the training dataset is given not in the form of a fixed data batch, but as the sequence $x(1), \ldots, x(k), \ldots, x(N), x(N+1), \ldots$ that increases in size over time. This automatically increases the number of neurons in the pattern layer, which means the network in the training process "builds up" its architecture. Therefore, the configuration of such a network should occur in a sequential mode with the help of those or other recurrent learning algorithms. Assume that the training dataset received an observation $x(N+1)$ that belongs to the class Cl_j. At the same time in the pattern layer, one extra neuron is added to the group that "corresponds" exactly for this class. The procedure for clarifying the centroid of this class can be represented as

$$c_j(N_j+1) = \frac{1}{N_j+1}\sum_{\tau_j} c_{\tau_j} \qquad (15)$$

or in a recurrent form

$$c_j(N_j+1) = \begin{cases} c_j(N_j) + (N_j+1)^{-1}(x(N+1) - c_j(N_j)), \\ if\ x(N+1) \in Cl_j, \\ c_j(N_j), if\ x(N+1) \notin Cl_j. \end{cases} \qquad (16)$$

It's easy to see that (16) is completely coincides with the T. Kohonen's self-organizing maps (SOM) training rule [17], "The winner takes all" (WTA).

The increasing number of neurons in the pattern layer may be unwanted because the network may become too cumbersome, especially if the information comes in the form of a stream of images. In this case, it is advisable to fix the number of neurons for each class by some number N_j (basically, this number may be the same for all classes), and the training procedure can be implemented on a "sliding window" in the form

$$c_j(N_j+1) = \begin{cases} c_j(N_j) + \dfrac{1}{\overline{N_j}}(x(N+1) - x(N-\overline{N_j}+1)), \\ if\ x(N_j+1) \in Cl_j, \\ c_j(N_j), if\ x(N+1) \notin Cl_j. \end{cases} \qquad (17)$$

In fuzzy cases, when classes can overlap, a situation when the observation $x(N+1)$ belongs to the Cl_l is placed closer to the centroid $Cl_j(N)$ than to the "native" $Cl_l(N)$, that is $d_F(x(N+1), c_j(N)) < d_F(x(N+1), c_l(N))$ can arise. However if $x(N+1) \in Cl_l$ and if in the procedure (16), the specified centroid $Cl_l(N+1)$ is "tightened" to $x(N+1)$, in this case, $c_j(N+1)$ should "push" from $x(N+1)$.

In this situation, the tuning procedure for centroid c_j can be written in the form

$$c(N_j+1) = \begin{cases} c_j(N) + \eta(N+1)(x(N+1) - c_j(N_j)), \\ if\ x(N+1) \in Cl_j,\ d_F(x(N+1), c_j(N)) < \\ < d_F(x(N+1), c_l(N)), \\ c_j(N) - \eta(N+1)(x(N+1) - c_j(N_j)), \\ if\ x(N+1) \in Cl_l,\ d_F(x(N+1), c_j(N)) < \\ < d_F(x(N+1), c_l(N)), \\ c_j(N), if\ x(N+1) \notin Cl_j, Cl_l \end{cases} \qquad (18)$$

where $0 < \eta(k+1) < 1$ - the parameter of the learning rate, which is selected according to A. Dvoretzky's conditions [18].

It's easy to see that procedure (18) is an algorithm for learning vector quantization (LVQ) [19, 20], which is widely used in problems of image classification-recognition.

Thus, the introduced matrix probabilistic neural network is a hybrid of the standard PNN, the Kohonen's SOM, and LVQ, as well as the J. Bezdek's FCM, which is meant to solve the problems of images classification under conditions of classes that arbitrarily overlap in the space of features.

IV. RESULTS OF THE EXPERIMENT

For the implementation of the proposed MFPNN the two well-known datasets such as UCI dataset "ML hand-written digits datasets" and "Fashion MNIST" were used. The first one can be perfectly used for the fuzzy case, by the reason of low resolution of original images-matrices every of which is represented by a matrix of 8x8 (Fig. 1).

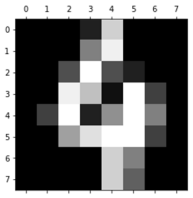


Fig. 1. The sample of the "ML hand-written digits datasets" dataset.

The second dataset has 60 000 observations in the training dataset and 6000 in the test dataset which allows evaluating the effectiveness in online mode.

Both of them were processed using the K-Nearest Neighbours (KNN), Convolutional Neural Network (CNN) [21] and MFPNN. The results of the experient are represented in the Table 1 and Table 2. The first table contains the results of the accuracy and time consumption evaluation of all mentioned approaches with ML hand-written digits dataset.

TABLE I. THE ACCURACY AND TIME CONSUMPTION OF ALGORITHMS WITH ML HAND-WRITTEN DIGITS DATASET

Algorithms for comparison	**Classification accuracies**	**Time, sec**
KNN	81,98	0.18
CNN	87,63	6.42
MFPNN	83,33	5.95

The accuracy of the matrix fuzzy-probabilistic neural network is higher compared to the KNN but yield to CNN. KNN and CNN algorithm were performed on the GPU and the proposed approach on the CPU. Due to hardware support, the speed in the first and second cases increases by up to 10 times [21]. Thus, the classification time in accordance with the proposed approach with the same hardware implementation would be comparable with the classification using the KNN and several times shorter than the classification time using the CNN. The Table 2 describes the accuracy of algorithms for a different number of elements of the Fashion MNIST dataset.

TABLE II. THE ACCURACY OF ALGORITHMS WITH FASHION MNIST DATASET

Algorithms for comparison	**Classification accuracies**					
	500	***1000***	***2000***	***6000***	***12000***	***60000***
KNN	50,98	71,3	76,8	80,96	82,72	83,33
CNN	62,8	73,11	77,01	83,68	86,01	90,99
MFPNN	67,33	75,86	80,13	81,5	82,33	84,74

From the presented data we can conclude that the accuracy of the MFPNN is growing faster comparing to the others mentioned approaches as the number of elements is increased. Therefore the proposed algorithm allows achieving higher accuracy when training on a smaller number of elements, which reduces the training time in online mode.

V. CONCLUSION

In the article, the fuzzy matrix modification of the probabilistic neural network is proposed, which is intended for solving the problems of classification-recognition of images in a stream of images, which sequentially come for processing. The introduced system and its learning algorithm combine the advantages of traditional probabilistic neural networks, self-organizing maps, learning vector quantization, and fuzzy clustering. Since the introduced system does not vectorize the images, but leaves them in a matrix form and is capable of operating under conditions of arbitrary shapes that overlap in the space of features, it has extended properties and increased performance compared to known systems for image recognition.

REFERENCES

[1] C.L. Mumford, Computational Intelligence. Berlin: Springer-Verlag, 2009.

[2] R. Kruse, C. Borgelt, F. Klawonn, C. Moewes, M. Steinbrecher, P. Held, Computational Intelligence. A Methodological Introduction. Berlin: Springer-Verlag, 2013.

[3] J. Kacprzyk, W. Pedrycz, Springer Handbook of Computational Intelligence, Berlin Heidelberg: Springer Verlag, 2015.

[4] Y. Bengio, Y. Le Cun, G. Hinton, "Deep learning," in Nature, vol. 521, iss. 7553, 2015, pp.436-444.

[5] J. Schmidhuber, "Deep learning" in Neural networks: An overview. Neural Networks, vol. 61 2015, pp.85-117.

[6] I. Goodfellow, Y. Bengio, A. Courville, Deep Learning. MIT Press, 2016.

[7] D. F. Specht "Probabilistic neural networks" in Neural Networks, vol. 3, 1990, pp. 109-118.

[8] D. F. Specht, "Probabilistic neural networks and polynomial ADALINE as complementary techniques to classification", IEEE Trans. on Neural Networks, vol. 1, pp. 111-121, March 1990.

[9] D.R. Zahirniak, R. Chapman, S.K. Rogers, B.W. Suter, M. Kabriski, V. Pyatti, "Pattern recognition using radial basis function network," in Aerospace Application of Artificial Intelligence, Proceedings, Dayton, Ohio, 1990, pp. 249–260.

[10] S.Y. Kung, Kernel Methods and Machine Learning. Cambridge: University, 2014.

[11] O. Nelles, Nonlinear Systems Identification. Berlin: Springer, 2001.

[12] Ye. Bodyanskiy, Ye. Gorshkov, V. Kolodyazhniy, J. Wernstedt, "Probabilistic neuro-fuzzy network with non-conventional activation functions," Lecture Notes in Artificial Intelligence, vol. 2774. Berlin Heidelberg New York: Springer, 2003, pp. 973-979.

[13] Ye. Bodyanskiy, Ye. Gorshkov, V. Kolodyazhniy, "Resource allocating neuro-fuzzy network," Proc. 3-rd. Int. Conf. of the European Society for Fuzzy Logic and Technology. Zittau/Görlitz: University of Applied Sciences (FH), 2003, pp.392-395.

[14] Ye. Bodyanskiy and O. Shubkina, "Semantic annotation of text documents using modified probabilistic neural network," 6th IEEE Int. Conf. on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications, Prague, Czech Republic, 2011, pp. 328-331.

[15] Ye. Bodyanskiy, O. Shubkina, "Semantic annotation of text documents using evolving neural network based on principle "Neurons at Data Points" ," 4th Int. Workshop on Inductive Modeling "IWIM 2011", Kyiv, Ukraine, 2011, pp. 31-37.

[16] Ye. Bodyanskiy, I. Pliss and V. Volkova, "Modified probabilistic neuro-fuzzy network for text document processing," in Int. J.Computing, Vol. 11, Iss. 4, 2012, pp. 391-396.

[17] T. Kohonen, Self-Organizing Maps. Berlin: Springer-Verlag, 1995.

[18] A. Dvoretzky, "On stochastic approximation," Proc. 3-rd Berkley Symp. Math, Statistics and Probability-Berkley, vol. 1. CA: University of California Press, 1956, pp. 39-55.

[19] T. Kohonen, "Improved version of learning vector quantization," Proc. Int. Joint Conf. on Neural Networks, vol. 1. San Diego, CA, 1990, pp. 545 - 550.

[20] J. S. Baras, A. La Vigna, "Convergence of Kohonens learning vector quantization". San Diego, CA, vol.3. 1990, pp. 17-20.

[21] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards realtimeobject detection with region proposal networks," in NIPS, 2015, pp. 91–99.

IEEE Third International Conference on Data Stream Mining & Processing
August 21-25, 2020, Lviv, Ukraine

The Algorithmic Classification Trees

Igor Povkhan
The Faculty of Information Technology
Uzhhorod National University
Uzhhorod, Ukraine
igor.povkhan@uzhnu.edu.ua

Maksym Lupei
The Faculty of Information Technology
Uzhhorod National University
Uzhhorod, Ukraine
maxim.lupey@gmail.com

***Abstract* — Information technology based on mathematical models of recognizing images in the form of classification trees is widely used in socio-economic, environmental and other information processing systems. This is explained by the fact that this approach eliminates a set of shortcomings of classical methods and achieves a fundamentally new result, efficiently and rationally using the power of computer systems. Each of the known algorithms and classification methods is restricted to the specificity of the application tasks, and this is by far the weakest point of not only these algorithms, but also recognition systems which are based on the respective concepts. The methods of building recognition systems, based on methods of logical (algorithmic) classification trees (decision trees), do not have this shortcoming. The peculiarity of the logical tree method (the algorithmic classification tree method) is the possibility of complex use for solving each specific task of constructing the recognition scheme of many known algorithms (methods) of recognition. The research is based on the single methodology – the optimal approximation of the training dataset with the help of a set of generalized features (autonomous classification algorithms) that are part of a scheme (an operator) built in the training process.**

Keywords — Classification system, discrete object, feature set, algorithmic classification tree, branching criterion.

I. INTRODUCTION

Today, there are more than 3600 recognition algorithms (based on different approaches and concepts) that have some limitations in their use (accuracy, speed, memory, unversality, reliability, etc.). It is known that the presentation of large-scale training samples (discrete information) in the form of logical tree structures (graph schemes) has significant advantages in terms of simple economic description of data and effective mechanisms (algorithms and methods) of working with them [1,2].

Therefore, the coverage of the initial training set (TS) with a set of ranked elementary features in the case of logical classification trees (LCTs), or the coverage of the training dataset with a fixed set of autonomous recognition and classification algorithms in the case of algorithmic classification trees (ACTs), generates a fixed tree structure of data (a data scheme); it, to some extent, enables even the compression and conversion of the initial dataset – and thus allows a significant optimization and savings in hardware resources of the information system (a classification system) which is under construction [3].

Therefore, the possibility of representing the recognition function (classification rules) in the form of a model of a certain classification tree has great advantages compared to other representations of classification schemes; and the method for generating algorithmic classification trees, proposed later in this study, according to the TS, complements the methodology of the decision tree approach (branched feature selection methods) and allows building simple and effective classification systems for discrete objects [4-8]. This paper focuses precisely on the description and peculiarities of the method for constructing an algorithmic classification tree model for the initial training set.

II. PROBLEM STATEMENT

Suppose that initially there is some training dataset in the standard form of a sequence of training pairs:

$$\left(x_1, f_R(x_1)\right), \dots, \left(x_M, f_R(x_M)\right). \quad (1)$$

It should be noted that here $x_i \in G$ ($G\ is$ some set of initial signals) and, respectively, $f_R(x_i) \in \{0,1,2, \dots, k-1\}, (i = 1,2, \dots, M)$, and $M\ is$ the total number of training pairs (objects of the known classification) of the initial training dataset in the problem.

Accordingly, $f_R(x_i) = l, (0 \le l \le k-1)$ means that the object $x_i \in H_l, H_l \subset G$. And here $f_R\ is$ some finitely significant function (a recognition function – RF) defining the initial R partitioning of $G\ set$ which consists of subsets (images, classes) $H_0, H_1, H_2, \dots H_{k-1}$ (of the specified TS).

Therefore, it can be stated that the initial training dataset is a collection (more precisely a sequence) of some sets (objects of the known classification), each set being a collection of values of some features and the value of some function (RF) that this set possesses. In other words, it can be pointed out that a collection of the values of features is a certain image (a discrete object), and the function value refers this image to the respective pattern [9,10].

Therefore, in view of the mentioned above, at this stage of the study the task is to construct the L structure of some classification tree (LCT / ACT) whose structural parameters p would be optimal $F(L(p, x_i), f_R(x_i)) \to opt$ in relation to the initial training dataset.

It is admitted that the main idea of the method of stepwise selection of elementary features (or a set of algorithms) is to maximize the value quality of the feature (algorithm) $W_M(f)$ [11]. The latter means that in the logical tree algorithms for the training set of type (1) there must be found such a generalized feature *f*, for which $W_M(f)$ value is as large as possible [12].

III. ALGORITHMIC CLASSIFICATION TREE

Suppose, in the beginning there is a training dataset of general type (1) – in the form of a sequence of training pairs $(x_i, f_R(x_i))$, with the power M, n as the dimensional feature space and a fixed set of classification algorithms of different types $(\alpha_1, \alpha_2, \dots, \alpha_m)$. It should be stressed that the operating of the constructed classification models is checked by means of the training dataset with the cardinality T (whose class is also known).

It is noted that here the initial training dataset determine some R partitioning into classes $(H_0, H_2, \dots, H_{k-1})$, and the corresponding algorithms α_i may be unrelated in terms of a single recognition concept, but implement different classification methods and approaches (for example, these can be ordinary geometric algorithms whose operating principle is to approximate the training set with the help of the corresponding geometric objects, algorithms for calculating estimates, potential functions, etc.) [13-15].

Admittedly, the result of the operating of each of the fixed (selected from the library of algorithms of some information system) autonomous classification and recognition algorithms a_i, within the corresponding step of generating an algorithmic classification tree, is one or more generalized features f_j (certain classification rules) which approximate the determined part of the initial training samples [2,4,8]. Thus, for the case of known geometric recognition algorithms geometric objects covering the TS in feature space of the n dimensional problem are the respective subsequent generalized features (for short, GFs).

It is clear that in real examples, there may be cases when the corresponding classification algorithm a_i cannot construct the generalized feature f_j due to the complex arrangement of the classes H_l in the feature space of the problem, or certain conceptual and implementation restrictions of the classification algorithm itself. Indeed, by analogy with a logical classification tree, there may be a case when the generalized features f_j constructed by the classification algorithm a_i do not fully approximate the initial training set, or such a situation is provided by the scheme of the algorithm of generating an algorithmic classification tree (for example, the presence of an initial restriction in the tree algorithm scheme – that is about generating no more than one GF f_j at each stage of constructing an algorithmic classification tree model).

It should be admitted that the objects of the initial TS which do not fall under the constructed scheme of sampling approximation by the sequence of the generalized feature$f_j$refer to failures (errors) of classification of the first type – En_{tr} and similarly for the TS incorrectly classified discrete objects are also referred to errors of the first type – Et_{tr}.

Therefore, in view of all the above, it can be assumed that the structure of an algorithmic classification tree (type I) will have a general construction of the form – (Fig. 1), where each tier of such a logical tree determines the stage of constructing an algorithmic classification tree in the form of approximation with the help of the current classification algorithm a_i of a certain part of the training set and owing to this approach it allows adjusting the final complexity (accuracy) of the obtained tree classification model.

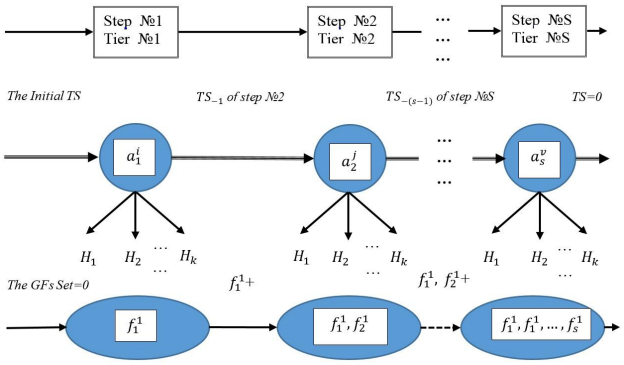


Fig. 1. The general scheme of the algorithmic classification tree structure.

It should be noted that within each step of generating the algorithmic classification tree model – (Fig. 1) its classification algorithm α_i as well as its respective TS (or the subset of the initial TS) is presented with the initial TS in its entirety being provided only in the first step; further with the subsequent stages of constructing the classification tree the power of the TS data will fall due to the set of the constructed GFs f_j that will cut (cover) some part of the initial TS. It is also important to note that, depending on the structure of the constructing scheme of the algorithmic classification tree and the specificities of the current algorithm α_i, it is possible to generate more than one GF f_j within each step.

At the next stage, for the method of algorithmic classification trees, there are introduced two basic criteria for constructing a classification tree model – the criterion for stopping the branching procedure K_{stop} (it actually regulates the complexity and accuracy of the obtained ACL model) and the selection criteria for branching $W(a)$ (choosing a classification algorithm within the current step) for the classification tree which is being constructed.

Thuswise, based on the above, it is feasible to introduce the stopping criterion K_{stop} of the branching process of the type (boolean) of the construction procedure of the algorithmic classification tree, which consists in checking the power $P_{pt}(TS)$ of the training set as follows:

$$K_{stop} = \{0, if\ P_{pt}(HB) = 0\ 1, if\ P_{pt}(HB) > 0 \quad . \quad (2)$$

It is worth noting that the procedure for constructing a classification tree continues until $K_{stop} = 1$, and the opposite situation – when $K_{stop} = 0$ signals the completion of the step of synthesyzing the algorithmic classification tree model and the need to go to the stage of test verification of the constructed classification rule using the TS and evaluation of the quality of the obtained classification tree model.

It should be emphasized that in the algorithmic classification tree method, the fundamental question is related to choosing a branching criterion (choosing the current classification algorithm a_i) in the structure of the classification tree model under construction.

It is clear that by analogy with the method of approximating the TS by means of a set of ranked elementary features as a branching criterion there can be suggested the initial estimate of the efficiency of the set of algorithms $(\alpha_1, \alpha_2, \dots, \alpha_m)$ which is as follows:

$$W(a_i) = \frac{\frac{1}{P_{pt}(HB)} * \sum_{j=1}^{k} (T_{y3} + S_{y3} + \frac{E_{y3}}{S_{y3}})}{k}. \quad (3)$$

Admittedly, in the proposed functionality (3) the entered values are interpreted in the following way:

1) k is the total number of classes of the current problem on the basis of partitioning R of the initial TS.

2) T_T characterizes the total time (hardware time) spent to build the current GF f_j;

3) E_T is the information capacity (structural complexity) of the built GF f_j within the current step of generating the algorithmic classification tree model;

4) S_T represents the total number of discrete objects x_i of the TS that are generalized (described) by the indicated GF f_j.

5) $P_{pt}(TS)$ is the power (volume) of the initial TS (or its fixed part) within the current step of the algorithm for constructing an ACT).

It should be underscored that in formula (3) summation occurs in view of all classes that are specified by the data of the initial TS (although there may be summation restrictions which are caused by the structure (parameters) of the algorithm of constructing a classification tree).

Taking into account the above mentioned, one of the possible algorithmic schemes of constructing the algorithmic classification tree (an ACT of type I) can be suggested.

The scheme of constructing an algorithmic classification tree (type I).

Stage 1. At the first stage of constructing the algorithmic classification tree a set of autonomous classification and recognition algorithms $(\alpha_1, \alpha_2, \dots, \alpha_m)$ is recorded in the library of algorithms of the information system (selected interactively, randomly or after the procedure of appropriate evaluation of quality (efficiency) according to the initial TS). It must be noted that both the classification algorithms themselves and their number (the value m) are selected depending on the conditions and aspects of the application task.

Stage 2. At the second stage of constructing the algorithmic classification tree, a set of classification algorithms $(\alpha_1, \alpha_2, \dots, \alpha_m)$ is estimated and ranked on the basis of the functional (3) in view of the TS data in the set according to their efficiency. It should be pointed out that here by analogy with the logical classification tree there are two options available – depending on the algorithmic scheme of constructing a classification tree:

One option is when the evaluating of the efficiency and ranking of the set of classification algorithms $(\alpha_1, \alpha_2, \dots, \alpha_m)$ are carried only once at this stage, and then within each step of constructing an algorithmic classification tree, the following algorithm a_i of the initial sequence is fixed to approximate the data. This approach significantly saves the hardware resources of the information system, but adversely affects the complexity of the resulting classification tree model.

Another option is when the evaluating of the efficiency and ranking of a set of classification algorithms $(\alpha_1, \alpha_2, \dots, \alpha_m)$ is carried out within each step of constructing the algorithmic classification tree in accordance with the respective data of the subsets (parts) of the initial TS in order to evaluate and identify the best quality (most efficient) classification algorithm for that part of the TS (the step of generating an ACT). This approach allows, with fewer steps, the completion of the TS approximation and obtaining a more cost-effective construction of the algorithmic classification tree in comparison with option (a), however, requires significantly more hardware resources of the information system for the second stage of the scheme of constructing the algorithmic classification tree and requires considerable attention and introducing a set of constraints related to the initial selecting of the set of classification algorithms $(\alpha_1, \alpha_2, \dots, \alpha_m)$.

Stage 3. At the third stage of the construction scheme of the algorithmic classification tree, there is recorded the initial top of the algorithmic classification tree, that is the classification algorithm α_i of the highest efficiency among the selected set $(\alpha_1, \alpha_2, \dots, \alpha_m)$, and the initial TS is presented in the form of a sequence of training pairs on the input. The α_i algorithm provides the generation of one or more GFs f_i of the first tier (the number of GFs generated within each step are determined by the parameters of the very algorithm of constructing an ACT), which approximate (classify) a certain part of the TS.

Stage 4. At the fourth stage, the next, in terms of efficiency, classification algorithm α_i of the ranked sequence $(\alpha_1, \alpha_2, \dots, \alpha_m)$ is selected as the vertex of the second tier, and the procedure of constructing the GF of the third stage is repeated with the only difference that on the the input the TS is already restricted, id est, without training pairs, which are approximated by the GF of the vertex of the first tier and so on.

Therefore, the procedure of constructing an ACT comes down to the repetition of this stage for the following, in view of efficiency, algorithm α_i of the sequence $(\alpha_1, \alpha_2, \dots, \alpha_m)$, to the constant cutting of the TS parts and checking of the branch stop criterion (the empty TS) which in fact signals the completion procedure of constructing an ACT model, and on the output, to the obtaining of the α_i algorithmic tree of classification as well as the tree of GFs f_i.

It must be underlined that there are other possible implementation options of schemes for constructing an ACT of the first type, which differ from the proposed scheme by variations in the number of GFs that are built within each step, criteria and sequence of stages for assessing the quality of classification algorithms, the possibility of using a limited number of algorithms (even one), the possibility to provide an approximation of each of the TS classes with a set of its selected algorithms, with the possibility of varying the criterion for stopping branching in an ACT.

Finally, it is necessary to emphasize that the fundamental peculiarity of such a scheme of constructing an algorithmic classification tree is the possibility to adjust the accuracy (efficiency) of the classification tree model, which is being built during the basic procedure of constructing an ACT. Moreover, it is crucially important to be able to construct an ACT model with the predetermined accuracy with respect to the TS data. Such an opportunity is achieved by limiting the number of steps in the procedure of generating an ACT, by the system of limitations on information capacity, the number and the parameters of generalizing (of the TS area which is approximated) a set of the generalized features that are

constructed at the respective stages of constructing the resultant classification tree [16-21].

IV. EXPERIMENTAL PART

It must be stressed that the proposed scheme for constructing an ACT allows regulating the complexity of a tree model under construction, building models with the predetermined accuracy, yet the structure of the classification tree consists of different-type autonomous classification algorithms as building modules (components). Moreover, the task of selecting a classification tree model among a set of constructed algorithmic classification trees for a specific problem is determined by a set of parameters that are of decisive importance with respect to the current application task (of the TS).

It is apparent that in order to compare and select a particular algorithmic classification tree model from among a fixed set, it is necessary to point to their most important parameters (the dimension of feature space, the number of vertices, transitions, algorithms, etc.) and to identify their error with respect to input array.

It is essential at this stage of the research to consider the quality criteria of the obtained information models, which depend on the error of the model, the power of the initial array of TS, the volume of the test set (the number of training pairs and the feature space dimension of the problem), the number of parameters of the model, etcetera [22,23].

It is evident that critically important parameters of the constructed algorithmic classification tree model that need to be minimized are the model errors, correspondingly, in the TS and test set data, and in each of the classes determined by the initial specification of the current problem.

It should be noted that the key issue which remains is the one related to reducing the complexity of an algorithmic classification tree structure (meaning the number of features, algorithms in the structure of an ACT, the total number of vertices of an algorithmic classification tree model and the total number of transitions in the structure of an ACT), the parameters of the total usage of memory and processing time of the information system. So an important indicator of the quality of the constructed model in the form of a classification tree, taking into account the parameters of the structure of a LCT model, is the overall integral quality indicator presented in the following way:

$$Q_{Main} = \frac{Fr_{All}}{V_{All} \cdot \sum_i p_i} \cdot e^{-\frac{Er_{All}}{M_{All}}}. \qquad (4)$$

Admittedly, in formula (4) the set of parameters p_i represents the most important characteristics of the constructed classification tree, which is estimated:

Er_{All} is the total number of errors of an algorithmic classification tree model within the initial test and training datasets;

M_{All} is the total capacity (volume) of training and test datasets;

Fr_{All} characterizes the number of vertices of the obtained LCT/ACT model with the resulting values f_R (a recognition function, that is, the leaves of the classification tree);

V_{All} represents the total number of all types of vertices in the structure of a LCT/ACT model;

O_{Uz} is the total number of generalized features used in the classification tree model;

P_{All} is the general number of transitions between the vertices in the structure of the constructed classification tree model;

N_{Alg} is the total number of different autonomous classification algorithms which are used in the classification tree model;

It must be pointed out that this integral indicator of the quality of an algorithmic classification tree model will range from zero to one. The lower it is, the worse the quality of the constrctued classification tree is and the higher the index, the better the obtained model is [24-29].

Let us further consider the following example of constructing an algorithmic classification tree model with the corresponding initial parameters:

The fixed set of different-type classification and recognition algorithms $(\alpha_1, \alpha_2, \dots, \alpha_5)$, $(m = 5)$.

The initial training dataset contains information about partitioning R into non-intersecting classes $(H_1, H_2, \dots, H_4)$, $(k = 4)$.

The power of the initial training dataset of the form (1) is 2000 training pairs (the objects of the known classification), $(M = 2000)$.

Each of discrete objects of the training dataset x_i is characterized by a set of features, attributes – $(x_i^1, x_i^2, \dots, x_i^{20})$, $(n = 20)$.

To check the obtained algorithmic classification tree model there is given the test dataset with the capacity elements 500, $(T = 500)$.

The initial training dataset, introduced in this example, contains the data of the component chemical analysis of the content of diesel (carbohydrate) fuel (the task of assessing the quality of fuel) in the simplified version (the number of classes of the training dataset of the problem is reduced to four, the dimension of feature space ranges from 38 to 20, and the number of classification algorithms is also limited at the initial stage by selecting only geometric classifiers) in order to demonstrate the very concept of an algorithmic tree.

It must be also said that in this case, the test set is actually a separate part of the training dataset with objects of the known classification, i. e., the evaluation of the efficiency of the constructed system can also be based on the test dataset, excluding the procedure of approximation of data by means of a set of generalized features (geometric algorithms) and calculation of the integral quality index of the constructed algorithmic classification tree model.

At the first stage of the procedure of constructing a classification tree, we will evaluate the effectiveness of each of the selected classification algorithms, on the basis of which a general classification scheme (an algorithmic classification tree model) will be constructed with regard to the initial training dataset (by the number of generalized features generated by the current algorithm and classification failures).

The cells in Table I show the efffeciency of each of the classification algorithms selected for the current task with regard to the classes of the initial training dataset, the first number being responsible for the number of objects that were denied in terms of classification by the corresponding algorithm (errors, classification failures), and the second –for the number of generalized features (for this type of algorithm – geometric objects) which approximate the corresponding class of the initial dataset. Depending on the initial selection of the algorithm as the vertex of a classification tree (an algorithmic classification tree model), the procedure of constructing the resulting classification scheme may end with a different number of steps.

TABLE I. TABLE I. THE EVALUATION OF EFFICIENCY OF FIXED ALGORITHMS FOR CLASSIFICATION OF DISCRETE OBJECTS WITH REGARD TO THE INITIAL TRAINING DATASET

(Class number / Algorithm type)	***Algorithm*** a_1	***Algorithm*** a_2	***Algorithm*** a_3	***Algorithm*** a_4	***Algorithm*** a_5
Class H_1	0/32	0/12	0/11	0/9	18/10
Class H_2	0/16	12/17	1/16	14/6	12/8
Class H_3	0/8	0/10	0/17	0/12	0/10
Class H_4	0/11	15/3	9/16	16/6	14/11

Table 1 illustrates that the efficiency of all algorithms, with the exception of algorithm a_5 (a geometric algorithm of hyperplanes) regarding class H_1 is 100%, that is why for its approximation an arbitrary algorithm can be applied (certainly, except a_5). At all subsequent stages of constructing the recognition scheme (tiers of an algorithmic classification tree structure), it is advantageous to fix again the algorithm a_1 (a geometric algorithm of hyperspheres), which proved to be the most efficient and economical in relation to all other classes of the initial dataset; its peculiarity is a high degree of universality in terms of the possibility of constructing a generalized feature even in those cases when other geometric algorithms cannot do this and represent a high percentage of classification failures (errors) of classification for the objects of the initial dataset (a case of complicated, confusing arrangement of classes in feature space of the problem).

An important point is also the fact that each of the generalized features generated by the classification algorithm a_1 represents a set of coordinates of the center of hypersphere (in feature space of the problem) and its radius requires the minimum amount of memory of the information system for its storage and simple mechanisms for working with a set of such GFs. One of the possible constructed algorithmic classification tree schemes is presented in (Fig. 2).

It should be noted that a classification tree model (Fig. 2) was constructed with the help of algorithms whose efficiency was estimated with respect to the number of GFs (which approximate the initial training dataset), and only three classification algorithms were sufficient for the complete approximation of the training dataset of this problem. So for the approximation of class H_1 (from the first step to the third step) two algorithms were applied – initially algorithm a_5 (a geometric algorithm of hyperplanes) built a generalized feature S_1^1 which only partially described it. At the second stage, the algorithm a_4 (a geometric algorithm of hyper-parallelepipeds) is applied – features P_1^2 and P_1^3, which ultimately completed the recognition (approximation) of the class H_1.

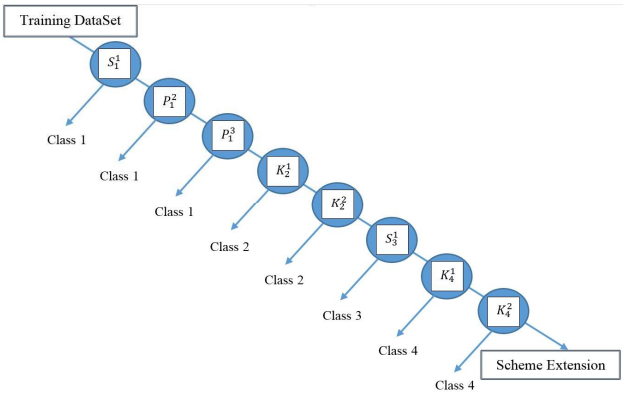


Fig. 2. An example of a constructed algorithmic classification tree model.

At the next stages of constructing an ACT model, the classification algorithm a_1 (a hyperspheres algorithm) (features $K_1^2, K_2^2, K_4^1, K_4^2$) and the algorithm a_5 (the generalized feature S_3^1) are applied again.

Attention should be paid here to the fact that, for constructing this classification scheme (Fig. 2), there were applied three different recognition algorithms (out of five selected initially), which do not affect each other's work directly, and it means that in their place there could have been completely different in principle and ideology classification algorithms from which one can construct a recognition scheme (an ACT model) of arbitrary complexity and efficiency. Only the efficiency of each of them within a fixed dataset and the information capacity of the generalized features generated by them are important [30-34].

It must be emphasized that the algorithmic tree method manipulates only the ready-made (constructed) generalized features, and it may be of no importance for this method in what way or with the help of what algorithm (rule, method) they were obtained. Moreover, an ACT scheme (Fig. 2) shows a step-by-step scheme of generating not single but entire sets of GFs K_i^j, S_i^j, P_i^j, where j is the number of the GF for the corresponding approximation class (the number of the stage of generating an ACT regarding the class), and i is the step number of the procedure of constructing a tree (of the class which is approximated).

It is obvious that this recognition scheme (an algorithmic classification) constructed on the basis of the logical tree method (a logical classification tree), and can be represented as a certain algorithm scheme (operator) constructed with the help of some algorithm of minimization or maximization of the corresponding functional, on the basis of which there are estimated the importance of a feature, group of features or the efficiency of an autonomous classification algorithm which is definitely related to classification errors (goes along with methods of logical classification trees).

It should be stressed that the method of an algorithmic tree, based on the input data (a training dataset) and an assorted set of algorithms for forming GFs stored in its library, constructs (generates) the optimal, in view of memory consumption (complexity) and recognition efficiency (system), a certain

scheme (a classification tree). By the scheme in this case we understand a set of numerical parameters for the set of GFs K_i^j, S_i^j, P_i^j, which approximate the array of the initial data in the best possible way. Thus, in our case, the arguments of the constructed recognition scheme are features of classes (hyperspheres, hyper-parallelepipeds, and others) or interclass features (hyperplanes). The parameters of these features and the overall structure of an algorithmic classification tree (a classification scheme) are stored in the memory of the computer (the information system).

Further in this research, we evaluate and compare the constructed models of classification trees (LCTs and ACTs) for the classification problem presented above. Thus, a fragment of the main results of the above experiments (of the comparative tests of the methods for constructing LCTs models using the dataset of this application problem), presented in (Table. II), and the synthesized models of different-type classification trees provided the necessary level of accuracy specified by the problem, speed and the working storage consumption of the information system, but showed a different structural complexity of constructing classification trees (models) and a set of generalized features (in the case of an ACT model).

TABLE II. THE COMPARATIVE TABLE OF MODELS / METHODS OF CLASSIFICATION TREES.

№	**Method of synthesis of a logical tree structure**	**Integral indicator of the quality of the model Q_{Main}**	**Total number of errors of the model on the TS and testing dataset Er_{All}**
1	The Full LCT method on the basis of elementary features selection	0,004789	2
2	The LCT model with one-time evaluation of feature importance	0,002263	3
3	The limited method of constructing the LCT	0,003181	2
4	The algorithmic tree method (type I)	0,005234	0
5	The algorithmic tree method (type II)	0,002941	0

V. CONCLUSIONS

The practical value of the obtained results lies in the fact that the proposed method of constructing ACT models makes it possible to build economical and efficient classification models of the specified accuracy (this method was implemented in the library of algorithms of the system "ORION III" for solving various applied classification problems) which are characterized by a high degree of universality regarding a wide range of applications. It should also be noted that the practical application has confirmed the functionality of the constructed models of classification trees and the developed software. As part of the follow-up research activities, there can be mentioned the further development of ACT methods (the introduction of new types of classification trees), optimization of software implementations of the proposed ACT method, as well as its practical testing on many real classification and recognition problems.

Therefore, in view of all the above said, the following points can be made.

The use of the ACT concept in constructing models of classification trees allows achieving good results in terms of expanding the range of application problems (universality requirement), the possibility to build classification models, whose accuracy can be adjusted in the process of construction (or to build systems with the predetermined accuracy), the possibility of rational use of the already accumulated potential of recognition methods and algorithms.

The ACT concept allows constructing models of classification trees of different types (they are not limited to the proposed, in the specified, first type of ACT graph schemes), yet all of them are based on the simple idea of approximating the initial training dataset using a set of autonomous classification algorithms and the representation of the obtained model in the form of some tree-like scheme.

An important point is that the ACT model which was obtained using different classification algorithms and methods, in turn, represents a new recognition algorithm, id est, the ACT concept is the method of synthesizing new classification algorithms of the specified accuracy with regard to the training dataset based on the set of already known ones.

REFERENCES

[1] J.R. Quinlan, "Induction of Decision Trees", Machine Learning, no. 1, 1986, pp. 81–106.

[2] I. Povhan, "General scheme for constructing the most complex logical tree of classification in pattern recognition discrete objects." Collection of scientific papers Electronics and information technologies. Lviv, vol. 11, 2019, pp. 112–117.

[3] R. Srikant and R. Agrawal, "Mining generalized association rules", Future Generation Computer Systems, vol. 13, No. 2, 1997, pp. 161–180.

[4] Y.A. Vasilenko, E.Y. Vasilenko and I.F. Povkhan, "Conceptual basis of image recognition systems based on the branched feature selection method", European Journal of Enterprise Technologies, No. 7(1), 2004, pp. 13–15.

[5] I.F. Povkhan "Features random logic of the classification trees in the pattern recognition problems", Scientific notes of the Tauride national University. Series: technical Sciences, Vol. 30(69), No. 5, 2019, pp. 152–161.

[6] V.O. Laver and I.F. Povkhan "The algorithms for constructing a logical tree of classification in pattern recognition problems", Scientific notes of the Tauride national University. Series: technical Sciences, Vol. 30(69), No. 4, 2019, pp. 100–106.

[7] P.E. Vtogoff "Incremental Induction of Decision Trees", Machine Learning, No. 4, 1989, pp. 161–186.

[8] Y.A. Vasilenko, E.Y. Vasilenko and I.F. Povkhan, "Branched feature selection method in mathematical modeling of multi-level image recognition systems", Artificial Intelligence, No. 7, 2003, pp. 246–249.

[9] I.F. Povkhan "Features of synthesis of generalized features in the construction of recognition systems using the logical tree method", Materials of the international scientific and practical conference "Information technologies and computer modeling ITKM-2019". Ivano-Frankivsk, pp. 169–174, 2019.

[10] Y.A. Vasilenko, E.Y. Vasilenko and I.F. Povkhan, "Defining the concept of a feature in pattern recognition theory", Artificial Intelligence, No. 4, 2002, pp. 512–517.

[11] S.A. Subbotin "Construction of decision trees for the case of low-information features", Radio Electronics, Computer Science, Control, No. 1, 2019, pp. 121–130.

[12] Y.A. Vasilenko, F.G. Vashuk, I.F. Povkhan, M.Y. Kovach, O.D. Nikarovich "Minimizing logical tree structures in image recognition tasks", European Journal of Enterprise Technologies, 3(9), 2004, pp. 12–16.

[13] T. Hastie, R. Tibshirani and J. Friedman, “The Elements of Statistical Learning”. Berlin, Springer, 2008.

[14] I. Povhan, “Designing of recognition system of discrete objects”, IEEE First International Conference on Data Stream Mining & Processing (DSMP), Lviv, Ukraine, pp. 226–231, 2016.

[15] L.L. Breiman, J.H. Friedman, R.A. Olshen and C.J. Stone, “Classification and regression trees.” Boca Raton, Chapman and Hall/CRC, 1984.

[16] I.F. Povhan “The problem of general estimation of the complexity of the maximum constructed logical classification tree”, Bulletin of the national technical University Kharkiv Polytechnic Institute, No. 13, 2019, pp. 104–117.

[17] Y.A. Vasilenko, F.G. Vashuk and I.F. Povkhan, “The problem of estimating the complexity of logical trees recognition and a general method for optimizing them”, Scientific and technical journal “European Journal of Enterprise Technologies”, No. 6/4(54), 2011, pp. 24–28.

[18] Y.A. Vasilenko, F.G. Vashuk and I.F. Povkhan, “General estimation of minimization of tree logical structures”, European Journal of Enterprise Technologies, No. 1/4(55), 2012, pp. 29–33.

[19] H. Deng, G. Runger and E. Tuv, “Bias of importance measures for multi-valued attributes and solutions”, Proceedings of the 21st International Conference on Artificial Neural Networks (ICANN), pp. 293–300, 2011.

[20] B. Kamiński, M. Jakubczyk and P. Szufel, “A framework for sensitivity analysis of decision trees”, Central European Journal of Operations Research, No. 26 (1), 2017, pp. 135–159.

[21] K. Karimi and H.J. Hamilton, “Generation and Interpretation of Temporal Decision Rules”, International Journal of Computer Information Systems and Industrial Management Applications, vol. 3, 2011, pp. 314–323.

[22] I.F. Povkhan “The problem of functional evaluation of a training sample in discrete object recognition problems”, Scientific notes of the Tauride national University. Series: technical Sciences,vol. 29(68), No. 6, 2018, pp. 217–222.

[23] Y. Bodyanskiy, O. Vynokurova, G. Setlak G. and I. Pliss, “Hybrid neuro-neo-fuzzy system and its adaptive learning algorithm”, Xth Scien. and Tech. Conf. "Computer Sciences and Information Technologies" (CSIT), Lviv, pp. 111–114, 2015.

[24] S.B. Kotsiantis “Supervised Machine Learning: A Review of Classification Techniques”, Informatica, No. 31, 2007, pp. 249–268.

[25] S. Subbotin and A. Oliinyk, “The dimensionality reduction methods based on computational intelligence in problems of object classification and diagnosis”, Recent Advances in Systems, Control and Information Technology, [eds.: R. Szewczyk, M. Kaliczyńska]. Cham, Springer, 2017, pp. 11–19. (Advances in Intelligent Systems and Computing, Vol. 543).

[26] S.A. Subbotin, “Methods and characteristics of localitypreserving transformations in the problems of computational intelligence”, Radio Electronics, Computer Science, Control, No. 1, 2014, pp. 120–128.

[27] H. Koskimaki, I. Juutilainen, P. Laurinen and J. Roning, “Two-level clustering approach to training data instance selection: a case study for the steel industry”, Neural Networks: International Joint Conference (IJCNN-2008), Hong Kong, pp. 3044–3049, 1-8 June 2008. DOI: 10.1109/ijcnn.2008.4634228

[28] S. Subbotin, “The neuro-fuzzy network synthesis and simplification on precedents in problems of diagnosis and pattern recognition”, Optical Memory and Neural Networks (Information Optics), vol. 22, No. 2, 2013, pp. 97–103. DOI: 10.3103/s1060992x13020082

[29] S.A. Subbotin, “Methods of sampling based on exhaustive and evolutionary search”, Automatic Control and Computer Sciences, vol. 47, No. 3, 2013, pp. 113–121. DOI: 10.3103/s0146411613030073

[30] R. L. De Mántaras, “A distance-based attribute selection measure for decision tree induction”, Machine learning, vol. 6, No. 1, 1991, pp. 81–92.

[31] H. Deng, G. Runger and E. Tuv, “Bias of importance measures for multi-valued attributes and solutions”, 21st International Conference on Artificial Neural Networks (ICANN), Espoo, Berlin, Springer-Verlag, vol. 2, pp. 293–300, 14–17 June 2011.

[32] A. Painsky and S. Rosset, “Cross-validated variable selection in tree-based methods improves predictive performance”, IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 39, No. 11, 2017, pp. 2142–2153. DOI:10.1109/tpami.2016.2636831

[33] M. Miyakawa, “Criteria for selecting a variable in the construction of efficient decision trees”, IEEE Transactions on Computers, Vol. 38, No. 1, 1989, pp. 130–141.

[34] M. Lupei, A. Mitsa, V. Repariuk and V. Sharkan, “Identification of authorship of Ukrainian-language texts of journalistic styleu sing neural networks”, Eastern-European Journal of Enterprise Technologies, 1, 2020, pp. 30–36. doi: https://doi.org/10.15587/1729-4061.2020.195041

IEEE Third International Conference on Data Stream Mining & Processing
August 21-25, 2020, Lviv, Ukraine

Development of an Adaptive Module of the Distance Education System Based on a Hybrid Neuro-Fuzzy Network

Mykola Pikuliak
Department of Information Technology
Vasyl Stefanyk Precarpathian National University
Ivano-Frankivsk, Ukraine
mykolapikulyak@gmail.com

***Abstract* – A fuzzy neural network method for constructing an adaptive module of a distance knowledge transfer system has been suggested. Experimental modeling with the use of fuzzy ANFIS neural network in MATLAB program based on a hybrid algorithm, which allowed to make it easier to train the network and improve the quality of distance education.**

Keywords – neuro-fuzzy network; educational system; adaptive module; fuzzy rules; learning procedure; Anfis network.

I. INTRODUCTION

Presently for creation of the adaptive educational systems that engage in the construction of individual educational trajectories, the wide spectrum of methods, models and tools is used. It is such technologies as probability theory and mathematical statistics, theory of the graphs, theory of fuzzy sets and fuzzy logic, theory of making decision and analysis of operations, combinatorics topology and theory of fractals and many other. However, for most systems created on the basis of the adopted technologies, characteristic is an insufficient level of adaptivity of serve of educational content and taking into account of individual characteristics of every student.

At the same time, technologies that allow to solve a wide range of research results through the use of hybrid systems are becoming more and more widespread in intellectual training systems, since they allow the use of traditional tools and methods of artificial intelligence on the one hand, and the use of new models on the other hand and instrumental tools based on the integration of already known techniques.

Such technologies are based on the theoretical directions of informatics which are rapidly developing such as distributed computing, theory of intellectual and multi agent systems, neural network and evolutionary algorithms, methods of effective information retrieval intellectual data analyses.

II. ANALYSIS OF RECENT RESEARCH AND PUBLICATIONS

As shows the analysis of recent publications, many intelligent systems built on the hybrid approach are known today, and neuro-fuzzy systems are becoming more widely used to solve structural synthesis problems. Such systems differ in architecture, as well as in algorithms and models, but they are characterized by the following key properties:

- the ability to automatically generate a set of output rules;
- the application of different training algorithms;
- the ability of online training in the process of receiving data;
- the possibility of structure change;
- the preserving the knowledge embedded in the system of knowledge in the process of parametric optimization or learning of new rules [1].

In particular, hybrid neural-fuzzy systems and wavelet-neuro-fuzzy systems, such as Wang-Mendel architectures, are increasingly being used to solve a wide range of problems [2], adaptive neuro-phases of the Takagi-Sugeno-Kang system [3] are increasingly used to solve a wide range of problems of wavelet-neuro-phase networks [4], adaptive wavelet-neuro-phase systems with W -neurons [5].

Hybrid systems allow to combine formalized and non-formalized knowledge more efficiently, resulting in higher accuracy of calculations, without losing the ability to their function in real time. Therefore, the task of researching of such systems for construction of individual trajectories of education systems is actual.

In particular, in the intellectual educational systems, application of hybrid methods allows more qualitatively to execute working of parameters of student model, that represent the basic level of education, material assimilation rate, time spent studying the course, etc. Due to it in the educational system the automatic construction of flexible student trajectories will be realized with the corresponding filling of educational material.

Therefore, the task of researching of hybrid systems for construction of individual trajectories of education systems is actual.

III. FORMULATING THE GOALS OF THE ARTICLE

The purpose of this study is to use the apparatus of neuro-fuzzy sets (NFS) to build an adaptive module in an

automated learning system. Since the possible learning trajectories in the work are researched and known, and their number is fixed, from a mathematical point of view, the processing of learning outcomes is a classification problem. Therefore, to solve this problem, a hybrid system ANFIS is proposed.

The fact is that neural networks are usually trained as a rule on the basis of the use of universal algorithms, but they have limitations on the number of input variables. When fuzzy output systems are used, time-consuming processes of formalization of concepts, construction of membership functions and formation of rules of output are used, so the process of work of the training network is difficult to understand. The application of the NFS allows you to train the system based on the use of a neural network, and to adjust the accessory functions using a fuzzy output system.

The implementation of the proposed technology on the basis of a hybrid architecture in an adaptive learning system makes it possible to solve the following problems:

- to analyze the current state of student's success with the use of fuzzy inference system;

- to predict possible variants of a student's educational trajectory by using of neuron network;

- to design a new structural and functional solution of the student's educational behavior by using a hybrid algorithm for processing the rules of a fuzzy base of rules and parameters of membership functions.

IV. PRESENTING THE MAIN MATERIAL

The fuzzy model of the adaptive module proposed in the paper is represented by a system of fuzzy logical inference, which allows to control the student's educational trajectories, according to his current level of mastering new knowledge.

The basis for the construction of such a model is the analysis of 5 parameters P_i (P_1 – the total level of assimilation of educational material; P_2 – the depth of knowledge; P_3 – the degree of assimilation of the material; P_4 – the quality of assimilation of the material; P_5 – the time spent on training), the numerical values of which we get from the student module [6]. Each of the parameters of the student model P_i takes values from the set $\{L, M, H\}$ defined in the interval [0..1] ($L \rightarrow$ "low level" – [0..0,4), $M \rightarrow$ "medium level" – [0,4..0,8) and $H \rightarrow$ "high level" – [0,8..1] respectively).

Each fuzzy rule is described by a linguistic variable R_j, the term-set of which is a set of $\varphi(R_j)$, values, which reflects the probabilities of deciding whether to choose a training mode of overtraining R_1, beforeretraining R_2 or training R_3 [6].

Such a model should be presented in the form of a terminal quantum of the second level [7], which simplifies its further processing with the help of formulas of mathematical logic:

$$vk_2R_j = \begin{bmatrix} P_1 \left| M : P_2 \right| L : \ldots : P_i \left| L : \varphi(R_1)(\rightarrow) = 0.1 \right. \\ \ldots \\ P_1 \left| M : P_2 \right| H : \ldots : P_i \left| M : \varphi(R_2)(\rightarrow) = 0.4 \right. \\ \ldots \\ P_1 \left| H : P_2 \right| M : \ldots : P_i \left| H : \varphi(R_3)(\rightarrow) = 0.9 \right. \end{bmatrix} \quad (1)$$

Each row of matrix (1) represents a relationship between the possible values of the P_i parameters and the implication values R_j corresponding to them. For example, entry:

$$P_1 \left| L : P_2 \right| M : P_3 \left| L : P_4 \right| L : P_5 \left| M : \varphi(R_1)(\rightarrow) = 0.6 \right. \quad (2)$$

means that with the values of parameters $P_1 = L$ and $P_2 = M$ and $P_3 = L$ and $P_4 = L$ and $P_5 = M$, the probability of choosing a overtraining mode is $\varphi(R_1) = 0.6$.

For modeling, the proposed approach uses the Adaptive-Network-Based Fuzzy Inference System (ANFIS) structure [8], implemented in the MATLAB Fuzzy Logic Toolbox extension package, which allows to combine neural network approach (learning abilities) with mathematical apparatus of fuzzy logic.

The advantage of using the ANFIS editor is the ability to automatically synthesize neural networks with experimental data. In addition, ANFIS has, in comparison to other methods, high learning speeds and simplicity of algorithms used for modeling.

The use of this structure in adaptive systems makes it possible to build a model of automatic formation of educational mode for an individual student, depending on the current parameters of his success. In this case, the membership functions are tuned (trained) to minimize the variance between fuzzy modeling results and expert data.

In the built neuro-fuzzy model, the ANFIS network consists of the following five consecutive layers:

the first layer determines the fuzzy terms of the input parameters, that is, it converts the input signal to an input signal into a fuzzy variable using accessory functions (phasing operation). At its input comes a 5-dimensional vector $P_i(P_1, P_2, P_3, P_4, P_5)$, which reflects the current level of acquired knowledge.

Therefore, the first layer contains 5 inputs of layers and 5n (n=3 – number of terms L, M, H) of the membership functions φ_{ij}, $i = 1,2,3,4,5$, $j = 1,\ldots,3$, and performs the phasing of the input data – Fig. 1.

At the output of the first layer we get the value of the function of accessory at specific values of the inputs P_i. Thus, if a P_i vector is given at the input to the network, then the elements of the first layer carry out its phasing, calculating the levels of accessories $0 < \varphi_{ij} \leq 1$.

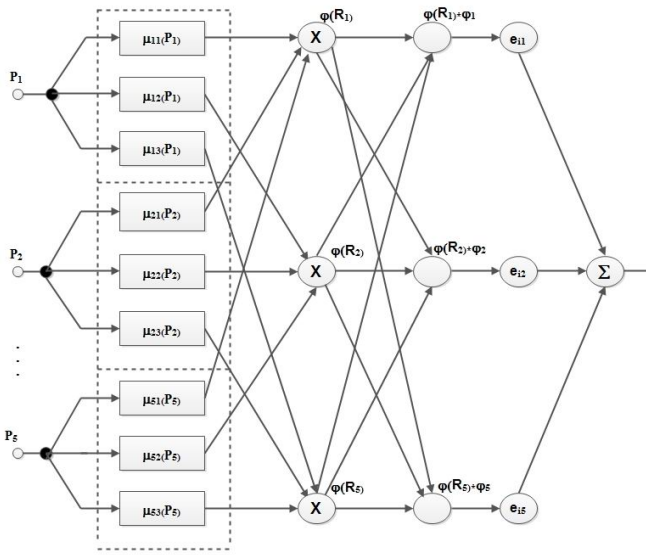


Fig. 1. General structure of the ANFIS network

In the second layer are neurons that characterize the truth values of the input variables $\mu(P_i)$ at specific values of the inputs P_i. Each vertex of this layer is adaptive, that is, the truth parameters of the input variables $\mu(P_i)$ can be changed during the learning process of the network.

Each node of the second layer is connected to those nodes of the first, which form the truth indicators of the corresponding rule. The rules implement a fuzzy logical operation "AND". The outputs of the neurons of this layer are the degrees of veracity of the links of each rule of the system, calculated by the formulas:

$$\varphi(R_1) = \min\{\mu_{11}(P_1), \mu_{21}(P_2), \mu_{31}(P_3), \mu_{41}(P_4), \mu_{51}(P_5)\}$$
$$\varphi(R_2) = \min\{\mu_{12}(P_1), \mu_{22}(P_2), \mu_{32}(P_3), \mu_{42}(P_4), \mu_{52}(P_5)\}$$
$$\varphi(R_3) = \min\{\mu_{13}(P_1), \mu_{23}(P_2), \mu_{33}(P_3), \mu_{43}(P_4), \mu_{53}(P_5)\}.$$

The third layer normalizes the degree of implementation of the rules of the second layer. In other words, the output of a neuron in this layer calculates the relative degree (weight) of the execution of a fuzzy rule by the formula:

$$\bar{\varphi}(R_j) = \frac{\varphi(R_j)}{\sum_{j}^{k} \varphi(R_j)}$$

Adaptive nodes of the fourth layer calculate the contribution of each fuzzy rule to the output of the network by the formula:

$$e_{ij} = \bar{\varphi}(R_j) \cdot \varphi_i; i = 1,...,5; j = 1,...,3,$$

where e_{ij} is the contribution of each fuzzy rule to the network output; $\bar{\varphi}(R_j)$ is the relative degree of implementation of the j fuzzy rule; φ_i is a clear number that determines the output of each i rule:

$$\varphi_i = \sum_{i=j}^{n} R_j(P_i), i = 1,...,k.$$

Each unit of the fourth layer is connected to one unit of the third layer, as well as all with all the inputs of the network.

Finally, the fifth (source) layer implements a defragmentation function, which determines the contribution of all the rules. In this layer, the output is generated based on fuzzy rules (4th layer):

$$e = \sum e_{ij}$$

where e is network output; $\sum e_{ij}$ is the total contribution of all the fuzzy rules.

The suggested adaptive function is resembles fuzzy inference mechanism that, for each variant of the rule base, either builds a trajectory for one of the learning modes or issues an inability to continue learning.

V. THE RESULTS OF THE RESEARCH

After comparing the results of the training of the constructed network with different algorithms, a hybrid algorithm was used, which used a combination of gradient descent in the form of an error propagation algorithm with error rate 0 and number of cycles 100 and the least squares method [9] – Fig. 2.

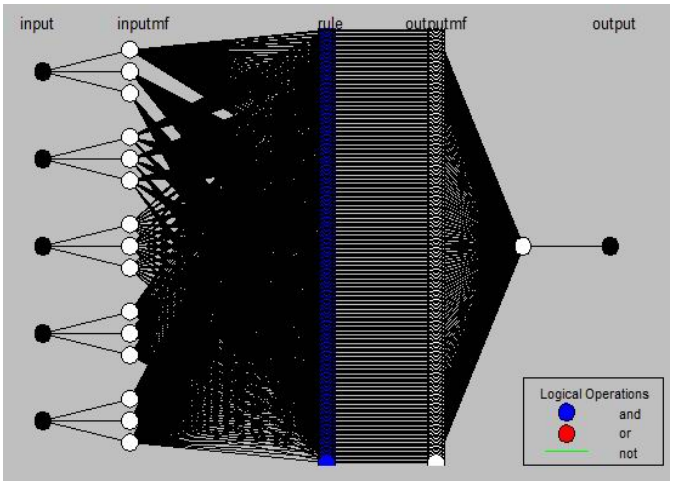


Fig. 2. The architecture of the neuro-fuzzy output of the ANFIS editor

The algorithm of the reverse extension of an error adjusts the parameters of antecedents of rules (accessory functions). With the method of the smallest squares the coefficients of conclusions of rules is estimated because they are linearly connected with the output of the network. Each iteration of the tuning procedure is performed in two steps:

1) a training sample is presented to the input of the network with fixed values of nonlinear parameters P_1, P_2, P_3, P_4, P_5 the first layer of neurons.

This signal is propagated over the network in the forward direction, and the values of the output signals of the intermediate layers and the output signal of the fourth layer of the network are calculated sequentially.

Based on the deviation between the desired and existing (real) behavior of the network using the iterative method of the smallest squares, the values of the linear parameters e_j of the fourth layer of the network are searched;

2) when using the error propagation method, the initial response $\varphi_a(R_j)$ is compared with the reference value $\varphi_e(R_j)$:

$$E = \frac{1}{2}\sum_{j=1}^{m} (\varphi_e(R_j) - \varphi_a(R_j))^2,$$

where m is the number of rules in the training sample;

$\varphi_e(R_j)$ is the desired (reference) value of the membership function of the j rule;

$\varphi_a(R_j)$ is calculated by the neural network value of the level of activation of the j rule (the resulting value of the membership function).

The learning process is to change the weighting coefficients between the neurons so that the total mean-square error of the neural network for vector P_i is minimal (9).

At the same time, the coefficients of the conclusions of the rules found at the first stage do not change. The iterative tuning procedure continues until the deviation exceeds the preset error value.

Consider the work of the proposed model on the basis of a demonstration example. As input, we take the values of the parameters of the vector P_i, which are proposed in formula (2):

$P_1 = 0.2$ – corresponds to the low term L of the total level of assimilation of educational material;

$P_2 = 0.4$ – corresponds to the mean term M of the depth of knowledge;

$P_3 = 0.25$ – corresponds to the low term L of the degree of assimilation of the material;

$P_4 = 0.35$ – corresponds to the low term L of the quality of assimilation of the material;

$P_5 = 0.52$ – corresponds to the mean term M of the time spent on training.

As a result of the experimental studies of different groups and students of different levels of knowledge, each parameter P_i is analyzed in terms of the influence on the likelihood of choosing a training mode and the degree of connection with other parameters.

According to the modeled educational sample modeled in this way, a student with parameter values corresponding to formula (2) should be sent to the educational mode R_1 (overtraining) with the possible involvement of additional portions of knowledge.

The appearance of the ANFIS editor with the downloaded data is shown in Fig. 3.

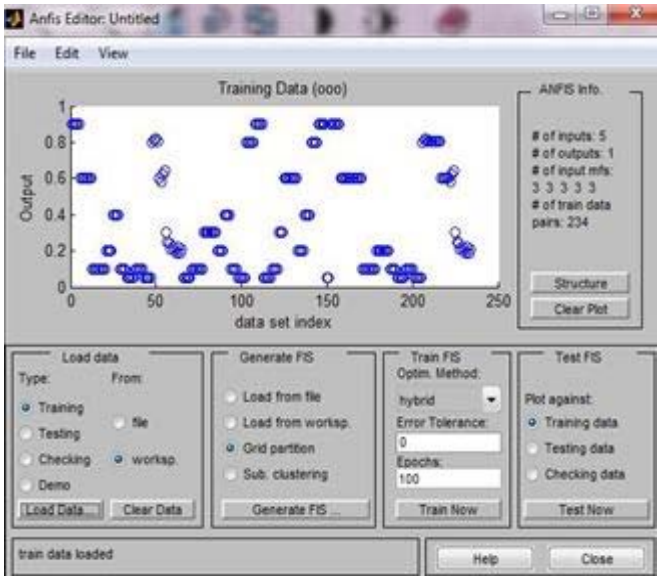


Fig. 3. The graphical interface of the ANFIS editor after downloading training data

The learning results of the network and the interface of the generated rules window are shown in Fig. 4 and Fig. 5 respectively:

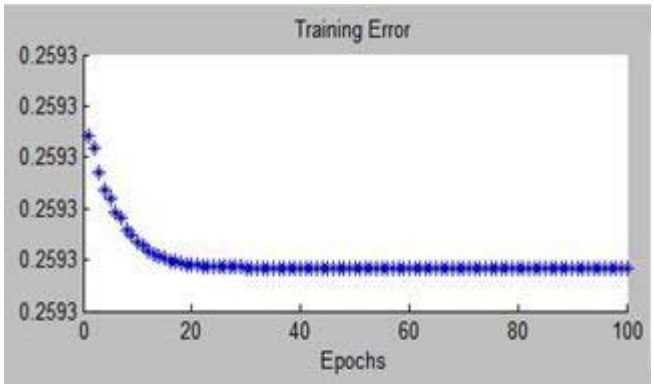


Fig. 4. Schedule of learning error versus the number of learning cycles

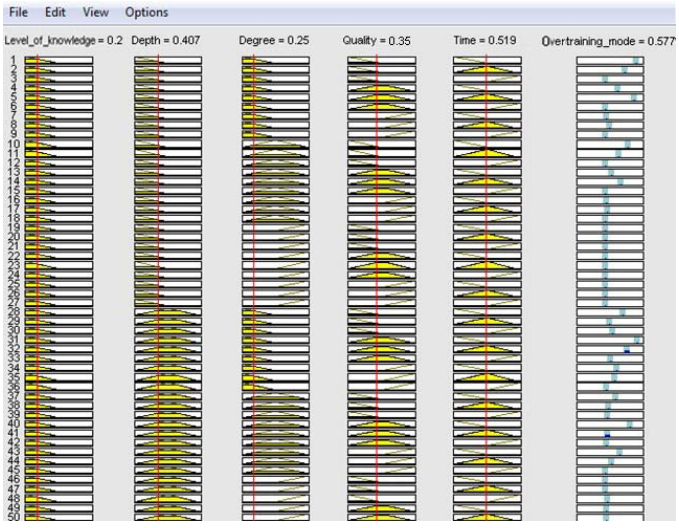


Fig. 5. The interface of the generated rules window in the ANFIS editor

As the analysis of the error chart shows, the training of the network was completed almost after the 20th step. The

value of the original variable is 0.577 (error 0.023), which is almost the same as the existing (experimental) result.

In Fig. 6 the surface of fuzzy logic output, which allows to trace the dependence of different parameters of the student model for the overtraining mode are shown.

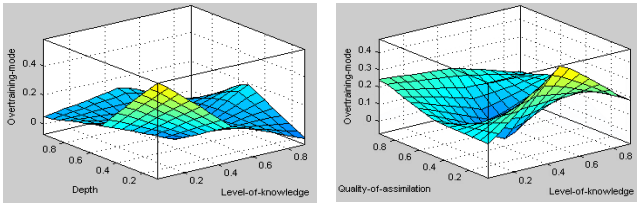


Fig. 6. ANFIS network fuzzy output surfaces using hybrid algorithm for overtraining mode

As can be seen from the figure, a uniform and monotonous plot of the dependence of the "output surface" indicates the good "quality" of the mechanism of output and the sufficiency and consistency of the used rules of output.

An analogous method was used to analyze the effectiveness of the proposed model for beforeretraining and training modes. The deviation errors between the experimental data and those calculated in ANFIS did not exceed the allowable values.

Thus, the proposed neuro-fuzzy inference system is an effective tool for modeling adaptive trajectories in training systems, because it allows to make control effects of the system depending on different values of the input parameters of the student model.

And using the hybrid method allows you to divide the learning process into two separate stages. This leads to a reduction in the number of mathematical operations and, accordingly, an increase in the speed of calculations. This makes this method much more efficient than, for example, the conventional gradient method, whereby all parameters of the network are refined in parallel and simultaneously.

The achieved results give grounds for claiming the possibility of practical application of the proposed approach of constructing the adaptive module at the stage of tuning the input fuzzy parameters of the training system.

It should be noted that the initial set of rules is rather excessive and may contain conflicting rules with the same left part and different right part. Therefore, the use of optimized simulated rules which can be done both on the basis of expert evaluations of the rules and by adapting to existing training data has been foreseen in the system.

In order to reduce the number of rules in the work, the theory of partitioning of natural numbers into the whole inalienable items, which allows us to construct 21 basic rules and group into 6 categories (Table I) [10].

The table shows which configurations of values of P_i parameter describe separate group. For example, the third group includes rules that characterize the average level of acquired units of information, since there is no single parameter from the L set, but possible values from the M set and the B set, but maximum only two parameters can be from the H set.

TABLE I. GROUP CLASSIFICATION OF STUDENTS BOUND TO THE MODE OF STUDY

№	Group	$L+M+H=5$	Probability of mode selection		
			R_1	R_2	R_3
1	1	5 0 0	0,9	0,05	0,05
2		4 1 0	0,8	0,1	0,1
3		3 2 0	0,6	0,3	0,1
4		2 3 0	0,3	0,6	0,1
5		4 0 1	0,8	0,1	0,1
6	2	3 1 1	0,6	0,2	0,2
7		2 2 1	0,4	0,4	0,2
8		2 1 2	0,4	0,2	0,4
9		1 4 0	0,1	0,8	0,1
10		1 3 1	0,2	0,6	0,2
11	3	0 5 0	0,05	0,9	0,05
12		0 4 1	0,1	0,8	0,1
13		0 3 2	0,1	0,6	0,3
14	4	1 1 3	0,2	0,2	0,6
15		1 2 2	0,2	0,4	0,4
16	5	0 0 5	0,05	0,05	0,9
17		0 1 4	0,1	0,1	0,8
18		1 0 4	0,1	0,1	0,8
19		0 2 3	0,1	0,3	0,6
20	6[a]	3 0 2	0,05	0,05	0,05
21		2 0 3	0,05	0,05	0,05

[a.] Class of non-existent (anomalous) rules

The 5-category built-in grouping allows to simplify the further data processing and making decisions by applying a learning trajectory related to the repetition or in-depth learning of undrained portions of knowledge to each group. This enables the system to manage the most typical student transition probabilities across learning modes.

VI. CONCLUSIONS

The proposed information technology for the use of neuro-fuzzy networks for the construction of the adaptive module is in our opinion effective because it improves the quality of training systems due to the high speed of calculations and the possibility of on-line training in the process of receiving new data.

Iterative use of the described hybrid algorithm in learning network layers provides a high enough chance of achieving the required learning precision for a finite number of iterations.

The developed grouping of rules allows to eliminate difficulties connected with building of a complete training selection and the necessity for using of a large number of neurons.

The built-in matrix of input rules and the obtained surfaces of fuzzy output allow, on the basis of information about the current level of knowledge of the student, to simulate individual trajectories and to realize automatic control of learning modes in adaptive systems.

REFERENCES

[1] A.A. Kuzkin, “Assessing the performance and efficiency indicators of IT processes using hybrid neuro-fuzzy networks,“ Naukovedenie, vol. 1, January–February 2014, URL: http://naukovedenie.ru/PDF/57TVN114.pdf.

[2] L.Wang, Adaptive Fuzzy Systems and Control: Design and Stability Analysis. New Jersey: Prentice Hall, 1994.

[3] J.-S. R. Jang, C.-T. Sun, E. Muzutani, Neuro-Fuzzy and Soft Computing: A Computational Approachto Learning and Machine Intelligence. Upper Saddle River, N.J.: PrenticeHall, Inc., 1997.

[4] R.H. Abiyev, O. Kaynak, "Fuzzy waveletneural networks for identification and control of dynamic plants – A novel structure and a comparativestudy," IEEE Trans. OnIndustrial Electronics, vol. 55 (8), pp. 3133–3140, 2008.

[5] Ye.V. Bodyanskiy, O.A. Vynokurova, "Intelligent data processing based on hybrid wavelet-neuro-phase system on adaptive W-neurons," Scientific papers. Computer Technology. Mykolayiv: Publishing house Petro MogylaChSU, vol. 104 (117), pp. 88–98, 2009.

[6] M.V. Pikuliak, "Information technology of adaptive testing model construction in distance learning systems," Bulletin of Khmelnitsky National University. Series: Technical Sciences, vol. 4, pp. 153-158, 2018.

[7] I. Siroja, Quantum models and artificial intelligence methods for decision making and management. Kiev: Scientific thought, 2002.

[8] S.D. Stovba, Designing fuzzy systems with MATLAB. Moscow: Hotline – Telecom, 2007.

[9] MATLAB Fuzzy Logic Toolbox User's Guide. The MathWorks, Inc, 2008.

[10] M.V. Pikuliak, "Application of vehicle fuzzy is sets is for the construction of model of the adaptive education al system," Polish journal of science, vol.1, n. 14, pp. 39-44, 2019

IEEE Third International Conference on Data Stream Mining & Processing
August 21-25, 2020, Lviv, Ukraine

New Approaches in the Learning of Complex-Valued Neural Networks

Vladyslav Kotsovsky
Department of Information Management Systems and Technologies
Uzhhorod National University
Uzhhorod, Ukraine
kotsavlad@gmail.com

Anatoliy Batyuk
Department of Automated Control Systems
Institute of Computer Science and Information Technologies
Lviv Polytechnic National University
Lviv, Ukraine
abatyuk@gmail.com

Maksym Yurchenko
Department of Information Management Systems and Technologies
Uzhhorod National University
Uzhhorod, Ukraine
maxskywalker94@gmail.com

***Abstract*—We consider neural networks with complex weights and continuous activation functions. The complex generalization of the backpropagation learning algorithm is studied in the paper. We introduce new kinds of activation functions, namely, the complex modifications of the rational sigmoid and the ReLU activation function, the use of which has two main benefits. The first is that the application of these functions allows to avoid the splitting of transfer functions. The second is that they are fast to compute. In order to improve the performance and to increase the training speed we use the complex version of the modern optimizers instead of the classical techniques based on the application of gradient descent. The design of a complex-weighted neural networks for multiclass classification is also treated. The simulation results confirm the assumption that the combination of complex version of ReLU-like activation functions and Adam optimizer can considerably speed up the training of complex-valued neural networks.**

Keywords—complex-valued neuron, complex-valued neural network, learning, backpropagation

I. INTRODUCTION

In the last two decades, there have been important advances in the area of the design of artificial intelligence systems [1, 2]. One of the most promising approach in the field is based on the application of neural-like models, structures, and methods [3]. Neural networks are the effective tools for solving the problems of function approximation [4], prediction of the behavior of dynamic systems [5, 6], classification [7, 8], pattern recognition [8], Data Mining [9], optimal control, and so on.

At present, many different architectures of real-weighted artificial neural networks have been developed in data science and machine learning [10, 11]. But the numerous limitations of the real-weighted neural networks [12] renew the interest in the study of generalized concepts and models of neural units and networks. One of the possible step in this direction is to extent the area of the application of neural computing to the systems with complex-valued data.

The complex-valued neural devices deal with information in complex domain using both complex data, as well as complex parameters. The notion of a complex neuron was introduced by N. Aizenberg [12, 13]. The different kinds of discrete complex activation functions were studied in [14]. The complex-valued neural networks with continuous activation function were introduced in early 90's in [15], [16]. Several authors rediscovered the backpropagation learning algorithm for such networks [15–19]. The reasons for the difficulty in finding a nonlinear complex activation function in networks design are given in [19].

The further progress in research in the area of complex-valued neural network is stated in [13]. Numerous important applications of complex-valued neural-like systems in fields such as adaptive designing of patch antennas, neurophysiological analysis, bioinformatics and related image processing, associative memory, time-sequential processing, communications, traffic-dependent optimal control, and complex-valued signal processing are described in [12, 13, 19, 20]. Despite the significant advances in the field, the questions concerning the application of modern learning techniques such as rectifiers [21], spectrum of the function [22], and Adam optimizer [23] are still open for complex-valued networks.

The paper has the following structure. First, we give the notion of the complex-valued neuron and consider the architecture of feed-forward complex-valued neural networks. Next, we describe the complex modification of the backpropagation learning algorithm and propose two new kinds of activation functions for the hidden units and one more function, which can be useful for a neural network classifier. Then, we consider the use of modern optimizers improving the quality of learning. Finally, we give the results of simulation and discuss them.

II. COMPLEX NEURAL NETWORKS

A. Complex-valued neural unit

The complex-valued neural unit (complex neuron) is the functional unit with n inputs $z_1,\dots,z_n$ and one output *out*, which is calculated in the following way:

$$out = f\left(\sum_{j=1}^{n} w_j z_j + w_0\right),$$

where complex numbers $z_1,\dots,z_n$ are input signals, w_0, $w_1,\dots,w_n$ are complex weight coefficients (similarly to [4] the last term w_0 can be seen as the bias of our neuron), $f:\mathrm{C}\to\mathrm{C}$ is a nonlinear complex activation function, which has partial derivatives with respect to both the real and imaginary parts of its argument.

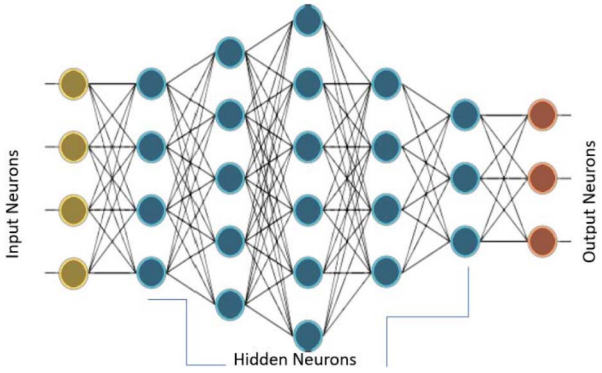


Fig. 1. Fully connected feed-forward complex neural network

B. Feed-forward complex neural network

There many different ways to organize complex neurons in a neural network. We confine ourselves to studying only multilayer fully connected feed-forward complex-valued neural networks (see Fig. 1) satisfying the following condition: each neuron of the network is connected to every neuron in the next layer. The first layer is called the input layer, internal layers are called hidden ones, and the last Lth layer is the output layer. The proceeding of neural network can be described by the following equation:

$$out_{kl} = f\left(\sum_{j=1}^{n_{l-1}} w_{jkl} z_{jkl} + w_{0kl}\right),\ z_{kj,l+1} = out_{kl},$$

where $z_{jkl} = x_{jkl} + i\,y_{jkl}$ is the jth input signal of kth neuron in the lth layer, n_{l-1} is the number of its inputs, i.e. the number of neurons in the previous $(l-1)$-th layer, $w_{jkl} = u_{jkl} + i\,v_{jkl}$ is the value of the respective weight coefficient of the neuron, $l \in \{2,\ldots,L\}$, $k = 1,\ldots,n_l$. We say that our L-layer network has a $n_1 - n_2 - \ldots - n_L$ topology. Henceforth, we shall consider only networks with at least one hidden layer, so $L \geq 3$.

III. LEARNING ALGORITHM

A. Backpropagation learning rule

We consider supervised learning consisting of presenting input patterns and changing the network parameters to bring the actual outputs closer to the desired teaching or target values [4].

Let the collection of pairs $\{(\mathbf{z}^1,\mathbf{t}^1),\ldots,(\mathbf{z}^m,\mathbf{t}^m)\}$ be our training sample, where m is a sample size, $\mathbf{z}^r \in \mathrm{C}^{n_1}$ are training patterns, $\mathbf{t}^r = (t_1^r,\ldots,t_{n_L}^r) \in \mathrm{C}^{n_L}$ are training targets, $(r = 1,\ldots,m)$. We start with an untrained neural network, present patterns $\mathbf{z}^r$ to the input layer, pass the signals through the network and determine the corresponding output vectors $\mathbf{out}^r = (out_1^r,\ldots,out_{n_L}^r)$. Let us consider the notion of the training error E on our training sample:

$$E(\mathbf{w}) = \frac{1}{2}\sum_{r=1}^{m}\left\|\mathbf{out}^r - \mathbf{t}^r\right\|^2 = \frac{1}{2}\sum_{r=1}^{m}\sum_{k=1}^{n_L}\left|out_k^r - t_k^r\right|^2, \quad (1)$$

where the vector $\mathbf{w}$ represents all weights in the networks, $|z| = z\cdot\bar{z}$ denotes the norm of the complex number z, $\|\mathbf{z}\| = \sqrt{(\mathbf{z},\bar{\mathbf{z}})}$ is the norm of the complex vector $\mathbf{z}$. Error function (1) is the scalar real-valued function of the weights and is minimized when the network outputs $\mathbf{out}^r$ match the desired outputs $\mathbf{t}^r$ $(r = 1,\ldots,m)$. During the learning the weights are adjusted to reduce the measure of error (1).

It is evident that $E(\mathbf{w}) = E(\mathbf{u},\mathbf{v})$, where $\mathbf{u}$ is the vector of the real parts of all coefficients of our neural network, $\mathbf{v}$ is the vector of the imaginary parts of all coefficients of the network. Assuming that the activation function f is differentiable, we can apply the complex modification of the backpropagation learning rule based on gradient descent [4]. The weights of the network are initialized with random values, and then they are changed in a direction that will reduce the network error (1):

$$\Delta\mathbf{u}^p = -\eta_p \operatorname{grad} E\left(\mathbf{u}^{p-1}\right),\ \mathbf{u}^p = \mathbf{u}^{p-1} + \Delta\mathbf{u}^p, \quad (2)$$

$$\Delta\mathbf{v}^p = -\eta_p \operatorname{grad} E\left(\mathbf{v}^{p-1}\right),\ \mathbf{v}^p = \mathbf{v}^{p-1} + \Delta\mathbf{v}^p, \quad (3)$$

where $p = 1,2,\ldots$ is the iteration step and the positive factor η_p is the learning rate indicating the relative size of the change in weights, $\mathbf{u}^0$ and $\mathbf{v}^0$ are some initial weight vectors.

It should be mentioned that we use only real values for the learning rate. The question concerning the use of arbitrary complex learning rates in the case of a single complex neuron was studied in [14].

Let $s_{kl}^r = \sum_{j=1}^{n_{l-1}} w_{jkl} z_{jkl}^r + w_{0kl}$, $a_{kl}^r = \operatorname{Re} s_{kl}^r$, $b_{kl}^r = \operatorname{Im} s_{kl}^r$, $f(z) = g(x,y) + ih(x,y)$, where $z = x + iy$. Now we can give the equations for the components of gradient in (2), (3). Because the error is not explicitly dependent upon w_{jkl}, we must use the chain rule for differentiation:

$$\frac{\partial E}{\partial u_{jkl}} = \sum_{r=1}^{m}\left(\frac{\partial E}{\partial g_{kl}^r}\left(\frac{\partial g_{kl}^r}{\partial a_{kl}^r}\frac{\partial a_{kl}^r}{\partial u_{jkl}} + \frac{\partial g_{kl}^r}{\partial b_{kl}^r}\frac{\partial b_{kl}^r}{\partial u_{jkl}}\right) + \frac{\partial E}{\partial h_{kl}^r}\left(\frac{\partial h_{kl}^r}{\partial a_{kl}^r}\frac{\partial a_{kl}^r}{\partial u_{jkl}} + \frac{\partial h_{kl}^r}{\partial b_{kl}^r}\frac{\partial b_{kl}^r}{\partial u_{jkl}}\right)\right), \quad (4)$$

$$\frac{\partial E}{\partial v_{jkl}} = \sum_{r=1}^{m}\left(\frac{\partial E}{\partial g_{kl}^r}\left(\frac{\partial g_{kl}^r}{\partial a_{kl}^r}\frac{\partial a_{kl}^r}{\partial v_{jkl}} + \frac{\partial g_{kl}^r}{\partial b_{kl}^r}\frac{\partial b_{kl}^r}{\partial v_{jkl}}\right) + \frac{\partial E}{\partial h_{kl}^r}\left(\frac{\partial h_{kl}^r}{\partial a_{kl}^r}\frac{\partial a_{kl}^r}{\partial v_{jkl}} + \frac{\partial h_{kl}^r}{\partial b_{kl}^r}\frac{\partial b_{kl}^r}{\partial v_{jkl}}\right)\right). \quad (5)$$

Let us describe equations for partial derivatives in (4) and (5) (we shall omit the index r for the sake of brevity):

$$\frac{\partial a_{kl}}{\partial u_{jkl}} = x_{jkl},\ \frac{\partial a_{kl}}{\partial u_{0kl}} = 1,\ \frac{\partial b_{kl}}{\partial u_{jkl}} = y_{jkl},\ \frac{\partial b_{kl}}{\partial u_{0kl}} = 0, \quad (6)$$

$$\frac{\partial a_{kl}}{\partial v_{jkl}} = -y_{jkl},\ \frac{\partial a_{kl}}{\partial v_{0kl}} = 0,\ \frac{\partial b_{kl}}{\partial v_{jkl}} = x_{jkl},\ \frac{\partial b_{kl}}{\partial v_{0kl}} = 1. \quad (7)$$

Equations (4)–(7) are suitable for all neurons in the network. But the formulas for derivatives $\frac{\partial \mathrm{E}}{\partial g_{kl}}$ and $\frac{\partial \mathrm{E}}{\partial h_{kl}}$ depend on the index of the layer *l*. For the output layer we have

$$\frac{\partial \mathrm{E}}{\partial g_{kL}} = g_{kL} - \mathrm{Re}\, t_k,\ \frac{\partial \mathrm{E}}{\partial h_{kL}} = h_{kL} - \mathrm{Im}\, t_k. \quad (8)$$

Therefore, equations (4)–(8) allow us to compute weight corrections for the output layer.

For the other neurons we cannot use (8), because there is no specified desired response for them. The error signal for a hidden neuron can be determined recursively in terms of the error signals of all the neurons to which this hidden neuron is directly connected [4], i.e. during the training an error must be propagated from the output layer back to the hidden ones [13].

Suppose that for the *l*th layer $(l > 2)$ the all necessary in (4)–(7) partials derivatives are already calculated. Let us show how we can calculate the weight corrections for the *k*th neuron in the previous $(l-1)$-th layer. We can write for the *l*th layer:

$$\frac{\partial E}{\partial x_{jkl}} = \sum_{r=1}^{m} \frac{\partial E}{\partial g_{kl}^r}\left(\frac{\partial g_{kl}^r}{\partial a_{kl}^r}\frac{\partial a_{kl}^r}{\partial x_{jkl}} + \frac{\partial g_{kl}^r}{\partial b_{kl}^r}\frac{\partial b_{kl}^r}{\partial x_{jkl}}\right) + \sum_{r=1}^{m} \frac{\partial E}{\partial h_{kl}^r}\left(\frac{\partial h_{kl}^r}{\partial a_{kl}^r}\frac{\partial a_{kl}^r}{\partial x_{jkl}} + \frac{\partial h_{kl}^r}{\partial b_{kl}^r}\frac{\partial b_{kl}^r}{\partial x_{jkl}}\right), \quad (9)$$

$$\frac{\partial E}{\partial y_{jkl}} = \sum_{r=1}^{m} \frac{\partial E}{\partial g_{kl}^r}\left(\frac{\partial g_{kl}^r}{\partial a_{kl}^r}\frac{\partial a_{kl}^r}{\partial y_{jkl}} + \frac{\partial g_{kl}^r}{\partial b_{kl}^r}\frac{\partial b_{kl}^r}{\partial y_{jkl}}\right) + \sum_{r=1}^{m} \frac{\partial E}{\partial h_{kl}^r}\left(\frac{\partial h_{kl}^r}{\partial a_{kl}^r}\frac{\partial a_{kl}^r}{\partial y_{jkl}} + \frac{\partial h_{kl}^r}{\partial b_{kl}^r}\frac{\partial b_{kl}^r}{\partial y_{jkl}}\right). \quad (10)$$

In the last two equations the values of partial derivatives $\frac{\partial E}{\partial g_{kl}}, \frac{\partial E}{\partial h_{kl}}$ $\frac{\partial g_{kl}}{\partial a_{kl}}, \frac{\partial g_{kl}}{\partial b_{kl}}, \frac{\partial h_{kl}}{\partial a_{kl}}, \frac{\partial h_{kl}}{\partial b_{kl}}$ are already calculated by assumption. The remaining partial derivatives are given by

$$\frac{\partial a_{kl}}{\partial x_{jkl}} = u_{jkl},\ \frac{\partial b_{kl}}{\partial x_{jkl}} = v_{jkl},\ \frac{\partial a_{kl}}{\partial y_{jkl}} = -v_{jkl},\ \frac{\partial b_{kl}}{\partial y_{jkl}} = u_{jkl}.$$

But the partial derivatives of *E* with respect to the input values x_{jkl} and y_{jkl} for the *l*th layer implicitly coincide with the partial derivatives of the error function (1) with respect to the real and imaginary parts of the output values of the previous layer. Hence,

$$\frac{\partial E}{\partial g_{j,l-1}} = \sum_{k=1}^{n_l} \frac{\partial E}{\partial x_{jkl}},\ \frac{\partial E}{\partial h_{j,l-1}} = \sum_{k=1}^{n_l} \frac{\partial E}{\partial y_{jkl}}. \quad (11)$$

Equation (9)–(11) must be used instead of (8) for the hidden layers. It provides the transition from the calculation of the gradient components corresponding to the neurons of the current layer to the ones of the previous layer (the method of the fast backward gradient computation).

The presented learning algorithm based on the corrections of weight coefficients according to equations (2)–(11) is the complex modification of the well-known backpropagation learning algorithm [4].

B. On-line version of learning

We have just described the off-line or batch version of the network learning with the error function (1). But in the case of large training samples it needs the expenses of considerable computational resources and may have unacceptable performance. It is known that sometimes one can speed up the network training significantly by the use of small batches of teaching patterns instead of the entire sample [4].

In the case where batch size is 1, we come to the on-line learning protocol. A teacher provides input patterns in random order and compute the gradient of the error function (1) with respect to the only current training pair $(\mathbf{z}^r, \mathbf{t}^r)$. In this case we can simplify equations (4) and (5):

$$\frac{\partial E}{\partial u_{jkl}} = \frac{\partial E}{\partial g_{kl}}\left(\frac{\partial g_{kl}}{\partial a_{kl}}\frac{\partial a_{kl}}{\partial u_{jkl}} + \frac{\partial g_{kl}}{\partial b_{kl}}\frac{\partial b_{kl}}{\partial u_{jkl}}\right) + \frac{\partial E}{\partial h_{kl}}\left(\frac{\partial h_{kl}}{\partial a_{kl}}\frac{\partial a_{kl}}{\partial u_{jkl}} + \frac{\partial h_{kl}}{\partial b_{kl}}\frac{\partial b_{kl}}{\partial u_{jkl}}\right),$$

$$\frac{\partial E}{\partial v_{jkl}} = \frac{\partial E}{\partial g_{kl}}\left(\frac{\partial g_{kl}}{\partial a_{kl}}\frac{\partial a_{kl}}{\partial v_{jkl}} + \frac{\partial g_{kl}}{\partial b_{kl}}\frac{\partial b_{kl}}{\partial v_{jkl}}\right) + \frac{\partial E}{\partial h_{kl}}\left(\frac{\partial h_{kl}}{\partial a_{kl}}\frac{\partial a_{kl}}{\partial v_{jkl}} + \frac{\partial h_{kl}}{\partial b_{kl}}\frac{\partial b_{kl}}{\partial v_{jkl}}\right).$$

C. Choice of the activation function

Let us pass to the problem of the choice of the activation function. Unfortunately, the complex version of the popular real-valued logistic sigmoid or hyperbolic tangent are discontinuous as the functions of complex variable. Therefore, they cannot be applied in learning algorithms for complex-valued networks which use the gradient vector [19]. The traditional approach to process complex signals by neurons is to use 'split' complex activation by uncoupled real sigmoidal functions for the real and the imaginary components of the weighted sum [13, 17]. But it leads to the complex-valued network which seems likely to perform as the pair of real-valued networks.

In order to obtain the satisfactory performance without the use of splitted activation functions we will use the complex version of the rational sigmoid

$$f(z) = \frac{z}{|z| + q}, \quad (12)$$

where *q* is a real positive parameter.

For the complex rational sigmoid (12) we have

$$g(x,y)=\frac{x}{\sqrt{x^2+y^2}+q},\quad h(x,y)=\frac{y}{\sqrt{x^2+y^2}+q}.$$

It is necessary to notice that the values of the function (12) lay in the unit disk centered on the origin. In addition, the complex rational sigmoid squashes proportionally the real and imaginary parts of its argument, enhances the "weak" input signals, and depresses the "strong" ones. The parameter q in (12) can be useful in fitting the shape of function near the origin.

It is easy to obtain the following expressions for partial derivatives of the function (12) in the case $s_{kl}\neq 0$:

$$\frac{\partial g_{kl}}{\partial a_{kl}}=\frac{b_{kl}^2+q\,|s_{kl}|}{\left(|s_{kl}|+q\right)^3},\ \frac{\partial g_{kl}}{\partial b_{kl}}=-\frac{a_{kl}b_{kl}}{|s_{kl}|\left(|s_{kl}|+q\right)^2},\quad (13)$$

$$\frac{\partial h_{kl}}{\partial a_{kl}}=-\frac{a_{kl}b_{kl}}{|s_{kl}|\left(|s_{kl}|+q\right)^2},\ \frac{\partial h_{kl}}{\partial b_{kl}}=\frac{a_{kl}^2+q\,|s_{kl}|}{\left(|s_{kl}|+q\right)^3}.\quad (14)$$

Equations (13), (14) is not true in the origin, where

$$\frac{\partial g_{kl}}{\partial a_{kl}}=\frac{\partial h_{kl}}{\partial b_{kl}}=\frac{1}{q},\ \frac{\partial g_{kl}}{\partial b_{kl}}=\frac{\partial h_{kl}}{\partial a_{kl}}=0.\quad (15)$$

Equations (2)–(15) describe in detail one step of the learning in the batch mode.

Nowadays, a very simple activation function $\mathrm{ReLU}(x)=\max\{0,x\}$ became popular in the real-valued network learning, because it provides very good experimental results and often overperform the classical sigmoidal functions. Let us consider the following generalization of ReLU to the complex domain:

$$\mathrm{CRelu}(z)=\begin{cases}z, \text{ if } \mathrm{Re}\,z>0 \text{ and } \mathrm{Im}\,z>0,\\ 0, \text{ otherwise.}\end{cases}$$

Both the real and the complex parts of CReLU have no partial derivatives in the origin, but can be easily extended. For definiteness, extend them to zero. We can then write

$$\frac{\partial g_{kl}}{\partial a_{kl}}=\frac{\partial h_{kl}}{\partial b_{kl}}=\begin{cases}1, \text{ if } \mathrm{Re}\,z>0 \text{ and } \mathrm{Im}\,z>0,\\ 0, \text{ otherwise.}\end{cases}\quad (16)$$

$$\frac{\partial h_{kl}}{\partial a_{kl}}=\frac{\partial h_{kl}}{\partial b_{kl}}=0\ .\quad (17)$$

It should be mentioned that there possible many other complex generalizations of ReLU deriving some its features.

If the complex-valued neural network is designed to perform as multiclass classifier, then we must modify the output layer activation function in the appropriate way. It can be attained by applying the modified softmax function to the weighted sums yielded in the last layer. Let us define the network outputs in the following way:

$$out_{kL}=\frac{\exp\left(|s_{kL}|\right)}{\sum_{j=1}^{n_L}\exp\left(|s_{jL}|\right)},\ k=1,\ldots,n_L$$

where n_L is the number of classes. The network error (1) must also be changed. For example, we can use the cross entropy

$$E(\mathbf{w})=-\frac{1}{m}\sum_{r=1}^{m}\sum_{k=1}^{n_L}t_k^r\log\left(out_{kL}^r\right)$$

and modify (8) and $\frac{\partial g_{kl}}{\partial a_{kl}}$, $\frac{\partial g_{kl}}{\partial b_{kl}}$, $\frac{\partial h_{kl}}{\partial a_{kl}}$, $\frac{\partial h_{kl}}{\partial b_{kl}}$ appropriately.

D. Modifications of the learning rule

It is well-known that the learning algorithm based on the gradient descent can be trapped in the plateaus on the error surfaces (regions in which the $\|\mathrm{grad}\,E\|$ is very small). The utilization of momentum optimization can increase the rate of learning. Let us modify (2), (3) as follows:

$$\Delta\mathbf{w}^p=\beta\Delta\mathbf{w}^{p-1}-\eta_p\,\mathrm{grad}\,E\left(\mathbf{w}^{p-1}\right),\ \mathbf{w}^p=\mathbf{w}^{p-1}+\Delta\mathbf{w}^p,$$

where $\beta\in[0,1)$ is the momentum constant, $\mathrm{grad}\,E(\mathbf{w})=\mathrm{grad}\,E(\mathbf{u})+i\,\mathrm{grad}\,E(\mathbf{v})$.

An Adam optimizer [23] combines the idea of momentum optimizer and RMSProp. According to Adam algorithm we change rules (2), (3) by the following one:

$$\mathbf{w}^p=\mathbf{w}^{p-1}-\frac{\eta_p\left(1-\beta_2^p\right)}{1-\beta_1^p}\frac{\mathbf{m}^p}{\sqrt{\mathbf{s}^p+\varepsilon}},\quad (18)$$

where $\mathbf{m}^0=\mathbf{s}^0=\mathbf{0}$, $\beta_1\in(0,1)$, $\beta_2\in(0,1)$, $\varepsilon>0$, β_1^p and β_2^p denote β_1 and β_2 to the power p,

$$\mathbf{m}^p=\beta_1\mathbf{m}^{p-1}+(1-\beta_1)\,\mathrm{grad}\,E\left(\mathbf{w}^{p-1}\right),$$

$$\mathbf{s}^p=\beta_2\mathbf{s}^{p-1}+(1-\beta_2)\,\mathrm{grad}\,E\left(\mathbf{w}^{p-1}\right)\otimes\mathrm{grad}\,E\left(\mathbf{w}^{p-1}\right),$$

the $\otimes$ symbol represents the element-wise multiplication, the sum $\mathbf{s}^p+\varepsilon$ denotes the element-wise sum (ε is added to all components of the vector $\mathbf{s}^p$), and the vector $\frac{\mathbf{m}^p}{\sqrt{\mathbf{s}^r+\varepsilon}}$ is the element-wise quotient of the numerator and the denominator.

IV. SIMULATION

We used the complex backpropagation algorithm for solving the very hard XOR problem in the case of 50 features. The multiplicative representation of XOR [24] was considered in $\{-1,1\}$ base. The full domain of definition is too large, so we used randomly generated samples containing 10,000 training pairs (8,000 instances for training and 2,000 patterns for testing). The validation set was not used.

We compared the performance of the real-valued and complex-valued four-layer networks consisting of neurons

with sigmoidal and ReLU-like activation functions. We applied the complexification of real inputs in complex cases, i.e. networks had 25 inputs, to which of them corresponded the coupled pair of two real features. We used 50-20-10-1 topology in the real case and 25-20-10-1 topology in the complex case. Each network was trained in on-line mode on 100 different learning samples. The number of epochs was equal to 1000. We did not employ scheduling of the learning rate and η_p was set to 0.1. We used hyperbolic tangent as a real sigmoid, applied (12)–(15) in the case of complex rational sigmoid (with $q=1$), and (16), (17) for CReLU. Every sigmoid was scaled by a factor $\alpha=1.716$ according to [4].

TABLE I. SIMULATION RESULTS

Activation function	Average accuracy (in %)			
	Training set SGD	*Test set SGD*	*Training set Adam*	*Test set Adam*
Real sigmoid	68.15	53.24	74.03	62.12
Splitted complex sigmoid	71.92	62.25	82.88	67.03
Complex rational sigmoid	73.01	62.19	85.35	67.76
CReLU	74.96	67.05	92.43	81.51

The parameter β_1 and β_2 in (18) were initialized to 0.9 and 0.999, respectively [23]. We employed weight initialization techniques from [21]. Table I presents average performance measures in the case of learning using stochastic gradient descent (SGD) and Adam optimizers, respectively.

V. CONCLUSIONS

Feed-forward complex-valued neural networks give a simple and powerful architecture providing an adequate approximation of nonlinear mappings. The new activation functions and the complex modifications of effective modern optimizers improve the quality of learning. Using the results of simulation, we can conclude that the combination of CReLU activation and Adam optimizer is preferable under conditions of our experiment.

Nevertheless, the new simulation experiments are necessary for tuning better parameters in the learning framework, and new theoretical research are needed to establish the convergence rate and the training complexity for such networks.

REFERENCES

[1] Y. Bodyanskiy, A. Dolotov, D. Peleshko, Y. Rashkevych, and O. Vynokurova, "Associative probabilistic neuro-fuzzy system for data classification under short training set conditions," in Advances in Intelligent Systems and Computing vol. 761. Heidelberg: Springer, 2019, pp. 56–63.

[2] 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, vol. 689, N. Shakhovska and V. Stepashko, Eds. Cham: Springer, 2018, pp. 538–557.

[3] I. Izonin, R. Tkachenko, N. Kryvinska, K. Zub, O. Mishchuk, and T. Lisovych, "Recovery of incomplete IoT sensed data using high-performance extended-input neural-like structure," Procedia Computer Science, vol. 160, pp. 521–526, 2019.

[4] S. Haykin, Neural Networks and Learning Machines, 3rd ed. Upper Saddle River, NJ: Pearson Education, 2009.

[5] F. Geche, V. Kotsovsky, A. Batyuk, S. Geche, and M. Vashkeba, "Synthesis of time series forecasting scheme based on forecasting models system," in CEUR Workshop Proceedings, vol. 1356, 2015, pp. 121–136.

[6] Y. Bodyanskiy, A. Dolotov, D. Peleshko, Y. Rashkevych, and O. Vynokurova, "Online time series changes detection based on neuro-fuzzy approach," in Predictive Maintenance in Dynamic Systems, E. Lughofer, M. Sayed-Mouchaweh Eds. Cham: Springer, 2019, pp. 131–166.

[7] M. Lupei, A. Mitsa, V. Repariuk, and V. Sharkan, "Identification of authorship of Ukrainian-language texts of journalistic style using neural networks," Eastern-European Journal of Enterprise Technologies, vol. 1, no. 2 (103), pp. 30–36, 2020.

[8] A. Kazarian, V. Teslyuk, I. Tsmots, and M. Tykhan, "Implementation of the face recognition module for the "smart" home using remote server," in Advances in Intelligent Systems and Computing III. AISC 871 (Selected Papers from the International Conference on Computer Science and Information Technologies, CSIT 2018. Lviv, Ukraine, September 11–14, 2018). Springer, 2019, pp. 17–27.

[9] I. Izonin, N. Kryvinska, R. Tkachenko, K. Zub, and P. Vitynskyi, "An extended-input GRNN and its application," Procedia Computer Science, vol. 160, pp. 578–583, 2019.

[10] I. Tsmots, V. Teslyuk, T. Teslyuk, and I. Ihnatyev, "Basic components of neuronetworks with parallel vertical group data real-time processing," in Advances in Intelligent Systems and Computing vol. 689, N. Shakhovska and V. Stepashko, Eds. Cham: Springer, 2018, pp. 558–576.

[11] R. Tkachenko and I. Izonin, "Model and principles for the implementation of neural-like structures based on geometric data transformations," in Advances in Intelligent Systems and Computing, vol. 754, Z. Hu, S. Petoukhov, I. Dychka, and M. He, Eds. Springer, 2019, pp. 578–587.

[12] I. Aizenberg, Complex-Valued Neural Networks with Multi-Valued Neurons. Springer, 2011.

[13] A. Hirose, Complex-Valued Neural Networks, 2nd ed. Springer, 2012.

[14] V. Kotsovsky, F. Geche, and A. Batyuk, "Artificial complex neurons with half-plane-like and angle-like activation function," in Proceedings of the International Conference on Computer Sciences and Information Technologies, CSIT 2015. Lviv, Ukraine, September 14–17, 2015, pp. 57–59.

[15] H. Leung and S. Haykin, "The complex backpropagation algorithm," IEEE Transactions on Signal Processing, vol. 39, no. 2, pp. 2101‑2104, September 1991.

[16] N. Benvenuto and F. Piazza, "On the complex backpropagation algorithm," IEEE Transactions on Signal Processing, vol. 40, no. 4, pp. 967–969, 1992.

[17] G. M. Georgiou and C. Koutsougeras, "Complex domain backpropagation," IEEE Transactions on Circuits and Systems II, vol. 39, no. 5, pp. 330–334, 1992.

[18] V. Kotsovsky, "Feed-forward neural networks with complex weights," Herald of Taras Shevchenko National University of Kyiv, series Physics and Mathematics, no. 3, pp. 147–150, 2007.

[19] T. Kim and T. Adali, "Fully complex multi-layer perceptron network for nonlinear signal processing," Journal of VLSI Signal Processing 32, pp. 29–43, 2002.

[20] T. Adalı, P. Schreier, and 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.

[21] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on ImageNet classification," in Proceeding 2015 IEEE International Conference on Computer Vision, ICCV'15. Santiago, Chile, December 13–16, 2015, pp. 1026–1034.

[22] V. Kotsovsky, F. Geche, and A. Batyuk, "Finite generalization of the offline spectral learning," in Proceedings of the 2018 IEEE Second International Conference on Data Stream Mining & Processing, DSMP 2018. Lviv, Ukraine, August 21–25, 2018, pp. 356–360.

[23] D. Kingma and J. Ba, "Adam: a method for stochastic optimization," in Conference Track Proceedings of 3rd International Conference on Learning Representations, ICLR 2015. San Diego, CA, USA, May 7–9, 2015.

[24] V. Kotsovsky, F. Geche, and A. Batyuk, "On the computational complexity of learning bithreshold neural units and networks," in Advances in Intelligent Systems and Computing, vol 1020, V. Lytvynenko et al., Eds. Heidelberg: Springer, 2020, pp. 189–202.

IEEE Third International Conference on Data Stream Mining & Processing
August 21-25, 2020, Lviv, Ukraine

Trainable Neural Networks Modelling for a Forecasting of Start-Up Product Development

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

Olga Mezentseva
Technology Management Department
Faculty of Information Technologies
Taras Shevchenko National University
of Kyiv
Kyiv, Ukraine
olga.mezentseva.fit@gmail.com

Maksym Proskurin
Technology Management Department
Faculty of Information Technologies
Taras Shevchenko National University
of Kyiv
Kyiv, Ukraine
proskuryn69@gmail.com

***Abstract* — In this article, components of the model of forecasting of IT products development of innovative start-up projects are considered based on the analysis of formed datasets of the interactions of prospective clients. We offered the algorithm of formation of initial datasets based on Customer Journey Map (CJM), which are the tool of fixing of events of the interaction of clients with the system. We propose to use the trainable neural network as a mechanism for processing big data sets and building IT product development strategies. We used a simple linear regression analysis to model the relationship between a single explanatory variable and a continuous response variable (dependent variable). An exploratory data analysis method was applied to the available data to find repetitive patterns and anomalies. In the course of the research, we constructed a model of linear regression implementation using the gradient optimisation approach. The linear models of *the scikit-learn* library for the regression task were also applied, and the stabilisation regression method was implemented. We have modelled and analysed the obtained results.**

Keywords — start-up, information interaction, customer journey map, forecasting, neural networks

I. INTRODUCTION

At the current stage of development of information systems engineering technologies, innovations in the form of start-up projects are becoming increasingly important [1]. Such projects are often formalised as separate commercial enterprises to obtain funding and financial profitability. Often, such companies are represented with the implementation of SaaS (Software as a Service) distribution model [2] or B2B (Business to Business) transactions [3]. Such business models have problems with long sales cycles, as the decision of the client is collective and depends on many factors. The success of the product development and, accordingly, the efficiency of the enterprise depends on the indicator of the quality of customer service at information interaction with the IT system. However, to define directions of development of such products, it is not simple. To do this, it is necessary to form data sets, which characterize the points of preferable information interaction of different kinds of clients, to analyse and predict the directions of such development.

The method of information interaction can be used to solve these problems. It proved to be effective in many projects. It is based on the analysis of the "journey" (interaction with IT product) of the prospective client (Customer Journey Map - CJM) [4, 5]. Also, it is necessary to combine millions of events, which will provide the necessary analysis of the impact on the customers' journeys and will determine the future content of projects to develop such IT products.

By analysing [6] millions of real-time interaction data points, it will be possible to find the most critical customer events in the system [7] and prioritise those opportunities. This can have a significant impact on business goals, such as increasing revenue, reducing customer churn, improving customer service [8] and developing innovations in the company.

However, constructive and technological difficulties of construction of development programs of complex IT products demand use of the project approach [9 - 12]. In such case, methods and information technologies of project management are applied in processes of creation and development of such systems.

Experience shows that for analytical processing of considerable volumes of the information of interaction of clients with IT system use of methods of artificial intelligence (AI) will have a significant influence on the efficiency of development programs (of projects) creation.

The search for optimal variants of distribution of resources at the preparation of development programs in innovative projects can reduce terms of performance of project tasks and, as a result, decrease their cost. At the same time, predicting the impact of interaction with clients [13] on the variants of development programs in such projects is a multidimensional task that can be solved using technologies of modern artificial neural networks [14]. Besides, it is necessary to take into account numerous changes [15] that affect various parameters of innovative projects and significantly affect the negative results of their implementation.

Thus, consideration of the possibilities of experimental use of the Customer Journey Map (CJM) and Artificial Neural

Networks (ANN) in the research of improving the quality of interaction of numerous clients based on optimal development programs of start-up projects is an actual challenge.

II. Analysis of recent research and publications

Application of the project approach to start-up development and effective use of products of these projects at the creation of modern IT were considered in research papers [9 - 16]. In [9], new models of management of innovations in projects based on the use of Markov processes are offered. This approach can be used for further development of start-up projects. However, there is no assessment of the value of products for customers of such projects. The [10] contains a description of the competency models that can be applied to Start-up project teams by the customer. However, there are no methods of interaction between users and the customer's team here.

The [11] provides a detailed description of the process approach for managing any kind of projects, and the specifics of innovation project management are considered separately. However, the processes of management of the creation and development of a product are not given. At the same time, [12] describes integrated methodologies for organisational and product management of development projects. However, methods of an estimation of the efficiency of communication interaction are insufficiently described. In [13, 14, 16] approaches to the management of complex IT projects based on proactive (anticipatory, predictive) management with the use of methods of artificial intelligence, in particular, with the use of artificial neural networks, have been proposed. However, studies based on interaction with clients have not been conducted.

The use of methods to assess customer interactions in SaaS and B2B business models based on Customer Journey Map has been highlighted in [4, 17 - 19]. As noted in these papers, such developments should be based on new tools to assess the productivity (value) of customer interaction with the IT product as feedback. Here we can use well-proven customer journey models though it is not specified, how to process rather big data on numerous indicators analytically.

Then, considering aspects of application of intellectual tools for the analysis and forecasting based on the processing of a big data set from the interaction of clients, it is possible to consider papers [20 - 27]. In these papers, it is noted that the most acceptable for us, in terms of intellectual forecasting of the composition of programs for the development of complex IT products, is the use of trainable neural networks. Furthermore, it will be useful in our case to build a model of linear regression implementation [23] using the gradient optimisation approach [27].

The goal of the article is to justify and develop a model for predicting the composition of the program for the development of complex products of start-up projects using the analysis of customer interaction data based on the Customer Journey Map and the application of trainable neural networks.

The main objectives of this study are as follows:

1) Identify the interaction elements that form the basis of the Customer Journey Map to create survey datasets within one year.

2. Development of a model for forecasting the composition of development programs, taking into account maximum customer loyalty and retention in their interaction with the IT system.

3. Development of algorithms for modelling the neural network training processes and setting up the forecasting processes.

III. Model development and use of modelling method

A. Constructing customer interaction models with IT product based on journey maps

As mentioned above, one of the effective tools to assess the quality of user interaction with the IT product is the customer journey map (CJM). It is a tool for visualising of the interaction of the consumer with a product or service. Creating a CJM is both a process of analysis and a method for generating ideas to improve a product or service.

CJM displays a time-bound interaction broken down into small components. The components of interactions refer both to the process (consumers' goals and objectives, their actions, expected results, problems and barriers preventing the transition to the next step, touchpoints, materials, tools, equipment, KPI from the business point of view, etc.) and to the psycho-emotional state of the consumer (thoughts, feelings, emotions).

This approach to the development of products characterized by multi-channel interaction is particularly useful, that is, for those cases where the customer and the product have many "touchpoints". The customer always has a sum of impressions of the interaction with the product through all available channels, so that even one negative experience can vilify the entire product in the eyes of the client and force him to cancel the subscription.

As an example, we can consider the CJM depicted in Fig. 1, which shows the customer's journey from awareness, consideration, purchase and further use of the “instant server” in a particular web application (IaaS distribution model) of the multinational telecommunications company - "Telefónica" (Fig. 1).

B. Building predictive models using neural networks

Regression models are used to predict target variables on a continuous scale, which makes them useful for many scientific issues and information industry applications, such as understanding relationships between variables, assessing trends, or making forecasts.

One example of their application can be the prediction of company sales in the coming months:

$$y = w_0 + w_1 x, \quad (1)$$

where w_0 – weight, that is the intersection point of the Y-axis;

w_1 - the explanatory variable coefficient.

The most frequently used is multiple linear regression ***y:***

$$y = w_0 x_o + w_1 x_1 + \cdots + w_m x_m = \sum_{i=1}^{m} w_i x_i = w^T x \,, \quad (2)$$

where w_0 is the intersection point of the Y-axis at $x_0 = 1$.

m – the number of regression members;
w^T - the explanatory variable coefficient.

To determine the number of linear relationships between features, we will now create a correlation matrix.

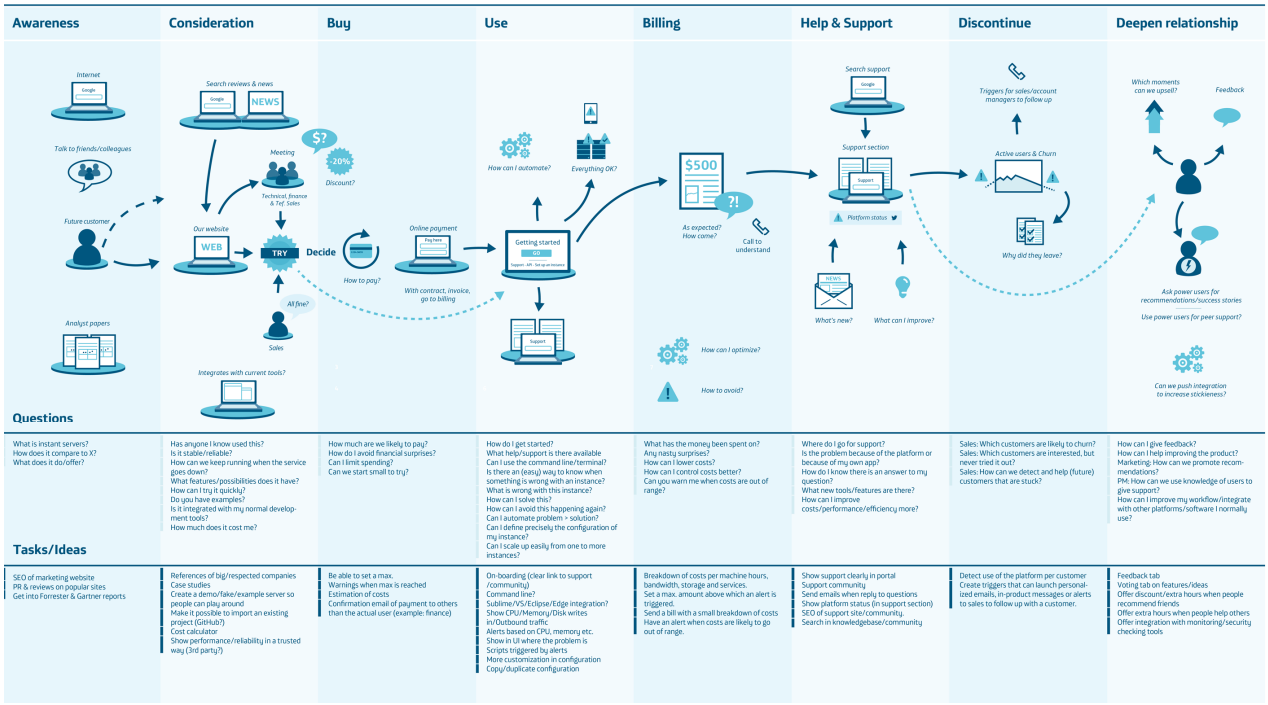


Fig. 1. Example of Customer Journey Map.

The correlation coefficients are limited to the range [-1, 1]. Two attributes have, respectively, absolutely positive correlation if r = 1, no correlation if r = 0, and absolutely negative correlation if r = -1. As mentioned earlier, Pearson's correlation coefficient can be calculated simply as the covariance between two attributes **x** and **y** - numerator, divided by the product of their standard deviations (denominator):

$$r = \frac{\sum_{i=1}^{n}\left[(x^i - \mu_i)(y^i - \mu_y)\right]}{\sqrt{\sum_{i=1}^{n}(x^i - \mu_x)^2}\sqrt{\sum_{i=1}^{n}(y^i - \mu_y)^2}} = \frac{\sigma_{xy}}{\sigma_x \sigma_y} \,, \quad (3)$$

where n – the number of attributes;
μ - the empirical average of these attributes;
σ – covariance between attributes.

As can be seen in the resulting figure, the correlation matrix provides us with another final diagram, which can help us select attributes based on their corresponding linear correlations (Fig. 2).

To fit a linear regression model, we are interested in those features that have a high correlation with our target variable MEDV. Looking at the above correlation matrix, we can see that our target variable MEDV shows the highest correlation with variable LSTAT (-0.74). The correlation between RM and MEDV is also relatively high (0.70). In the presence of a linear relationship between these two variables that we have observed in the dot matrix as an explanatory variable, the artificial neuron uses a linear activation function. Moreover, we have defined the JS(w) cost function that we have minimised for weight extraction thanks to optimisation algorithms such as gradient descent (GD) [23] and stochastic gradient descent (SGD) [24].

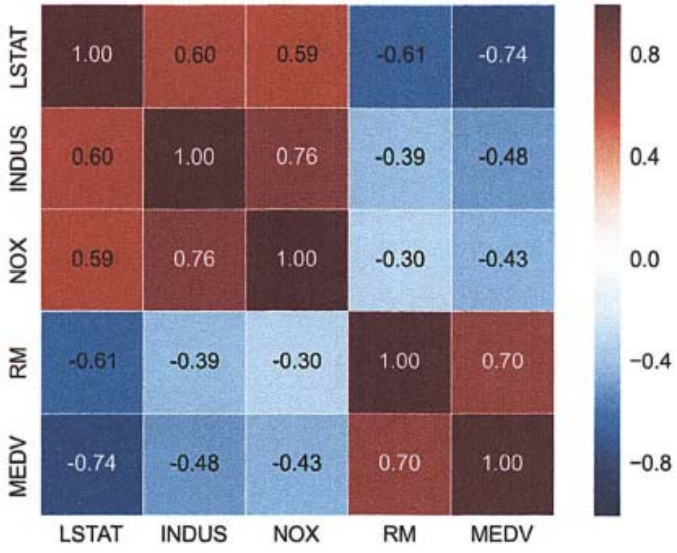


Fig. 2. An example of a correlation matrix.

This cost function in ADALINE is the sum of squared errors (SSE). It is identical to the least squared value function (MLS), which we defined as follows:

$$JS(w) = \frac{1}{2}\sum_{i=1}^{n}(y^i - \hat{y}^i)^2 \quad , \qquad (4)$$

where $\hat{y} -$ it is a predicted value $\hat{y} = w^T x$; (note that the coefficient 1/2 is used simply for the convenience of obtaining an update rule for gradient descent).

IV. EXPERIMENTATION

As our target variable, which we want to predict using one or more of these 51 explanatory variables, we will consider the tariffs for interaction with the system (a product of the start-up project) called MEDV.

Before we continue the exploration of this data set, we will put it into *DataFrame* data table of the *pandas* library [27] from the UCI repository (Fig. 3).

```
import pandas as pd

df = pd.read_csv(url, header=None, sep='\s+')
df.columns = ['CRIM', 'ZN', 'INDUS', 'CHAS',
              'NOX', 'RM', 'AGE', 'DIS', 'RAD',
              'TAX', 'PTRATIO', 'B', 'LSTAT', 'MEDV']
df.head()
```

Fig. 3. Downloading the source data.

During the training of a linear regression model, it is not required that explanatory or target variables are distributed normally.

First, let us create a matrix of dotted charts that allows visualising pairwise correlations between different attributes in one place in this dataset. To prepare the matrix of point graphs, let's use the *pairplot* function from Python *seaborn library* [30], which is developed based on the *matplotlib* library for building statistical charts (Fig. 4).

```
import matplotlib.pyplot as plt
import seaborn as sns
sns.set(style='whitegrid', context='notebook')
cols = ['LSTAT', 'INDUS', 'NOX', 'RM', 'MEDV']
sns.pairplot(df[cols], size=2.5)
plt.show()
```

Fig. 4. An example of a matrix of correlation links and point graphs

Using the *corrcoef* function of the NumPy library on five columns of attributes and the heatmap function of the seaborn library, we will get the necessary relationships between the initial indicators

The following code is used to prepare an array of correlation matrices (Fig. 5).

```
import numpy as np
cm = np.corrcoef(df[cols].values.T)
sns.set(font_scale=1.5)
hm = sns.heatmap(cm,
                 cbar=True,
                 annot=True,
                 square=True,
                 fmt='.2f',
                 annot_kws={'size': 15},
                 yticklabels=cols,
                 xticklabels=cols)
plt.show()
```

Fig. 5. Example of preparing an array of correlation matrices in the form of a heatmap.

In essence, linear regression on MLS can be understood as ADALINE without a single step function, resulting in continuous target values instead of class labels 1 and 1 (Fig. 6).

```
class LinearRegressionGD(object):

    def __init__(self, eta=0.001, n_iter=20):
        self.eta = eta
        self.n_iter = n_iter

    def fit(self, X, y):
        self.w_ = np.zeros(1 + X.shape[1])
        self.cost_ = []

        for i in range(self.n_iter):
            output = self.net_input(X)
            errors = (y - output)
            self.w_[1:] += self.eta * X.T.dot(errors)
            self.w_[0] += self.eta * errors.sum()
            cost = (errors**2).sum() / 2.0
            self.cost_.append(cost)
        return self

    def net_input(self, X):
        return np.dot(X, self.w_[1:]) + self.w_[0]

    def predict(self, X):
        return self.net_input(X)
```

Fig. 6. Part of a machine learning program for a classification task without a single step function.

To see our *LinearRegressionGD* linear *regressor* in action, we will use the RM (Profit volume) variable from the IT product development prediction dataset of innovative start-up projects as an explanatory model training variable that can predict MEDV (Interaction Tariffs).

Moreover, we standardise the variables for better convergence of the gradient descent algorithm. The corresponding source code looks like this (Fig. 7).

```
X = df[['RM']].values
y = df['MEDV'].values

from sklearn.preprocessing import StandardScaler
sc_x = StandardScaler()
sc_y = StandardScaler()
X_std = sc_x.fit_transform(X)
y_std = sc_y.fit_transform(y)
lr = LinearRegressionGD()
lr.fit(X_std, y_std)
```

Fig. 7. Standardisation of variables for gradient descent algorithm.

Let us plot the value against the number of epochs to check the convergence of the linear regression. As we can see in the chart below, the algorithm of the gradient downturn converged after the fifth epoch (Fig.8).

Note that it is more efficient (fewer emissions) to work with non-standardized variables in the linear regression object *LinearRegression* library *scikit-learn*, which uses the

dynamic library LIBLINEAR and advanced optimisation algorithms. The RANSAC model is used in this case.

As we can see from the execution of the previous source code, the LinearRegression model of the *scikit-learn* library, fitted with non-standardized variables RM and MEDV, produced other model coefficients.

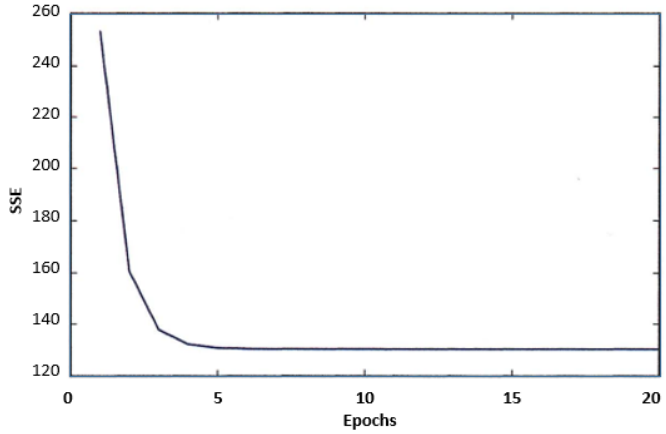


Fig. 8. Graph of cost to the number of epochs.

Let us compare it with our implementation based on gradient descent by constructing the MEDV chart to RM (Fig.9).

```
def lin_regplot(X, y, model):
    plt.scatter(X, y, c='blue')
    plt.plot(X, model.predict(X), color='red')
    return None
```

Fig. 9. A code fragment for plotting a point chart of linear regression

Having built a graph of training data and performed the model fitting by executing the above source code, we can now see that the overall result looks identical to our implementation based on the gradient descent (Fig. 10).

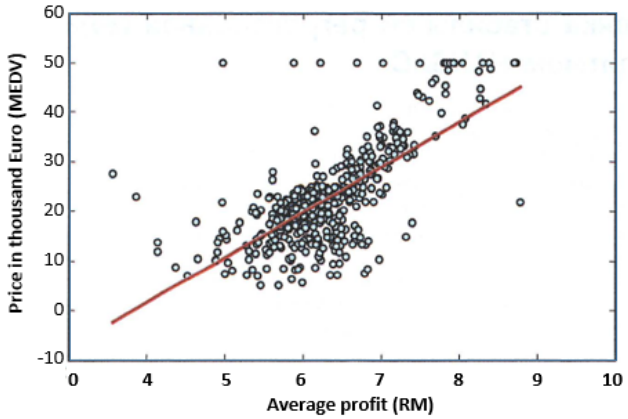


Fig. 10. The trend of IT projects development depending on the level of profit, based on training data.

Using the RANSAC model, we have reduced the potential impact of emissions in this dataset, but we do not know if this approach has a positive impact on predictive capacity on previously unseen data.

After executing the next source code, we should see a residual graph with a line passing through the beginning of the X-axis, as shown below (Fig.11).

In case of perfect prediction, the residues would be strictly zeros, which in real and practical applications we will probably never encounter.

However, we expect from a good regression model that the errors are distributed randomly, and the remains are randomly scattered around the midline.

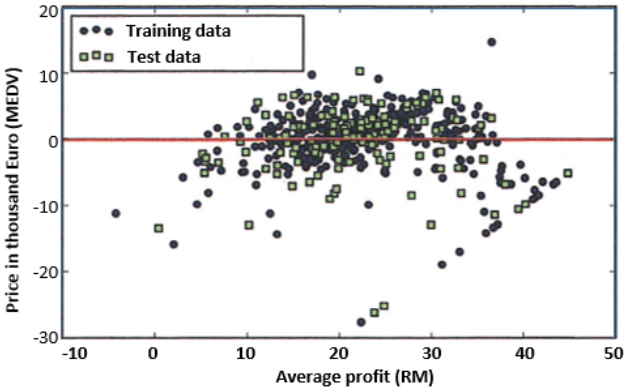


Fig. 11. Errors in fitting training data to the test data.

Another useful quantitative measure for assessing the quality of a model is the so-called weighted mean square error (mean squared error, MSE), that is, simply the average value of the SSE cost function, which we minimize to fit the linear regression model. MSE is useful for comparing different regression models or for fine-tuning their parameters by searching the parameter grid and cross-checking:

$$MSE = \frac{1}{N}\sum_{i=1}^{n}(y^i - \hat{y}^i)^2 \quad , \tag{5}$$

```
>>>
from sklearn.metrics import mean_squared_error
print('MSE тренировка: %.3f, тестирование: %.3f' % (
    mean_squared_error(y_train, y_train_pred),
    mean_squared_error(y_test, y_test_pred)))
```

Fig. 12. Errors in fitting training data to the test data.

We will see that the MSE on the training set is 19.96, and the MSE of the test set is much larger with a value of 27.20.

V. CONCLUSION

The analysis of the considered approaches to the definition of effective interaction of customers with a mass information system allows defining use of modern mechanisms for the analysis on the base of Customer Journey Maps. In this case, big data sets are generated, which cannot be analysed by traditional methods. To solve this problem, the authors used machine learning algorithms of neural network types.

An exploratory data analysis method was applied to the available data to find repetitive patterns and anomalies, which proved to be effective. It has allowed constructing a model of realisation of linear regression with the use of the approach based on gradient optimisation.

This approach allowed to construct a base for prediction of processes of the development program of the start-up project product.

REFERENCES

[1] Trends in the development of the global market for information technology. [Online]. Available: http://eir.pstu.edu/ handle/123456789/4299

[2] Euripidis Loukis, Marijn Janssen, Ianislav Mintchev, Determinants of software-as-a-service benefits and impact on firm performance, Decision Support Systems, Volume 117, 2019, pp. 38-47.

[3] B. Brown, K. Swani, “Introduction to the special issue: B2B advertising”, Industrial Marketing Management, February 2020, DOI: 10.1016/j.indmarman.2020.02.006

[4] Customer Journey Map: how to understand what the consumer needs. (2019) Available at: https://www.uplab.ru/blog/customer-journey-map/

[5] S.Gordon, M. Linoff, “Data mining techniques: for marketing, sales, and customer relationship management”. Published by Wiley Publishind Inc., 10475 Crosspoint Boulevard, Indianapolis, 830 p., 2011.

[6] IBM Institute for Business Value. Analytics: The real-world use of big data in consumer products, 2013.

[7] Seductive Interaction Design: Creating Playful, Fun, and Effective User Experiences (Voices That Matter) 1st Edition (2019) Available at: https://www.amazon.com/Seductive-Interaction-Design-Effective-Experiences/dp/0321725522/

[8] A. Zudov “Modeling of potential rule interactions in active databases”. In journal “Modern problems of science and education” (2015). № 1-1.; URL: http://www.science-education.ru/ru/article/view?id=17745.

[9] V. Gogunskii, O. Kolesnikov, G. Oborska, S. Harelik, D. Lukianov, “Representation of project systems using the Markov chain”, Eastern-European Journal of Enterprise Technologies, № 1/3 (85), 2017б pp. 25-32.

[10] International Project Management Association. Individual Competence Baseline Version 4.0. International Project Management Association, 432 p., 2015.

[11] A Guide to the project management body of knowledge (PMBoK Guide). Sixth Edition – USA: PMI Inc., 537 p., 2017.

[12] R. Turner, Guide to project-based management, tran. from English, Moskow, Grebennikov Publishing House, 552 p., 2007.

[13] V. Morozov, O. Kalnichenko, S. Bronin, “Development Of The Model Of The Proactive Approach in Creation Of Distributed Information Systems”. Eastern-European Journal of Enterprise Technologies, № 43/2 (94), pp. 6-15 (2018).

[14] A. Timinsky, O. Voitenko, I. Achkasov, “Competence-based knowledge management in project-oriented organisations in bi-adaptive contex”t. Proceedings of the IEEE 14th International Scientific and Technical Conference on Computer Sciences and Information Technologies (CSIT-2019). - Lviv, 2019, pp. 17-20.

[15] O. Maimon, L. Rokach, “Data Mining and Knowledge Discovery Handbook”, 2005, [Online]. Available: http://www.bookmetrix.com/detail/book/ae1ad394-f821-4df2-9cc4-cbf8b93edf40

[16] V. Morozov, O. Kalnichenko, M. Proskurin, O. Mezentseva, “Investigation of Forecasting Methods the State of Complex IT-Projects With Using Deep Learning Neural Networks”, Published in the book "Lecture notes in computational intelligence and decision making" (series "Advances in intelligent systems and computing"), vol. 1020, 2020, pp. 261-280,.

[17] E. Loukis, M. Janssen, I. Mintchev, “Determinants of software-as-a-service benefits and impact on firm performance”, Decision Support Systems, Volume 117, 2019, pp. 38-47.

[18] K. Swani, B. Brown, Su. Mudambi, “The untapped potential of B2B advertising: A literature review and future agenda”, Industrial Marketing Management, 2019.

[19] D. Herhausen, K. Kleinlercher, P. Verhoef, T. Rudolph, “Loyalty Formation for Different Customer Journey Segments”, Journal of Retailing, 2019.

[20] V. Morozov, O. Kalnichenko,M. Proskurin, “Methods of Proactive Management of Complex Projects Based on Neural Networks”, Proceedings of the 2019 10th IEEE International Conference on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications, 2019, pp. 964-969.

[21] D. Garaedagi, “System thinking. How to manage chaos and complex processes. Platform for modeling business architecture”, Grevtsov Buks (Grevtsov Publicher). 480 p., 2011.

[22] A. Novikov, Ezhov A.A. “Rosenblatt's multilayer neural network and its application for solving the problem of signature recognition”, Bulletin of TSU. Technical science. No. 2. pp. 188-197, 2016.

[23] D. Komashinsky, D. Smirnov, “Neural networks and their use in control and communication systems”, Hotline-Telecom. p. 94, 2003.

[24] Li, L., Fan, K., Zhang, Z., Xia, Z. “Community detection algorithm based on local expansion K-means”, Neural Network World, 26(6), 2016, pp.589-605.

[25] I. Prigogine, G.Nikolis, “Knowledge of the complex. Introduction”, Per. from English, M .: Lenar, 360 p., 2017.

[26] G. Carmantini, S. Rodrigues, P. Graben, M. Desroches, “A modular architecture for transparent computation in recurrent neural networks”, Neural Networks, #85, 2017,pp.85-105.

[27] A. Hosseini, “A non-penalty recurrent neural network for solving a class of constrained optimization problems”, Neural Networks, 2016, #73, pp.10-25

[28] A. Polaine, L. Løvlie, “Service Design: From Insight to Implementation”, Rosenfeld Media, 2013, 216 p.

[29] S. Anderson, “Seductive Interaction Design: Creating Playful, Fun, and Effective User Experiences”, New Riders; 240 p., 2011 https://www.amazon.com/Seductive-Interaction-Design-Effective-Experiences/dp/0321725522/

[30] Installing and getting started, [Online]. Available: https://seaborn.pydata.org/installing.html

IEEE Third International Conference on Data Stream Mining & Processing
August 21-25, 2020, Lviv, Ukraine

Using Hybrid Neural Networks to Detect DDOS Attacks

Olexander Belej
Department of Computer-Aided Design
Lviv Polytechnic National University
Lviv, Ukraine
Oleksandr.I.Belei@lpnu.ua

Liubov Halkiv
Department of Management of Organizations
Lviv Polytechnic National University
Lviv, Ukraine
Lubov.I.Halkiv@lpnu.ua

***Abstract*—The results of the development of a denial of service network attack detection method for various services of storing, processing, and transmitting data on the Internet are presented. The focus is on the detection of low-level denial of service attacks. The opinion is refuted that special means for detecting denial of service attacks are not required since the fact of a detecting denial of service attacks cannot be overlooked. It is shown that for effective counteraction it is necessary to know the type, nature, and other indicators of a denial of service attack, and distributed attack detection systems allow you to quickly obtain this information. Also, the use of this type of attack detection system can significantly reduce the time it takes to determine the fact of an attack from 2-3 days to several tens of minutes, which reduces the cost of traffic and the downtime of the attacked resource. A hybrid neural network based on the Kohonen network and a multilayer perceptron is used as a detection module. The work of the created prototype of the attack detection system, the methodology for the formation of the training sample, the course of experiments, and the topology of the experimental stand are described. The results of an experimental study of the prototype are presented, during which errors of the first and second kind, respectively.**

Keywords—attack detection; low-intensity detecting denial of service attacks; hybrid neural network; computer network security.

I. INTRODUCTION

One of the main trends in recent years in the field of computer crime is the increasing number and complexity of attacks on the availability of information, as one of the three main criteria, along with the confidentiality and integrity of information security object. These attacks form a denial of service (DDoS) class. If the attack is carried out simultaneously from a large number of computers, a DoS attack occurs.

In general, DDoS attacks target both the network as a whole and server clusters, and ultimately the hosts are tasked with maximizing the resources available to significantly degrade or discontinue service to regular users. The most commonly attacked resources are channel width, CPU time of servers and routers, and specific protocol implementations. Among them are: SYN-attack aimed at overflowing the TCP stack of the operating system; ICMP directional broadcasting - such packets are sent to the attacker and responses from him reduce network bandwidth; DNS-flood attacks that use a certain weakness of the DNS-protocol and are aimed at significantly increasing traffic to the attacked.

To minimize the consequences of DDoS attacks, their detection and classification are extremely important and at the same time a difficult task. The main way to recognize a DDoS attack is to detect anomalies in the traffic structure [1]. A fundamental prerequisite for detecting attacks is the construction of traffic control characteristics during network operation under normal conditions with the subsequent search for anomalies in the traffic structure [2]. An anomaly in network traffic is an event or condition in the network characterized by a statistical deviation from the standard traffic structure obtained based on previously collected profiles and control characteristics.

At the same time, the existing methods for detecting DDoS attacks that can efficiently recognize transport-level DDoS attacks are ineffective for detecting application-level low-rate DDoS attacks [3]. This class of DDoS attacks has arisen relatively recently and today represents the main threat to the availability of information in distributed computer networks [4]. It is quite difficult to distinguish the traffic generated during these attacks from legal HTTP traffic, also, the data transmission channels are practically not overloaded.

A common factor necessary for the implementation of Low-Rate DDoS attacks is the presence of a large number of compromised or voluntarily participating hosts and a gross “blocking” of the attacked node with packets. It is “rudeness” in the implementation of these attacks that can nullify the whole effect in the case of detecting large volumes of anomalous traffic by network monitors [5].

The algorithm integrates the ant colony optimization algorithm (COA) with the genetic algorithm (GA) [6]. The experiment indicates that the proposed algorithm has preferable performance both in balancing resources and guaranteeing the Quality of Service (QoS). To overcome the defects of the cloud computing data center in resource management, and ensure that the cloud computing can supply better QoS service, the ant colony optimization, COA will be applied to the cloud resource scheduling according to the actual QoS parameters requirement of the environment for the cloud computing [7].

DDoS attacks are aimed at reducing the bandwidth of low-intensity TCP traffic flows to avoid detection. Using the TCP stack retransmission vulnerability, zero bandwidth can be achieved by mixing specially selected DDoS traffic patterns with primary traffic. It is quite difficult to distinguish the traffic generated during these attacks from legal HTTP traffic, also, the data transmission channels are practically not overloaded. These attacks lead to loss of requests and responses, i.e. the actual failure of web servers based on Microsoft IIS, Apache, and other systems. Besides, the attack can be adapted to affect SMTP and even DNS servers. In the DDoS variant, the increased use of the DNS amplification technique for the avalanche-like amplification of the power of the network botnet is of particular danger. These facts determine the relevance of developing new mechanisms for detecting low-intensity distributed attacks of the application level such as a denial of service in computer networks using artificial intelligence methods.

II. Problem formulation

Bandwidth management in TCP is carried out on 2-time scales. On the small timeline of packets passing through the communication channel to the destination and back (RTT), usually from 10 to 100 milliseconds, the TCP stack uses additive-increase multiplicative-decrease (AIMD) control to transmit each traffic stream at the same speeds through the bottleneck, the so-called bottleneck. When the communication channel begins to “clog” and a large number of losses occur, the TCP stack starts working on the 2nd, larger timeline with marks of retransmission packet timeouts. To avoid channel “clogging”, the traffic flow is reduced to one packet, and after the RTO time has passed, the packet is sent again. With subsequent losses, the RTO time doubles with each subsequent timeout. If the packet is successfully received, the TCP stack begins to use AIMD control [8].

To conduct a Low-Rate DDoS attack, it is necessary to take traffic flows in the form of pulses and consider periodic impulse attacks, consisting of short peaks with a specially selected duration, repeating at a specific, specially selected frequency on a slow time scale. If for the first stream of TCP traffic, the total traffic during the peak is sufficient for packet loss to occur, then this stream will “fall off” by timeout and an attempt will be made to send a new packet over time RTO If the frequency of sending DDoS traffic coincides with the normal traffic RTO, regular traffic will constantly receive a timeout, as a result, the losses will approach 100% and the throughput will approach zero. Also, if the DDoS sending period is approximately equal, but lies outside the RTO range, then a significant (but not complete) decrease in bandwidth will be observed. This mechanism was considered in more detail in [9], and the TCP stack timeout mechanism in [10].
In Fig. 1 shows the dependence of web service available on the frequency of the attack. In an ideal model without errors in network delays and the performance of attacking machines, as well as losses on the Internet, it is safe to say that such an attack is carried out without serious costs for the communication channel bandwidth of the attacking machines.

Conducting a low-intensity attack on the webserver greatly affected the overall throughput. After the start of the attack, a decrease in the bandwidth of the communication channel was observed (Fig. 2), which led to a decrease in the quality of customer service by the Http-server.

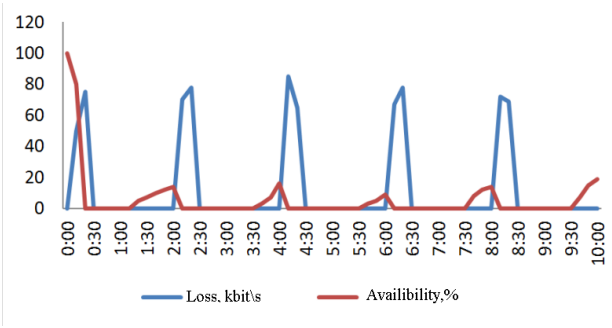


Fig. 1. Graph comparing the availability of the attacked system and losses during an attack

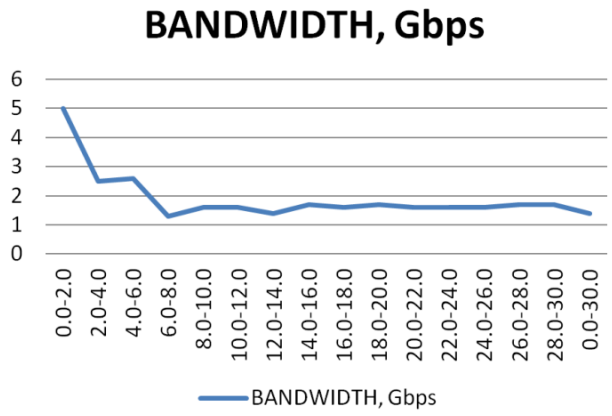


Fig. 2. Communication bandwidth graph during the attack

Existing methods of detecting DoS attacks, based on a statistical analysis of threshold values that allow efficient recognition of transport-level attacks, are ineffective for detecting low-level DoS attacks. This makes it urgent to develop new mechanisms for detecting low-intensity attacks of the application level such as "denial of service" in computer networks using artificial intelligence methods.

To detect an attack, it is necessary to identify the fact that a certain set of packets of the same type appears in incoming traffic periodically and then determine whether this set belongs to a certain class. The sequence of packets, in this case, does not play a leading role, temporary information is taken into account when breaking incoming traffic into windows.

A slow loris attack was modeled as part of the study. The attack is based on vulnerabilities in the HTTP protocol. A Slow HTTP POST attack works as follows: an attacker sends a POST header with a legitimate “Content-Length” field, which allows the webserver to understand how much data is being sent to it. As soon as the header is sent, the body of the POST message begins to be transmitted at a very slow speed, which allows the server to use resources much longer than necessary, and, as a result, interfere with the processing of other requests. Several thousand of these connections can disable the webserver for several minutes.

For recruiting bots, online games of unlucky players were used to send specially formed HTTP requests to the target system. The simplicity of the attack allows you to effectively use for this a simple Java applet that runs during an online game. As soon as the victim accepts a self-signed applet, he begins to attack while the user is playing the game. After exiting the game and closing the browser, the attack stops, and

the applet is deleted. Finding that the computer has become the source of the attack is quite problematic, since the computer does not become infected in the classical sense of the word, and it is difficult to distinguish attack traffic from legal Http-traffic. Also, the Internet channel is almost not overloaded.

The attack crashes web-servers with Microsoft IIS and Apache within the framework of HTTP or HTTPS protocols and any "secure" connections like SSL, VPN, and others. Also, the attack can be adapted to work with SMTP and even DNS-server. The load balancing software currently used to prevent similar DDOS attacks is ineffective against the new technique.

III. MODELING A LOW-INTENSITY DDOS ATTACK

We describe the method and architecture of the prototype attack detection system (ADS). It is believed that special tools for detecting DDoS attacks are not required since the fact of a DDoS attack cannot be overlooked. In many cases, this is true. However, quite often successful attacks were noted, which were noticed by the victims only after 2-3 days. It happened that the negative consequences of an attack consisted of excessive expenses for paying for traffic, which was found out only when an invoice was received.

Most modern attack detection systems detect attacks by monitoring behavior profiles or by searching for specific string signatures. Using these methods, it is practically impossible to create a complete database containing the signatures of most attacks. There are three main reasons for this:

1) New signatures must be created manually. Signatures of known attacks that are already included in the database cannot guarantee reliable protection without constant updates.

2) Theoretically, there is an infinite number of methods and variants of attacks, and for their detection, you will need a database of infinite size. Thus, there is the possibility that a certain attack, not included in the database, can be successfully carried out.

3) Modern detection methods cause a large number of false alarms. In this way, legitimate network developments.

The approach proposed in the article involves the use of a neural network to detect abuse, distributed over time, and based on the joint creation of an artificial data stream by several sources.

Timed attacks are attacks carried out over one long period. In a joint attack, several attackers are working in parallel from different sources distributed in space. Each of them individually can take actions that may seem harmless. Attacks become apparent only when all events are considered together.

The main advantage of ADS using neural networks is that the neural network is not limited to the knowledge that the programmer put into it. They have the opportunity to learn from previous events - both on abnormal and normal traffic. Due to this, high efficiency and adaptability of ADS are achieved (Fig. 3).

There are various options for the use of neural network attack detection systems (NNDS) - analysis of all network traffic in a protected network segment, analysis of commands entered by the user, analysis of state transitions, etc. This article proposes a detection method and architecture of SOA based on the use of a hybrid neural network as a traffic analyzer.

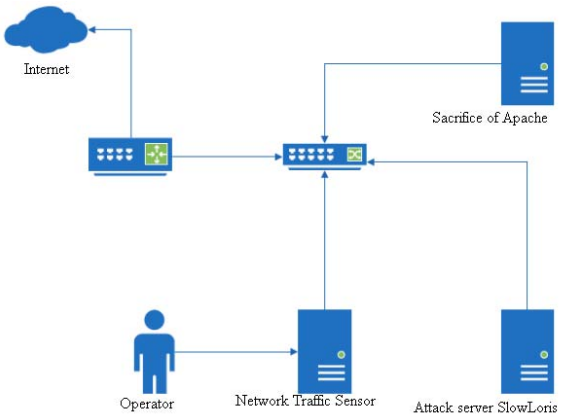


Fig. 3. The block diagram of intrusion detection

The data streams were divided into sets of 150 packets, and in addition to the data from the packet headers of the network and transport layers, the first 50 characters of the application-level payload were also taken into account. Besides, the presentation of destination port values is normalized to clarify the application protocol used.

As a rule, not a single event constituting the traffic of a DDoS attack separately looks suspicious. Also, even two or three events are not unusual. Nevertheless, taken together, in a certain sequence, repeated at a certain time, these events are an indication that an attack is being carried out. A further complication of the discovery is that these events, although in a certain sequence, can come from various sources and alternate with other, normal events.

To handle abnormal traffic, it is proposed to use a hybrid neural network consisting of a SOM and a multilayer perceptron. The prototype architecture of the proposed attack detection system is shown in Fig. 4.

Using the Kohonen map, clustering of 50-character events occurs in matrix nodes, in which events of similar numeric characters will be grouped. Individual nodes will represent specific attack scenarios.

The input vector for ADS contains the following components: 1–4 - bytes that make up the packet address (for IPv4), given in the range 0-1; 5 - packet port, given in the range 0–1; 6–55 - the first 50 bytes of packet data, listed in the range 0–1. After that, packet header data and grouping information are fed to the input of a multilayer perceptron trained to recognize anomalous traffic, but already taking into account event information, i.e. belonging of the package to one or another scenario group. This allows not only detecting anomalies in single packets but also revealing that the packet belongs to a time-distributed attack.

The input vector format for the neuro-network is as follows. The input sample is split into disjoint parts of 150 packets in length. All packets pass through the Kohonen network. The package is assigned the cluster number (the number of the winning neuron).

Clusters are enlarged – 5 neighboring clusters come together. It turns out 100 enlarged clusters.

For enlarged clusters, a histogram is considered to represent traffic. That is, the first component of the vector is the number of packets falling into the first enlarged cluster (clusters 1–5), and so on. The histogram is normalized; components are

reduced to a range of 0–1. After that, the delta between adjacent vectors is calculated.

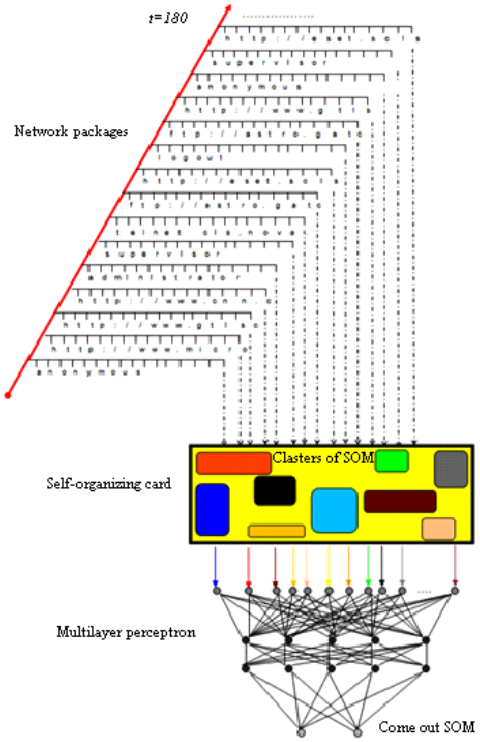


Fig. 4. The architecture of attack detection system

The proposed method takes into account two types of traffic activity. When analyzing individual cluster nodes, individual attack scenarios are identified. Also, sudden changes in the power of the network stream are analyzed to detect bursts of the load. A joint analysis of these two types of events allows a more accurate response to the appearance of DDoS traffic while reducing the number of false positives associated with a legitimate increase in bandwidth usage.

To train an artificial neural network, two types of network traffic were modeled - normal and abnormal. The first contained packets that appeared on the network during normal operation and the second imitated a distributed attack.

Data collection was carried out in 2 stages:

1) The stage of collecting normal behavior. A PHP script has been created on the attacked server that performs a "heavy" SQL query.

2) The stage of collecting abnormal behavior. To simulate traffic that occurs during an attack, 3 copies of the slow loris script are launched on the attacking server with the difference in the parameter responsible for the delay between reconnections.

The NNDS was trained and tested on disjoint windows since when using sliding windows during testing, the ability to control the degree of use of the network channel bandwidth is worsened.

Kohonen map training is carried out on separate packages, sequentially selected from the window. Before applying to the Kohonen network and the perceptron, all data were normalized to the range [0;1]. The Kohonen network has a hexagonal structure of neuron connections and sizes 25 by 20 clusters.

Window size is calculated by the formula:

$$\omega = u \times \frac{S_{byte}}{8 \times P_{\min}} \quad (1)$$

where S_{byte} is the speed of information transmission in the network in bytes per second; P_{min} is the minimum theoretical packet size; u is a coefficient showing the level of utilization of the information transmission channel.

The Ethernet standard requires that there be a 12-byte "silence" period between packets, so the end of one packet and the beginning of the next are determined. At the end of each packet, it is also necessary to transmit a CRC code to check the integrity of the transmission, and at the beginning of the packet, a mandatory preamble of 8 bytes. There is another limitation - the minimum packet size is 60 bytes.

Given all the restrictions, the packets should be at least 84 bytes each. For 1 Gbit/s, we get a theoretical limit of *1000000000/84*8=1488095* packets\second. For 100 Mbps - *100000000/84*8=148810* packets\second.

The choice of the coefficient u, which shows the level of utilization of the information transmission channel, depends on the network load in normal operation.

The normal mode is the channel load in the range from 3% to 10%. We believe that for a network with a maximum speed of 100 Mbit/s, from the effective operation of the attack detection system, values from normal to high are found in this range. The maximum workload at which analysis is possible without packet loss is 40%. Since the traffic generated during the low-intensity attack is minimal, we will experiment on the lower limit of congestion.

Accordingly, *0,01*148810~1500* packets\second with normal disposal at 1%.

Besides, we use two more values: window size 30 packets - the minimum value found in the IDS Snort rules for low-intensity attacks; window size 180 - corresponds to the arrival rate of 1 packet per second.

For the analysis of string data, the first 50 characters of the raw data of the packet payload are used, since the bulk of the data significant for the analysis in the collected network packets is contained at the very beginning of the packet. Characters are represented in ASCII code. The number of clusters in the Kohonen map som_size "=" 500.

The perceptron has the following structure - two hidden and an output layer, activation functions in the hidden layers - hyperbolic tangent, in the output layer - linear. The number of neurons in the hidden layers is 21 and 7. The training method - trail - is fast and expensive in memory, but with the input data considered, there were no problems with memory overrun.

IV. Discussion

As indicated above, in the general case, attacks can be directed at various network services. From packet processing by the stack, services are characterized by different values of the destination ports. In this case, the attack scenario does not change. Therefore, it is proposed to control each port separately, and, accordingly, use separate clusters and classifier. The prototype will consider one service running on port 80.

The system uses a sensor based on the libpcap library, which allows the use of a fast packet capture engine and a flexible filter expression system. With its help, the traffic

directed to the protected service is filtered out and transmitted to the analysis subsystem.

To train an artificial neural network, two types of network traffic were modeled - normal and abnormal. The first contained packets that appeared on the network during normal operation and the second imitated a distributed attack.

Data collection was carried out in 2 stages:

1) The stage of collecting normal behavior. A PHP script has been created on the attacked server that performs a "heavy" SQL query (Loop in N iterations). Apache benchmark was launched from the attacking server to create legitimate connections to the server. 10 threads of 100 requests.

2) The stage of collecting abnormal behavior. To simulate traffic that occurs during the attack, 3 copies of the slow loris script are launched on the attacking server with the difference in the parameter responsible for the delay between reconnections: the size of the normal set is 459565 packets; the size of the attack set is 428890 packets.

In Fig. 5 shows a graph of bandwidth congestion in normal operation at the stage of collecting normal behavior.

Fig. 6 shows the graphs for the stage of collecting attack traffic - packet loss due to the unavailability of the service for new connections (red line) and normal traffic processed by the service (blue line).

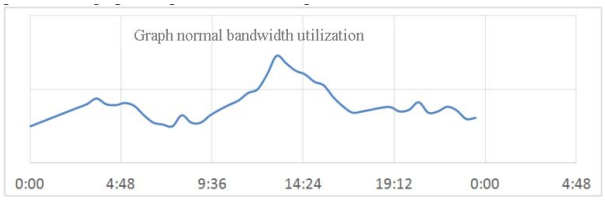


Fig. 5. The normal bandwidth utilization

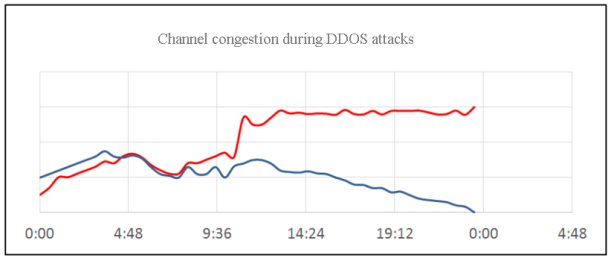


Fig. 6. A channel load under attack

In Fig. 5 and 6 on the abscissa axis the time of day is postponed starting from 00,00 hours, on the ordinate axis the maximum is 100% bandwidth utilization. The results of the experimental study are presented in Table I.

TABLE I. THE RESULTS OF THE PROTOTYPE OF THE INTRUSION DETECTION SYSTEM

N.	Line length	The value of the training set for SOM packages	The value of the training set for FFNET, windows	Win-dow size	Shift size	Studying time	The result of the test sample	
							Type 1 error	*Type 2 error*
1	20	5000	4000	30	3	00:01:34	8,1827e-04	0,0050
2	50	30000	24000	30	3	00:26:48	0,0367	0,0172
3	20	5000	800	180	18	00:01:48	0	3,4378e-04
4	50	30000	4800	180	18	00:18:33	0,0449	0,0449
5	20	5000	90	1500	150	00:01:11	0,0554	0,5287
6	50	30000	540	1500	150	00:20:33	0	0,0033
7	20	5000	900	30	15	00:01:18	0,0118	0,0900
8	50	30000	5400	30	15	00:22:44	0,0289	0,0232
9	20	5000	160	180	90	00:01:05	0,0011	0,1124
10	50	30000	960	180	90	00:16:06	0	0,1447
11	20	5000	20	1500	750	00:01:05	0,1154	0,8386
12	50	30000	120	1500	750	00:15:21	0	0,0471

During the experiments, the influence of the following values on the effectiveness of the method was investigated.

Size of string data: 20 bytes; 50 bytes.

The size of the training sample: 5000 packages; 30,000 packages.

Window size (in packages): 30; 180; 1500.

Sheer size: by 10%; by 30%.

Recognition was carried out on a test sample. Estimated proximity to the standard. Recognition was considered successful if the absolute difference between the reference and actual values for each component of the output vector did not exceed 0,3.

In Fig. 7 presents the results of the clustering.

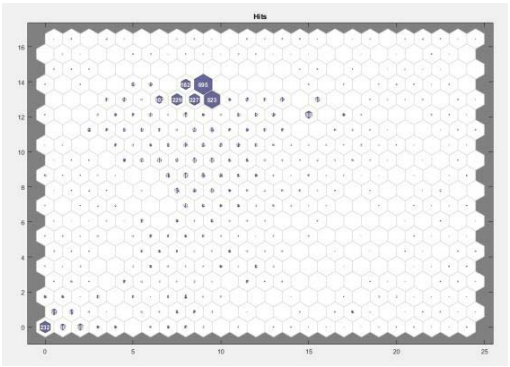


Fig. 7. Packet Clustering Results by the Kohonen Network

In Fig. 8 shows the error values in various samples during the training of the perceptron.

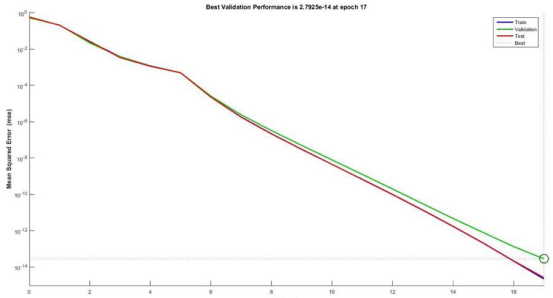


Fig. 8. Error on different samples when training perceptron

It can be seen that the number of false positives does not exceed 0,115% in the worst case, which was due to the minimum size of the analyzed data and the training sample. The goal skip value rose to 0,84% also in the eleventh experiment. The best results are shown in experiments 1, 6, 12, which confirms theoretical assumptions.

V. CONCLUSION

An analysis of the mechanism for implementing low-speed DDoS attacks using the example of a slow role-based attack showed that this type of attack can be organized with a

minimum ratio of the required channel power of the attacking computer and the attacked server, which was unattainable before. Analysis of vulnerability statistics also shows that a large number of sites are susceptible to this attack, but on the server-side, it is often even difficult to diagnose a site attack, since traffic does not exceed normal values.

The article describes the developed method for detecting low-intensity denial of service attacks based on the use of a hybrid neural network.

The method demonstrates a high percentage of attack detection by reducing the number of undetected attacks (errors of the second kind), while the speed of operation depends only on the processing speed of incoming packets. The proposed method for detecting low-intensity attacks makes it possible to efficiently determine traffic belonging to a time-distributed attack.

The architecture and prototype of a software package designed to detect low-intensity denial of service attacks were developed, and an experimental study of the effectiveness of the developed complex was carried out.

ACKNOWLEDGMENT

This paper has been written as a result of the realization of the “International Academic Partnerships Program”. The project is funded by The Polish National Agency for Academic Exchange (NAWA), the contract for refinancing no. PPI/APM/2018/1/00031/U/001.

REFERENCES

[1] M. Barati, A. Abdullah, N. I. Udzir, R. Mahmod, and N. Mustapha, "Distributed Denial of Service detection using hybrid machine learning technique," 2014 International Symposium on Biometrics and Security Technologies (ISBAST), Kuala Lumpur, 2014, pp. 268-273.

[2] W. Zhou W. Jia S. Wen Y. Xiang and W. Zhou "Detection and defense of application-layer DDoS attacks in backbone web traffic" Futur. Gener. Comput. Syst. Aug. 2013.

[3] S. Guha, S. S. Yau, and A. B. Buduru, "Attack Detection in Cloud Infrastructures Using Artificial Neural Network with Genetic Feature Selection," 2016 IEEE 14th Intl Conf on Dependable, Autonomic and Secure Computing, 14th Intl Conf on Pervasive Intelligence and Computing, 2nd Intl Conf on Big Data Intelligence and Computing and Cyber Science and Technology Congress (DASC/PiCom/ DataCom/CyberSciTech), Auckland, 2016, pp. 414-419.

[4] Z. Chiba, N. Abghour, K. Moussaid, A. E. Omri and M. Rida, "A Clever Approach to Develop an Efficient Deep Neural Network-Based IDS for Cloud Environments Using a Self-Adaptive Genetic Algorithm," 2019 International Conference on Advanced Communication Technologies and Networking (CommNet), Rabat, Morocco, 2019, pp. 1-9.

[5] T. Nguyen, N. Tran, B. M. Nguyen, and G. Nguyen, "A Resource Usage Prediction System Using Functional-Link and Genetic Algorithm Neural Network for Multivariate Cloud Metrics," 2018 IEEE 11th Conference on Service-Oriented Computing and Applications (SOCA), Paris, 2018, pp. 49-56.

[6] Y. Dai, Y. Lou and X. Lu, "A Task Scheduling Algorithm Based on Genetic Algorithm and Ant Colony Optimization Algorithm with Multi-QoS Constraints in Cloud Computing," 2015 7th International Conference on Intelligent Human-Machine Systems and Cybernetics, Hangzhou, 2015, pp. 428-431.

[7] Y. Zhu and H. Liang, "Research for the virtual machine-oriented cloud resource scheduling algorithm," 2013 6th International Conference on Information Management, Innovation Management and Industrial Engineering, Xi'an, 2013, pp. 133-136.

[8] Y. SU, X. MENG, Q. MENG and X. HAN, "DDoS Attack Detection Algorithm Based on Hybrid Traffic Prediction Model," 2018 IEEE International Conference on Signal Processing, Communications and Computing (ICSPCC), Qingdao, 2018, pp. 1-5.

[9] I. Abdulqadder, D. Zou, I. Aziz, B. Yuan, and W. Dai, "Deployment Of Robust Security Scheme In SDN Based 5G Network Over NFV Enabled Cloud Environment," in IEEE Transactions on Emerging Topics in Computing, 2019.

[10] J. Shen, D. Liu, Q. Liu, X. Sun and Y. Zhang, "Secure Authentication in Cloud Big Data with Hierarchical Attribute Authorization Structure," in *IEEE Transactions on Big Data,* 2017.

IEEE Third International Conference on Data Stream Mining & Processing
August 21-25, 2020, Lviv, Ukraine

Deep Neo-Fuzzy Neural Network and its Accelerated Learning

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

Tymofii Antonenko
Department of Artificial Intelligence
Kharkiv National University of Radio Electronics
Kharkiv, Ukraine
tymofii.antonenko@nure.ua

***Abstract*—The using of a neo-fuzzy neuron as the main component of a neural network is considered. A deep neo-fuzzy neural network architecture and corresponding backpropagation algorithm are described.**

Keywords—neo-fuzzy neuron, multilayer neural network, F-transform.

I. INTRODUCTION

Nowadays artificial neural networks (ANN) are widely used for solving tasks of processing information of different nature due to their universal approximation abilities which are achieved during the learning processes based on input datasets. More advanced results could be achieved with help of deep neural networks (DNN) [1-5], which outperform traditional ANNs quality-wise but their learning process depends on a set of computational issues and they are more time consuming. In the basis of ANN is a Rosenblatt's perceptron [6] and as an activation function traditional sigmoid is being used or hyperbolical tangent. Three layers perceptron guarantees high quality of approximation of complex functions defined in limited domain [7]. Attempts to improve the quality of results with increasing number of hidden layers of ANN faced a problem that is related to vanishing and exploding gradient [8] which caused by a form of sigmoidal activation functions. Instead of sigmoidal functions, DNN uses linear rectified family which are typical representatives of the rectified linear unit (ReLU), which doesn't suffer from computational issues of a gradient. Using such functions does not cause any problem related to computation of gradient and their constant derivatives allow to simplify delta-rule for learning of each independent neuron. However, usage of ReLU doesn't satisfy the requirements of approximation theorems which underlying ANN and, first of all the requirements of Cybenko's theorem [7]. Thereby DNN ensure piecewise linear approximation and its high quality is achieved by magnifying the number of hidden layers in the network. In this case, increasing the number of layers decreases the speed of the learning process and requires more data for training.

II. ARCHITECTURE OF DEEP NEO-FUZZY NEURAL NETWORK

The proposed network has traditional multilayer feedforward architecture, which in common case includes s layers of information processing.

As an input (zero) layer $x(k) \in R^n$ - vector of input signal is used, $x(k) = (x_1(k), x_2(k), \ldots, x_n(k))$ where $k = 1, 2, \ldots, N$ – is a number of observation in the training dataset or index of the current discrete time. An output signal of the network is a vector $\hat{y}(k) = (\hat{y}_1(k), \hat{y}_2(k), \ldots, \hat{y}_m(k))^T \in R^m$. Further, to simplify the notation we will use the next representations:

$$x(k) \equiv o^{[0]}(k) = (o_1^{[0]}(k), \ldots, o_{i_0}^{[0]}(k) \ \ldots, o_{n_0}^{[0]}(k))^T,$$

$$\hat{y}(k) \equiv o^{[s]}(k) = (o_1^{[s]}(k), \ldots, o_{i_s}^{[s]}(k), \ldots, o_{n_s}^{[s]}(k))^T.$$

Thus, the input signal of the p-th layer (p=1, 2, …, s) is a vector

$o^{[p-1]}(k) = (o_1^{[p-1]}(k), \ldots, o_{n_{p-1}}^{[p-1]}(k))^T \in R^{n_{p-1}}$, and as an output we have a vector $o^{[p]}(k) = (o_1^{[p]}(k), \ldots, o_{i_p}^{[p]}(k), \ldots, o_{n_p}^{[p]}(k))^T \in R^{n_p}$. Herewith, the neo-fuzzy neural network contains $\sum_{p=1}^{s} n_p$ of neurons.

A node of this architecture is a neo-fuzzy neuron (NFN) [10-12] with n_{p-1} of inputs and one output $o_{i_p}^{[p]}$. On Fig.1 the architecture of i_p-th $NFN_{i_p}^{[p]}$ of the p-th layer of the network is shown.

Every i_p-th ($i_p = 1,2, \ldots, n_p$) neo-fuzzy neuron of p-th ($p = 1,2, \ldots, s$) layer of the neo-fuzzy network contains n_{p-1} of nonlinear synapses $NS_{i_p i_{p-1}}^{[p]}$ where each of them contains h membership functions $\mu_{i_p i_{p-1} l}^{[p]}$ ($l = 1,2, \ldots, h$) and the same number of synaptic weights $w_{i_p i_{p-1} l}^{[p]}$, which are being tuned during the process of learning. Thereby, this architecture has $\sum_{p=1}^{s} n_p n_{p-1}$ nonlinear synapses and $h\sum_{p=1}^{s} n_p n_{p-1}$ membership functions $\mu_{i_p i_{p-1} l}^{[p]}(o_{i_{p-1}}^{[p-1]})$ and the same number of tuned weight coefficients $w_{i_p i_{p-1} l}^{[p]}$.

The output signal of each nonlinear synapse $NS_{i_p i_{p-1}}^{[p]}$ can be written as

$$f_{i_p i_{p-1}}^{[p]}\left(o_{i_{p-1}}^{[p-1]}\right) = \sum_{l=1}^{h} w_{i_p i_{p-1} l}^{[p]} \mu_{i_p i_{p-1} l}^{[p]}\left(o_{i_{p-1}}^{[p-1]}\right), \qquad (1)$$

and the output signal of the neo-fuzzy neuron in $NFN_{i_p}^{[p]}$ in general can be written as

$$o_{i_p}^{[p]} = \sum_{i_{p-1}=1}^{n_{p-1}} f_{i_p i_{p-1}}^{[p]}\left(o_{i_{p-1}}^{[p-1]}\right) = \\ = \sum_{i_{p-1}=1}^{n_{p-1}} \sum_{l=1}^{h} w_{i_p i_{p-1} l}^{[p]} \mu_{i_p i_{p-1} l}^{[p]}\left(o_{i_{p-1}}^{[p-1]}\right). \quad (2)$$

$$\hat{y}_j = o_{i_s}^{[s]} = \sum_{i_{s-1}=1}^{n_{s-1}} f_{i_s i_{s-1}}^{[s]}\left(o_{i_{s-1}}^{[s-1]}\right) = \\ = \sum_{i_{s-1}=1}^{n_{s-1}} \sum_{l=1}^{h} w_{i_s i_{s-1} l}^{[s]} \mu_{i_s i_{s-1} l}^{[s]}\left(o_{i_{s-1}}^{[s-1]}\right). \quad (3)$$

Then, by adding to consideration vectors of synaptic weights and membership functions

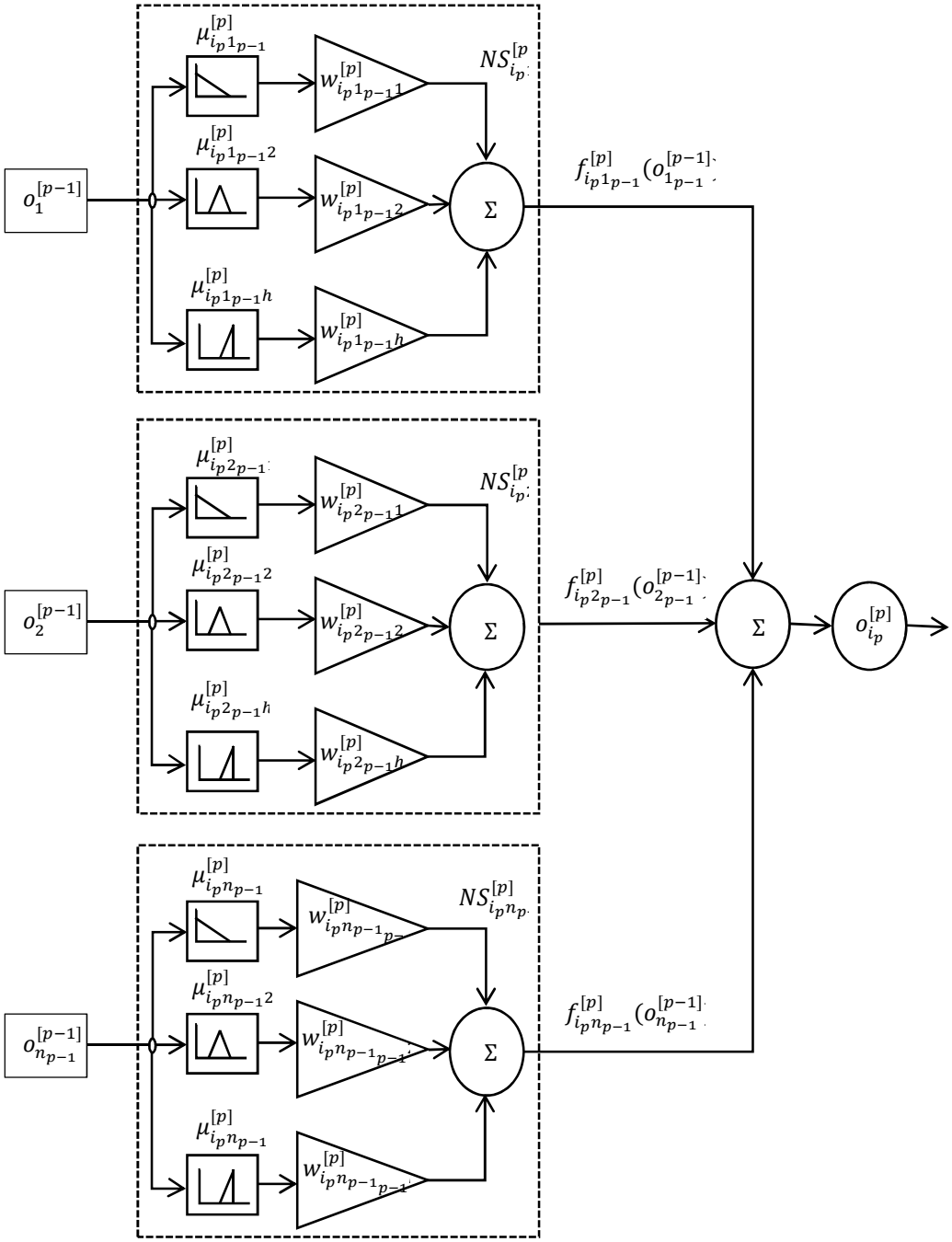


Fig. 1. i-th neo-fuzzy neuron of the p-th layer in a neo-fuzzy neural network

For the output layer of the network signal (2) can be written as

$$w_{i_p i_{p-1}}^{[p]} = (w_{i_p i_{p-1} 1}^{[p]}, \ldots, w_{i_p i_{p-1} l}^{[p]}, \ldots w_{i_p i_{p-1} h}^{[p]})^T,$$

$$\mu_{i_p i_{p-1}}^{[p]}\left(o_{i_{p-1}}^{[p-1]}\right) = \\ = (\mu_{i_p i_{p-1} 1}^{[p]}\left(o_{i_{p-1}}^{[p-1]}\right), \ldots \mu_{i_p i_{p-1} l}^{[p]}\left(o_{i_{p-1}}^{[p-1]}\right), \ldots \mu_{i_p i_{p-1} h}^{[p]}\left(o_{i_{p-1}}^{[p-1]}\right))^T$$

of dimension $(h \times 1)$, and also

$$w_{i_p}^{[p]} = (w_{i_p 1}^{[p]}, \dots, w_{i_p i_{p-1}}^{[p]}, \dots w_{i_p n_{p-1}}^{[p]})^T$$

and

$$\mu_{i_p}^{[p]}\left(o_{i_{p-1}}^{[p-1]}\right) = \\ = (\mu_{i_p 1}^{[p]}\left(o_{i_{p-1}}^{[p-1]}\right), \dots \mu_{i_p i_{p-1}}^{[p]}\left(o_{i_{p-1}}^{[p-1]}\right), \dots \mu_{i_p n_{p-1}}^{[p]}\left(o_{i_{p-1}}^{[p-1]}\right))^T$$

with dimension $(n_p h \times 1)$, we can write

$$f_{i_p i_{p-1}}^{[p]}\left(o_{i_{p-1}}^{[p-1]}\right) = w_{i_p}^{[p]T}\mu_{i_p}^{[p]}\left(o_{i_{p-1}}^{[p-1]}\right)$$

instead of (1),

$$o_{i_p}^{[p]} = w_{i_p}^{[p]T}\mu_{i_p}^{[p]}\left(o_{i_{p-1}}^{[p-1]}\right)$$

instead of (2) and

$$\hat{y}_j = o_{i_s}^{[s]} = w_{i_s}^{[s]T}\mu_{i_s}^{[s]}\left(o_{i_{s-1}}^{[s-1]}\right)$$

instead of (3).

As a membership functions of nonlinear synapses $NS_{i_p i_{p-1}}^{[p]}$ authors of the neo-fuzzy neuron [10-12] used traditional triangle functions which satisfy the requirements of a Ruspini partition:

$$\mu_{i_p i_{p-1} l}^{[p]}\left(o_{i_{p-1}}^{[p-1]}\right) = \\ = \begin{cases} \dfrac{o_{i_{p-1}}^{[p-1]} - \mathrm{c}_{i_p i_{p-1} l-1}^{[p]}}{\mathrm{c}_{i_p i_{p-1} l}^{[p]} - \mathrm{c}_{i_p i_{p-1} l-1}^{[p]}}, if\ o_{i_{p-1}}^{[p-1]} \in \left[\mathrm{c}_{i_p i_{p-1} l-1}^{[p]}, \mathrm{c}_{i_p i_{p-1} l}^{[p]}\right), \\ \dfrac{\mathrm{c}_{i_p i_{p-1} l+1}^{[p]} - o_{i_{p-1}}^{[p-1]}}{\mathrm{c}_{i_p i_{p-1} l+1}^{[p]} - \mathrm{c}_{i_p i_{p-1} l}^{[p]}}, if\ o_{i_{p-1}}^{[p-1]} \in \left[\mathrm{c}_{i_p i_{p-1} l}^{[p]}, \mathrm{c}_{i_p i_{p-1} l+1}^{[p]}\right), \\ 0, \quad else, \end{cases} \tag{4}$$

where $\mathrm{c}_{i_p i_{p-1} l}^{[p]}, l = 1,2, \dots, h$ - are the centers of triangular membership functions. In the case when all nonlinear synapses of the network $NS_{i_p i_{p-1}}^{[p]}$ have the same number h of centers which are distributed uniformly through all axes of input signals, condition (4) can be written in the more convenient way as:

$$\mu_{i_p i_{p-1} l}^{[p]}\left(o_{i_{p-1}}^{[p-1]}\right) = \\ = \begin{cases} \Delta_c^{-1}\left(o_{i_{p-1}}^{[p-1]} - \mathrm{c}_{i_p i_{p-1} l-1}^{[p]}\right), if\ o_{i_{p-1}}^{[p-1]} \in \left[\mathrm{c}_{i_p i_{p-1} l-1}^{[p]}, \mathrm{c}_{i_p i_{p-1} l}^{[p]}\right), \\ \Delta_c^{-1}\left(\mathrm{c}_{i_p i_{p-1} l+1}^{[p]} - o_{i_{p-1}}^{[p-1]}\right), if\ o_{i_{p-1}}^{[p-1]} \in \left[\mathrm{c}_{i_p i_{p-1} l}^{[p]}, \mathrm{c}_{i_p i_{p-1} l+1}^{[p]}\right), \\ 0, \quad else, \end{cases} \tag{5}$$

and at the same time:

$$\mu_{i_p i_{p-1} l-1}^{[p]}\left(o_{i_{p-1}}^{[p-1]}\right) + \mu_{i_p i_{p-1} l}^{[p]}\left(o_{i_{p-1}}^{[p-1]}\right) = \\ = \mu_{i_p i_{p-1} l}^{[p]}\left(o_{i_{p-1}}^{[p-1]}\right) + \mu_{i_p i_{p-1} l+1}^{[p]}\left(o_{i_{p-1}}^{[p-1]}\right) = 1. \tag{6}$$

Conditions (4)-(6) mean that in every moment k in every nonlinear synapse only two neighbored membership function can be fired. It caused that at the same moment only two neighbored weight coefficients are tuned, so on each step of learning only $2\sum_{p=1}^{s} n_{p-1} n_p$ of weight coefficients are being adjustable instead of $h\sum_{p=1}^{s} n_{p-1} n_p$.

Whereas every nonlinear synapse of the network in general implements F-transform [17] it allows to approximate any limited one-dimensional function with regards to big enough h. Increasing the number of membership functions doesn't add complexity to the learning of the network.

As follows, every NFN is a hybrid of the Wang–Mendel's system and F-transform, which provides us with high approximate capabilities of the network in general.

III. Learning of The Deep Neo-Fuzzy Neural Network

Learning of the multilayer neo-fuzzy network is based on error backpropagation algorithm.

The process of learning of the neo-fuzzy network is built on the search of synaptic weight coefficients $w_{i_p i_{p-1} l}^{[p]}, p = \\ = 1, 2, \dots, s; l = 1,2, \dots, h;\ i_p = 1,2, \dots, n_p;\ i_{p-1} = \\ = 1,2, \dots, n_{p-1}$ by minimization of the chosen loss function.

As the learning criterion we use standard squared function

$$E(k) = \frac{1}{2}\sum_{i_s=1}^{n_s} e_{i_s}^2(k) = \frac{1}{2}\sum_{j=1}^{m} e_j^2(k) \tag{7}$$

where

$e_j^2(k) = (y_j(k) - w_{i_s}^{[s]}(k-1)\mu_{i_s}^{[s]}(o_{i_{s-1}}^{[s-1]}(k)))^2$,

$y_j(k)$ – is an external learning signal.

Gradient minimization of (7) for each synaptic weight coefficient of the output layer $w_{i_s i_{s-1} l}^{[s]}$ can be written as recurrent procedure

$$w_{i_s i_{s-1} l}^{[s]}(k) = w_{i_s i_{s-1} l}^{[s]}(k-1) - \eta(k)\frac{\partial e_{i_s}^2(k)}{\partial w_{i_s i_{s-1} l}^{[s]}} \\ = \\ = w_{i_s i_{s-1} l}^{[s]}(k-1) - \eta(k)\mu_{i_s}^{[s]}(O_{i_{s-1}}^{[s-1]}(k)) \tag{8}$$

(where $\eta(k)$ is a learning rate parameter), or in vector form representation as

$$w_{i_s}^{[s]}(k) = w_{i_s}^{[s]}(k-1) - \\ -\eta(k)(y_j(k) - w_{i_s}^{[s]}(k-1)\mu_{i_s}^{[s]}\left(o_{i_{s-1}}^{[s-1]}(k)\right). \tag{9}$$

Optimal Kaczmarz-Widrow-Hoff algorithm and stochastic approximation as a hybrid can be chosen as a procedure for defining $\eta(k)$ [18, 19].

$$\begin{cases} w_{i_s}^{[s]}(k) = w_{i_s}^{[s]}(k-1) - r_{i_s}^{-1}(k)e_{i_s}^2(k)\mu_{i_s}^{[s]}\left(o_{i_{s-1}}^{[s-1]}(k)\right), \\ r_{i_s}^{-1}(k) = \alpha r_{i_s}(k-1) + \left\|\mu_{i_s}^{[s]}(o_{i_{s-1}}^{[s-1]}(k))\right\|^2 \end{cases}$$

where $0 \le \alpha \le 1$ – is forgetting factor.

Configuring of weights inside hidden layers is implemented with help of backpropagation algorithm. To achieve this, procedure (8) can be used in the form

$$w_{i_{s-1} i_{s-2} l}^{[s-1]}(k) = w_{i_{s-1} i_{s-2} l}^{[s-1]}(k-1) \\ - \eta(k)\frac{\partial E(k)}{\partial w_{i_{s-1} i_{s-2} l}^{[s-1]}} = \\ = w_{i_{s-1} i_{s-2} l}^{[s-1]}(k-1) - \\ -\eta(k)\sum_{i_s=1}^{n_s}\frac{\partial e_{i_s}^2(k)}{\partial O_{i_s}^{[s]}}\frac{\partial o_{i_s}^{[s]}(k)}{\partial o_{i_{s-1}}^{[s-1]}(k)}\frac{\partial o_{i_{s-1}}^{[s-1]}(k)}{\partial w_{i_{s-1} i_{s-2} l}^{[s]}}. \tag{10}$$

Because

$$\frac{\partial o_{i_s}^{[s]}}{\partial o_{i_{s-1}}^{[s-1]}} = \frac{\partial f_{i_s i_{s-1}}^{[s]}(o_{i_{s-1}}^{[s-1]})}{\partial o_{i_{s-1}}^{[s-1]}} = \sum_{l=1}^{h} w_{i_s i_{s-1} l}^{[s]} \frac{\partial \mu_{i_s i_{s-1} l}^{[s]}(o_{i_{s-1}}^{[s-1]})}{\partial o_{i_{s-1}}^{[s-1]}}$$

where

$$\frac{\partial \mu_{i_s i_{s-1} l}^{[s]}(o_{i_{s-1}}^{[s-1]})}{\partial o_{i_{s-1}}^{[s-1]}} =$$

$$= \begin{cases} \Delta_c^{-1}, & if\ o_{i_{p-1}}^{[p-1]} \in \left[c_{i_p i_{p-1} l-1}^{[p]}, c_{i_p i_{p-1} l}^{[p]}\right), \\ -\Delta_c^{-1}, & if\ o_{i_{p-1}}^{[p-1]} \in \left[c_{i_p i_{p-1} l}^{[p]}, c_{i_p i_{p-1} l+1}^{[p]}\right), \\ 0, & else, \end{cases} \quad (11)$$

algorithm (10) can be finalized as

$$w_{i_{s-1} i_{s-2} l}^{[s-1]}(k) = w_{i_{s-1} i_{s-2} l}^{[s-1]}(k-1) - \\ -\eta(k) \sum_{i_s=1}^{n_s} e_{i_s}(k) \sum_{l=1}^{h} w_{i_s i_{s-1} l}^{[s]}(k) * \\ * \frac{\partial \mu_{i_s i_{s-1} l}^{[s]}\left(o_{i_{s-1}}^{[s-1]}\right)}{\partial o_{i_{s-1}}^{[s-1]}} \mu_{i_{s-1} i_{s-2} l}^{[s-1]}\left(o_{i_{s-2}}^{[s-2]}(k)\right). \quad (12)$$

By analogy of (12) can be written algorithm of tuning of p-th hidden layer with considering the error of p-th layer in the next notation:

$$\frac{\partial E}{\partial o_{i_p}^{[p]}} = \sum_{i_{p+1}=1}^{n_{p+1}} \frac{\partial E}{\partial o_{i_{p+1}}^{[p+1]}} \frac{\partial o_{i_{p+1}}^{[p+1]}}{\partial o_{i_p}^{[p]}}$$

where delta-error is written as

$$\delta_{i_p}^{[p]} = \frac{\partial E}{\partial o_{i_p}^{[p]}} = \sum_{i_{p+1}=1}^{n_{p+1}} \delta_{i_{p+1}}^{[p+1]} \frac{\partial o_{i_{p+1}}^{[p+1]}}{\partial o_{i_p}^{[p]}}. \quad (13)$$

Thereby, for all weights we can write

$$\frac{\partial E(k)}{\partial w_{i_p i_{p-1} l}^{[p]}} = \sum_{i_{p+1}=1}^{n_{p+1}} \delta_{i_{p+1}}^{[p+1]} \frac{\partial o_{i_{p+1}}^{[p+1]}}{\partial o_{i_p}^{[p]}} \frac{\partial o_{i_p}^{[p]}}{\partial w_{i_p i_{p-1} l}^{[p]}}$$

where the delta-error is taken from the next layer and intermediate derivatives are calculated in the form:

$$\frac{\partial o_{i_{p+1}}^{[p+1]}}{\partial o_{i_p}^{[p]}} = \frac{\partial f_{i_{p+1} i_p}^{[p+1]}(o_{i_p}^{[p]})}{\partial o_{i_p}^{[p]}} = \sum_{l=1}^{h} w_{i_{p+1} i_p l}^{[p+1]} \frac{\partial \mu_{i_{p+1} i_p l}^{[p+1]}(o_{i_p}^{[p]})}{\partial o_{i_p}^{[p]}}$$

where the right part is received from (11).

A derivative of the output signal of a neo-fuzzy neuron by a weight coefficient is written as:

$$\frac{\partial o_{i_p}^{[p]}}{\partial w_{i_p i_{p-1} l}^{[p]}} = \frac{f_{i_p i_{p-1}}^{[p]}\left(o_{i_{p-1}}^{[p-1]}\right)}{\partial w_{i_p i_{p-1} l}^{[p]}} = \mu_{i_p i_{p-1} l}^{[p]}\left(o_{i_{p-1}}^{[p-1]}\right)$$

where the final expression

$$w_{i_p i_{p-1} l}^{[p]}(k) = w_{i_p i_{p-1} l}^{[p]}(k-1) - \eta(k) \frac{\partial E(k)}{\partial w_{i_p i_{p-1} l}^{[p]}}$$

can be written as:

$$w_{i_p i_{p-1} l}^{[p]}(k) = w_{i_p i_{p-1} l}^{[p]}(k-1) - \\ -\eta(k) \sum_{i_{p+1}=1}^{n_{p+1}} \delta_{i_{p+1}}^{[p+1]} \frac{\partial o_{i_{p+1}}^{[p+1]}}{\partial o_{i_p}^{[p]}} \mu_{i_p i_{p-1} l}^{[p]}(o_{i_{p-1}}^{[p-1]}) \quad (14)$$

It is not difficult to see that the learning of the neo-fuzzy neural network is simpler from the computational point of view than ANN and DNN based on traditional Rosenblatt's perceptrons because derivatives of the triangle membership functions are constants.

IV. Experimental Studies

The model was tested in different configurations on two well-known datasets – Fashion MNIST and Breast Cancer Wisconsin Data Set (BC dataset). Results were compared with a custom Multilayer perceptron (MLP) and AlexNet architecture. The goal of the experiment is to show that DFNN model is able to do approximation tasks in comparison to well-known models and approaches. A comparison of basic MLP and DFNN are described in Table I. Notation "number of layers X number of neurons in layer" is used for MLP and for DFNN. For DFNN we used fuzzy continuous space [0, 1] with a size of interval 0.1. It means that each neo-fuzzy neuron has at least 10 weights. Because of this, we used 100 neurons in one MLP layer to make results more comparable.

On BS dataset DFNN with 6 neo-fuzzy layers got overfitted on the 7th epoch. Accuracy on the test part of the dataset was 93.85% at this point. DFNN model with 3 neo-fuzzy layers got overfitted at 8th epoch and accuracy on the test data was 92.98%. MLP achieved similar results in accuracy and loss curves behavior but the loss value was smaller.

For the Fashion MNIST dataset results were contradictive. Everything's clear with AlexNet results but we had issues with training DFNN on the data with a relatively huge number of features.

TABLE I. Comparison of MLP and DFNN on BC dataset

Network type	**MLP**		**DFNN**	
Architecture	3**X**100	6**X**100	3**X**10	6**X**10
Test accuracy on 7th epoch	92.1	93.85	92.98	93.85
Test loss on 7th epoch	0.163	0.142	0.418	0.412
Train loss on 7th epoch	0.179	0.123	0.42	0.407

TABLE II. Comparison of MLP and DFNN on Fashion MNIST dataset

	Network type	**AlexNET**	**DFNN**
Epoch	Architecture	Default	**5X**10
1	Test accuracy	61.47	10.0
	Test loss	1.09	2.3041
	Train loss	2.21	2.3045
5	Test accuracy	82.6	10.0
	Test loss	0.46	2.3024
	Train loss	0.47	2.3027

100	Test accuracy	**89.4**	**38.88**
	Test loss	0.32	2.0012
	Train loss	0.24	2.0009

These training issues were related to the architectural approach in DFNN. Summing of results between layers forced us to come up with approach for handling to huge and not appropriate numbers for our universe [0, 1]. We handled it and make a trainable model but this issue is the main topic for our future research of DFNN.

V. CONCLUSIONS

The architecture and the learning algorithm of the neo-fuzzy neural network are proposed where the main difference from the traditional multilayer neural network with feedforward of information is using of neo-fuzzy neurons as nodes instead of traditional elementary Rosenblatt's perceptrons.

The proposed neo-fuzzy neural network has high approximation abilities with increased speed of learning of its synaptic weights which is caused by using triangular membership functions as nonlinear neurons synapses. Furthermore, the proposed neural network is quite simple from the implementation point of view.

REFERENCES

[1] Y. Bengio, Y. LeCun, G. Hinton, "Deep Learning", Nature, vol. 521, p.436-444, 2015.

[2] J. Schmidhuber. "Deep learning in neural networks: An overview", Neural Networks, vol. 1, pp. 85-117, 2015.

[3] I. Googfellow, Y. Bengio, A. Courville, Deep Learning. MIT Press, 2016.

[4] D. Graupe, Deep Learning Neural Networks: Design and Case Studies-New York: World Scientific, 2016.

[5] A.L. Caterini, D.E. Chang, Deep Neural Networks in a Mathematical Framework – Springer, 2018.

[6] A. Cichocki, R. Unbehauen, Neural Networks for Optimization and Signal Processing, Stuttgart: Teubner, 1993.

[7] K. Hornik, "Approximation capabilities of multilayer feedforward networks" Neural Networks, Vol. 4, Is. 2, pp. 251-257, 1991.

[8] Ch.C. Aggarwal, Neural Networks and Deep Learning. Cham: Springer, 2018.

[9] G. Cybenko "Approximation by superpositions of a sigmoidal function" in Math. Control Signals Systems, 2, pp. 303–314, 1989.

[10] T. Yamakawa, E. Uchino, T. Miki, H. Kusabagi, "A neo fuzzy neuron and its applications to system identification and predictions to system behavior." In Proc. 2nd Int. Conf. on Fuzzy Logic and Neural Networks, pp. 477-483, 1992.

[11] E. Uchino, T. Yamakava, "Neo-fuzzy neuron based new approach to system modeling with application to actual system" in Proc. Sixth Int. Conference on Tools with Artificial Intelligence", New Orlean, LA, USA, pp.564-570, 1994.

[12] T. Miki, T. Yamakawa, "Analog implementation of neo-fuzzy neuron and its on-board learning," in Computational Intelligence and Applications, Piraeus: WSES Press, pp. 144-149, 1999.

[13] V. Kolodyazhniy, Ye. Bodyanskiy, Fuzzy Kolmogorov's network – Lecture Notes in Computer Science. v. 3214, Heidelberg: Springer Verlag, pp.764-771, 2004.

[14] Ye. Bodyanskiy, V. Kolodyazhniy, P. Otto, Neuro-fuzzy Kolmogorov's network for time series prediction and pattern classification Lecture Notes in Artificial Intelligence, v. 3698, Heidelberg: Springer Verlog, pp.191-202, 2005.

[15] Ye. Bodyanskiy, S. Popov, T. Rybalchenko, Multilayer neuro-fuzzy network for short term electric load forecasting – Lecture Notes in Computer Science. v. 5010, Berlin-Heidelberg: Springer Verlag, pp. 339-348, 2008.

[16] 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.

[17] T. Perfilieva, "Fuzzy transforms: Theory and applications". Fuzzy Sets and Systems, v. 157, pp. 993-1023, 2006.

[18] Ye. Bodyanskiy, V. Kolodyazhniy, A. Stephan, "An adaptive learning algorithm for a neuro-fuzzy network" Ed. by B.Reush "Computitional Intelligence. Theory and Application", Berlin-Heidelberg: Ney York: Springer, pp. 68-75, 2001.

[19] P. Otto, Ye. Bodyanskiy, V. Kolodyazhniy. "A new learning algorithm for a forecasting neuro-fuzzy network" Integrated Computer Aided Engineering. Vol. 10(4), pp.399-409, 2003.

IEEE Third International Conference on Data Stream Mining & Processing
August 21-25, 2020, Lviv, Ukraine

Convolutional Neural Network with Multi-Valued Neurons

Igor Aizenberg
Manhattan College
Riverdale, NY, USA
igor.aizenberg@manhattan.edu

Alexander Vasko
Uzhhorod National University
Uzhhorod Ukraine
alexander.a.v.asko@gmail.com

***Abstract* — In this paper, a convolutional neural network (CNN) based on multi-valued neurons (MVNs) with complex-valued weights is presented. Convolutional neural networks are known as one of the best tools for solving such problems as image and speech recognition. The vast majority of researches and users so far employ a classical CNN, which operates with real-valued inputs/outputs and is based on classical neurons with a sigmoidal activation function. Recently complex-valued CNNs were introduced, but their neurons employ an activation function, which is a complex-valued generalization of a sigmoidal one. While such complex-valued neurons are more flexible when they learn and make it possible to process complex-valued data, their generalization capability is basically not higher than the one of real-valued neurons. At the same time there exists a complex-valued neuron (a multi-valued neuron – MVN) with a phase dependent activation function whose functionality is higher than the one of neurons with a sigmoidal activation function. It is also known that a multilayer neural network with multi-valued neurons (MLMVN) outperforms a classical multilayer peceptron in terms of speed of learning and generalization capability. Hence our goal was to develop a convolutional neural network based on multi-valued neurons (CNNMVN).**

We consider in detail a learning algorithm for CNNMVN, which is derivative free as well as the one for MLMVN. We also suggested the max pool operation for complex numbers. These results are illustrated by simulations showing that CNNMVN with even a minimal convolutional topology is able to solve image recognition problems with pretty high accuracy.

Keywords* — *Convolutional Neural Networks, Convolutional Neural Networks with Multi-Valued Neurons, CNNMVN, Complex-Valued Neural Networks, Multi-Valued Neuron, Multilayer Neural Network with Multi-Valued Neurons, MLMVN

I. Introduction

The ideas, which led to the creation of convolutional neural networks, were suggested in [1], [2]. A classical convolutional neural network as it is known today was introduced in the seminal paper [3]. Since then CNNs have shown wonderful results, first of all in solving image recognition [3] and speech recognition [4] problems.

CNN can actually be considered as the next generation of a feedforward neural network. While a classical feedforward neural network (a multilayer perceptron - MLP) and a classical CNN both are based on real-valued neurons with a sigmoid activation function, there exists another family of feedforward neural networks – complex-valued neural networks (CVNNs). CVNNs have a number of advantages over their real-valued counterparts (they learn faster, some of them generalize better and they are able to work with complex-valued data). These properties of CVNNs are considered, e.g. in [5], [6]. In CVNNs, mostly two types of neurons are used. There is a complex-valued neuron with a complex hyperbolic tangent activation function. This neuron, while learns faster than a neuron with a sigmoid activation function and has in general better generalization capabilities [5], has about the same functionality as a sigmoidal neuron. There is also another complex-valued neuron – a multi-valued neuron whose activation function depends only on the argument (phase) of its weighted sum and whose output is located on the unit circle [6]. This neuron is more functional, thus it may learn even some input/output mappings, which are not linearly separable [7].

A MVN [8] is a generic neuron, which is used in the multilayer feedforward neural network with multi-valued neurons (MLMVN) introduced in [9] and presented in detail in [10]. MLMVN is a feedforward neural network and it outperforms MLP when solving various problems [6], [10].

A number of researches pretty recently discussed and introduced complex-valued CNNs based on the complex-valued sigmoidal neurons [11], [12]. Such CNNs were successfully used for solving image recognition problems [13] and EEG analysis problems. [14].

Thus it becomes very attractive to design a complex-valued CNN with MVN as its generic neuron. It should be expected that such a network may have the same attractive properties, which MLMVN has. In particular, we should expect that a convolutional neural network with multi-valued neurons (CNNMVN) might have a derivative-free learning, learn as quickly as MLMVN and demonstrate good results when solving pattern recognition and particularly image recognition problems.

In this paper, we introduce a convolutional neural network with multi-valued neurons (CNNMVN) and its learning algorithm. We consider here this network as a tool for image recognition, but this doesn't limit its other potential applications. As it was mentioned above, CNNMVN is based on merging two fundamental ideas – the use of convolutional layers in a network and MVN as its basic neuron. As it is known from [3], convolutional layers in a network reflect the structure of human vision. They help to structurize most important information about an object to be classified. The most important contibution of this paper is the development of the convolutional operation for a MVN-based neural network and its derivative free learning consisting of the error backpropagation and weights adjustment based on the error correction learning rule adapted to batch learning.

978-1-7281-3214-3/20/$31.00 ©2020 IEEE

To illustrate how CNNMVN and its learning algorithm work, we will use it to solve two benchmark image recognition problems using the MNIST and Fashion MNIST datasets. It will be shown that this convolutional neural network with its simplest possible topology is able to recognize images with a pretty high accuracy.

II. Convolutional Layer: Fundamentals

To design any convolutional neural network, three components are essential: a convolutional layer, a pooling layer and a learning algorithm. Since we will use MVN as a basic neuron of CNNMVN, weights of all neurons in this network are complex-valued and all neurons produce outputs located on the unit circle employing continuous or discrete MVN activation function.

Let us start from a convolutional layer of CNNMVN.

Let us use the following notations. Let CNNMVN contain P convolutional layers, $F-1$ fully-connected layers, and one output layer. Each convolutional layer contains Q_p multi-valued kernels (MVK) of size $r_p, p=1,\dots,P$. Let S_p be the number of input images in the p^{th} convolutional layer, and $S_1=1$. Let us also assume without loss of generality that CNNMVN, which we are going to design, will be used for solving image recognition problems. Let us also further refer MVK as simply a kernel. The main purpose of a convolutional layer is to extract main features from input data to make it possible to recognize an object. Each convolutional layer consists of kernels (filters), which is a specific type of a neuron. Each filter slide over an image and at every location the element-wise matrix multiplication shall be performed followed by the summation of its results. This process of convolution is shown in Fig. 1(a).

The main peculiarity of kernels is that each of them, unlike a standard MVN, accepts not a single input vector, but several input vectors. One kernel can produce $(n-r+1)\times(m-r+1)$ outputs, were $n\times m$ is the input image size, and r is the kernel size. In general, CNNMVN, as any convolutional network, may contain multiple convolutional layers. Each convolutional layer may consist from many kernels. In such a case, each of them process the input images separately and independently and produce its own resulting convolved image. Therefore, we obtain convolved images, which become inputs to the next layer of CNNMVN (Fig 1(b), steps I-III).

The convolution operation in the first convolutional layer is determined as follows:

$$c_{ij}^{q}=\sum_{k=0}^{r_1-1}\sum_{l=0}^{r_1-1}x_{i+k.j+l}w_{k+1,l+1}^{q}, \tag{1}$$

where c_{ij}^{q} are the ij^{th} convolved pixels of q^{th} kernel, $q=1,\dots,Q_1$, $i=1,\dots,n-r_1+1$, $j=1,\dots,m-r_1+1$, $x_{i.j}$ are the intensities in the corresponding input pixels of an image to be learned, $w_{k+1,l+1}^{q}$ are the kernel weights, r_1 and Q_1 are the kernel size and the number of kernels in the first layer, respectively. To make (1) easier understandable, it can also be presented using the matrix notation as follows

$$\begin{bmatrix} c_{11}^{q} & \cdots & c_{1,m-r_1+1}^{q} \\ \vdots & \ddots & \vdots \\ c_{n-r_1+1,1}^{q} & \cdots & c_{n-r_1+1,m-r_1+1}^{q} \end{bmatrix}= \\ =\begin{bmatrix} x_{11} & \cdots & x_{1m} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{nm} \end{bmatrix}\otimes\begin{bmatrix} w_{11}^{q} & \cdots & w_{1r_1}^{q} \\ \vdots & \ddots & \vdots \\ w_{r_11}^{q} & \cdots & w_{r_1r_1}^{q} \end{bmatrix} \tag{2}$$

where $\otimes$ stands for convolution.

The output of the first convolutional layer, which is a matrix obtained according to (2) is forwarded to the next layer, which can be a pooling layer or the second convolutional layer. If the latter is the case, then this layer also consists of kernels. In general, there can be multiple convolutional layers in a network, followed by a pooling layer. The convolution in all following convolutional layers is determined as follows

$$d_{ij}^{pqs}=\sum_{k=0}^{r_p-1}\sum_{l=0}^{r_p-1}c_{i+k.j+l}^{s,p-1}w_{k+1,l+1}^{qp}, \tag{3}$$

where d_{ij}^{pqs} specifies the ij^{th} convolved pixel of the s^{th} resulting convolved image ($s=1,\dots,S_p$) created as an output at the preceding (p^{th}) convolutional layer, by its q^{th} kernel, $q=1,\dots,Q_p$. $c_{i+k.j+l}^{sp-1}$ are the "intensities" in the pixels of s^{th} convolved image, $w_{k+1,l+1}^{qp}$ are the weights of q^{th} kernel in p^{th} layer. As a result, we obtain S_pQ_p output images in p^{th} layer. To make (3) clearer, it can be presented using the matrix notation as follows

$$\begin{bmatrix} d_{11}^{pqs} & \cdots & d_{1,m-k+1}^{pqs} \\ \vdots & \ddots & \vdots \\ d_{n-r_p+1,1}^{pqs} & \cdots & d_{n-r_p+1,m-r_p+1}^{pqs} \end{bmatrix}= \\ =\begin{bmatrix} c_{11}^{s,p-1} & \cdots & c_{1m}^{s,p-1} \\ \vdots & \ddots & \vdots \\ c_{n1}^{s,p-1} & \cdots & c_{nm}^{s,p-1} \end{bmatrix}\otimes\begin{bmatrix} w_{11}^{qp} & \cdots & w_{1r}^{qp} \\ \vdots & \ddots & \vdots \\ w_{r_p1}^{qp} & \cdots & w_{r_pr_p}^{qp} \end{bmatrix} \tag{4}$$

After the convolution determined by (2)-(4) in any convolutional layer is complete and before proceeding it to the next layer, the continuous MVN activation function shall be applied to all elements of the corresponding resulting convolved image, which are analogs of weighted sums of a regular neuron, to project the layer`s output onto the unit circle.

All kernels in the corresponding convolutional layers of CNNMVN employ a continuous MVN activation function [9] projecting a complex number y onto the unit circle as follows

$$v=e^{i\arg(y)}=y/|y|,$$

where i is an imaginary unit.

Since every kernel creates multiple outputs and each convolutional layer may contain a number of kernels, the output of the last convolutional layer must be flattened in a single vector before forwarding it to a fully-connected layer. To perform flattening, the resulting convolved images shall be transformed into a single vector as it is shown in (Fig 1(b), steps III-IV).

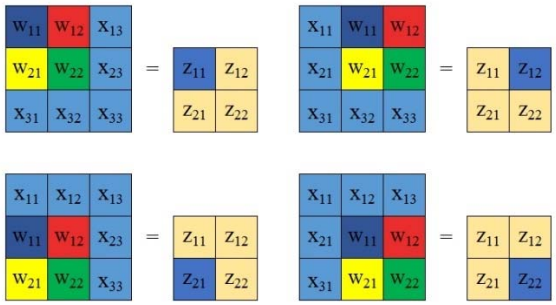


Fig. 1 (a) 2x2 Kernel slide over the 3x3 input image with stride 1. We obtain 2x2 image from 3x3 image.

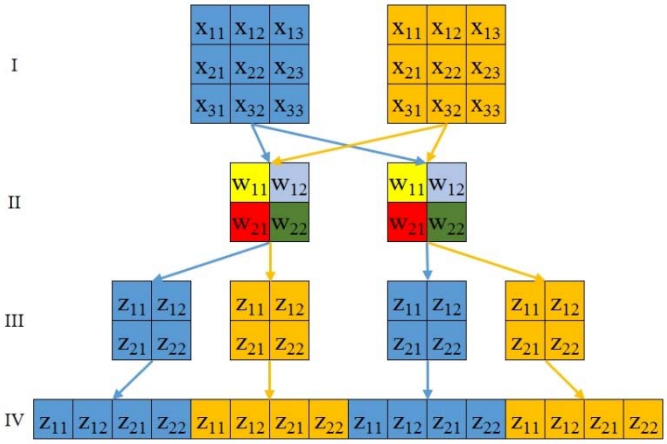


Fig. 1 (b) The last convolutional layer. At steps I-III, each kernel has 2 input images and 2 output convolved images. At steps III-IV, flattening is performed.

III. CNNMVN Learning Algorithm

A. Error Backpropagation

Since CNNMVN as any CNN has a traditional feedforward topology, its learning algorithm in general should rely on the same ideas as the one of MLMVN. The MLMVN learning algorithm is derivative free and is based on the generalization of the error-correction learning rule for a single MVN. This algorithm was presented in detail in [6], [10]. The MLMVN learning algorithm is derived from the consideration that the global error of the network depends on the local errors of all neurons and therefore must be shared among all neurons, since all of them contribute to this error by their local errors. In [15], a soft margins technique was added to the MLMVN learning algorithm. This technique improves MLMVN generalization capability when solving classification and pattern recognition problems. In [16], the batch learning algorithm was introduced for MLMVN. This algorithm is based on the correction of the network weights based on the errors found for the entire learning set and is reduced to solving of an overdetermined system of linear algebraic equations. In [17], this algorithm was improved and generalized for MLMVN with an arbitrary amount of output neurons.

Let us design the CNNMVN learning algorithm. Being based on the same general fundamentals as the MLMVN learning algorithm, this algorithm will be at the same time different from it.

Let us have CNNMVN with $M-1$ fully connected layers and an output layer. To calculate the errors of neurons in these layers, we should use the error backpropagation rule which was proposed in [9]. Before calculating errors in convolutional layers, we should mention that each kernel has multiple inputs and multiple outputs at a time accordingly. This means that to correct the weights of any kernel, it is necessary to calculate first the errors for all the kernel`s outputs. Also, since flattening is used right before proceeding to fully-connected layer, the order in which all convolved images were placed in a flattened vector (see Fig. 1(b)) shall be fixed.

To calculate the errors for all the kernel's outputs in the last (p^{th}) convolutional layer, the same formula, which is used for hidden neurons in fully connected layers shall be adapted as follows

$$\delta_{ij}^{pqs} = \frac{1}{Q_{p-1}} \sum_{k=1}^{F_1} \delta_{k,1} (w_{ijqs}^{k,1})^{-1} \qquad (5)$$

where $w_{ijqs}^{k,1}$ specifies the weight in the first fully-connected layer corresponding to ij^{th} convolved pixel produced by the q^{th} kernel from the s^{th} input image in a flattened vector. F_1 is the number of neurons in the first fully connected layer, and $\delta_{k,1}$ is the error of the k^{th} neuron in first fully-connected layer. Thus the errors for all kernels' outputs of the last convolutional layer shall be calculated according to (5).

The errors of kernels in all $p-1$ convolutional layers shall be calculated by using the backward convolution operation, which is shown in Fig 2. To perform backward convolution in the $p-1^{th}$ layer, we should flip kernels by 180° applying padding to the errors in p^{th} layer of size $Q_p - 1$, where Q_p is the kernel size in p^{th} layer, and slide over like it is a regular direct convolution.

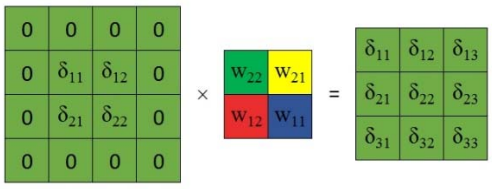


Fig. 2 Backward convolution. On the left we see errors of the kernel`s output with padding. To obtain errors of input image we should convolve errors of output image by flipped to 180° kernel.

The errors at the s^{th} output of the q^{th} kernel in the rest of the convolutional layers are

$$\delta_{ij}^{p-1,qs} = \frac{1}{Q_{p-2}} \sum_{k=0}^{r_p-1} \sum_{l=0}^{r_p-1} \sum_{m=1}^{Q_p} \delta_{i+k.j+l}^{mps} w_{r_p-k,r_p-l}^{mp}, \qquad (6)$$

where $\delta_{ij}^{p-1,qs}$ are the errors at the ij^{th} output element of the q^{th} kernel in the s^{th} convolved image, $q = 1, \dots, Q_{p-1}$, $s = 1, \dots, S_p$ w_{r_p-k,r_p-l}^{mp}, r_p are the m^{th} kernel`s weights and the

kernel size in p^{th} layer accordingly, and Q_{p-2} is the number of kernels in $p-2^{th}$ layer, $Q_1 = r_1^2$.

Finding errors according to (6) completes the process of the error backpropagation and calculation of errors of all neurons and kernels in CNNMVN.

B. Correction of the Weights

After the errors of all neurons in CNNMVN are calculated, we can correct the weights. To correct the weights of kernels in all convolutional layers, it is possible to use the same idea of the batch LLS learning algorithm, which was developed for MLMVN in [17]. The main idea of the batch algorithm is that the weights can be adjusted not for each learning sample separately and sequentially, but for a whole learning set as a batch at a time. The larger is a batch, the more suitable is this method. Since each kernel gets and produces multiple output matrices (vectors) at a time, this method should be very appropriate to correct its errors.

Let us for simplicity, but without loss of generality have a single $n \times m$ input image and a kernel of the size r. Let M be the number of input vectors of this kernel, thus $M = (n-Q+1) \times (m-Q+1)$. We can rewrite the convolution operation using the matrix-vector notation as follows

$$\begin{bmatrix} x_{1,1},\dots,x_{1,r} & \cdots & x_{r,1},\dots,x_{r,r} \\ \vdots & \ddots & \vdots \\ x_{1,j},\dots,x_{1,j+r-1,} & \cdots & x_{r,j},\dots,x_{r,j+r-1} \\ \vdots & \ddots & \vdots \\ x_{i,1},\dots,x_{i,r} & \cdots & x_{i+r-1,1},\dots,x_{i+r-1.r} \\ \vdots & \ddots & \vdots \\ x_{ij},\dots,x_{i,j+r-1,} & \cdots & x_{i+r-1,j},\dots,x_{i+r-1,j+r-1} \end{bmatrix} \begin{bmatrix} w_{1,1} \\ \vdots \\ w_{1,r} \\ \vdots \\ w_{r,1} \\ \vdots \\ w_{r,r} \end{bmatrix} = \tag{7}$$

$$= \begin{bmatrix} z_{1,1} \\ \vdots \\ z_{1,m-r+1} \\ \vdots \\ z_{n-r+1,1} \\ \vdots \\ z_{n-r+1,m-r+1} \end{bmatrix}$$

where $i = n-r+1, j = m-r+1$. After the errors for all resulting values z_{ij} are calculated, we can correct the weights.

To adjust the weights, the MVN error-correction learning rule [6] shall be used

$$W_{r+1} = W_r + \frac{C_r}{n+1}\delta\bar{X}, \tag{8}$$

where W_r and W_{r+1} are the weighting vectors before and after correction, respectively, $\bar{X}$ is the neuron`s input vector with complex-conjugated elements, n is the number of neuron inputs, r is the number of the learning step, C_r is the learning rate constant, which should be complex-valued in general, but was taken equal to 1 in all known applications. The error-correction rule (8) can be easily adapted to the kernel from a convolutional layer as follows

$$W_{r+1} = W_r + \frac{C_r}{n}\delta\bar{X}, \tag{9}$$

Since a kernel has no bias w_0, the error shall be distributed among all weights being divided by n.

Let us consider now the j^{th} convolved pixel. Let D_j be its desired output and δ_j is its error. In such a case, its weighted sum is

$$z_j = w_1 x_1^j + \dots + w_n x_n^j, j = 1,\dots,M \tag{10}$$

After the weights are adjusted according to (9), the updated weighted sum is (taking into account (10))

$$\begin{aligned} \tilde{z}_j &= \tilde{w}_1 x_1^j + \dots + \tilde{w}_n x_n^j = \\ &= \left(w_1 + \frac{\delta}{n}\bar{x}_1^j\right)x_1^j + \dots + \left(w_n + \frac{\delta}{n}\bar{x}_n^j\right)x_n^j = \\ &= w_1 x_1^j + \frac{\delta}{n} + \dots + w_n x_n^j + \frac{\delta}{n} = \\ &= w_1 x_1^j + \dots + w_n x_n^j + \delta = z + \delta, \end{aligned} \tag{11}$$

that is exactly what it should be: a corrected weighted sum was adjusted exactly by δ, (the error).

Let us use the following notation

$$\Delta w_i = \frac{\delta}{n}\bar{x}_i^j; i = 1,\dots,n. \tag{12}$$

Then for a desired output we ideally should obtain the following from (11) and (12)

$$\begin{aligned} D_j &= (w_1 + \Delta w_1)x_1^j + \dots + (w_n + \Delta w_n)x_n^j = \\ &= \underbrace{w_1 x_1^j + w_n x_n^j}_{z_j} + \underbrace{\Delta w_1 x_1^j + \Delta w_n x_n^j}_{\delta_j} = \\ &= z_j + \delta_j; j = 1,\dots,M. \end{aligned} \tag{13}$$

It follows directly from (13) that

$$\delta_j = \Delta w_1 x_1^j + \dots + \Delta w_n x_n^j; j = 1,\dots,M. \tag{14}$$

Thus, M kernels` input vectors determine (14), which can be considered as a system of linear algebraic equations in n unknowns $\Delta w_1, \Delta w_2, \dots, \Delta w_k$ over the field of complex numbers. These unknowns are the adjustments determined by (12). System (14) is typically overdetermined because the number of input vectors to the kernel is always larger than the kernel`s size. This means that such a system can pretty easily be solved using complex QR decomposition [18] or singular value decomposition (SVD) [18]. It should also be taken into account that there can be repetitive vectors $x_1,\dots,x_n$ in (14). The corresponding repetitive equations should be removed from the system.

Let a^q be a solution of (14) for the q^{th} kernel, $q = 1,\dots,Q$. w^q, $\tilde{w}^q$ are weighting vectors before and after correction respectively, then

$$\tilde{w}^q = w^q + a^q \tag{15}$$

Thus the kernel weights for all vectorized input matrices are corrected simultaneously as it was just shown. Finally, we should consider more general cases when there is more than one input image. Also, these cases are valid for arbitrary number of convolutional layers.

Let us have S input images and one kernel. This means that we should calculate errors of all S convolved images. In such a case system (15) contains SM equations. This system shall be solved in the same way as described above. After this system is solved, the weights shall be adjusted according to (15).

Let us have S input images and Q kernels in a convolutional layer. The algorithm is the same. It is just necessary to calculate the errors of all convolved images for each kernel and obtain Q different systems of SM equations for each kernel. Then these systems shall be solved for each kernel and the kernel's weighs shall be adjusted accordingly.

Finally, when the process of weights adjustment for convolutional layers is complete, we have to adjust the weights for the next layer(s), which are fully connected.

To adjust the weights of neurons in $F-1$ fully-connected hidden layers and in a single output layer, the MLMVN error-correction learning rule with soft margins [15] should be used. Once all the weights are adjusted for the entire learning set, a learning iteration is complete. The learning process should continue until the learning error satisfies the angular RMSE criterion [15].

The actual learning process should be organized in the incremental manner, using a step by step learning of sub-batches from the learning set (starting the process for each sub-batch from the weights obtained for the previous one). The process shall be considered converged when the testing error (for recognition of those images, which did not participate in the learning process) reached its minimum. The learning goal is in fact to train a network how to generalize on the testing data, not how to "memorize" a learning set.

C. Pooling layer

A pooling layer is usually located directly after the last convolutional layer and the main idea behind it is to reduce the size of an image to be learned while keeping its most specific important features. A pooling layer acts like a convolutional layer – it slides over its 2-D input "image", but at each step it determines and keeps only the most important pixel, which should be selected based on some rule.

There are two popular versions of pooling operations used in CNNs – the "Max" pooling and the "Average" one. The max pooling is reduced to selection of the maximum value in an input matrix, while its other elements remain ignored. The average pooling operation is reduced to averaging the values in an input matrix. This makes it possible to reduce the size of an image to be learned twice or even more, depending on the pool size and stride.

Since CNNMVN is dealing with complex numbers, the average pooling looks natural for this network. The max pooling operation can also be utilized if complex numbers are compared by their arguments (phases) located in the interval $[0,2\pi]$.

To perform the error backpropagation through a pooling layer, it should be adapted to the type of pooling. When the max pooling is used, the errors corresponding only to maxima should be calculated. In such a case (14) should consist only of the equations corresponding to these selected maxima. Other inputs should be ignored.

When the average pooling is used, the error shall be shared among all elements of a local pool. Since average pooling is reduced to finding an arithmetic mean of convolved pixels then to find the corresponding errors, the error of this arithmetic mean should be divided by the number of pixels containing in the pool.

IV. Simulation Results

We designed a CNNMVN software simulator. It is designed in Matlab and employs parallel processing using GPU[1]. This simulator was used to test the CNNMVN's recognition capability. We used popular image benchmark datasets MNIST [19] and Fashion MNIST [20]. Each dataset contains 70,000 samples. Creators of both datasets designated 60,000 samples for learning and 10,000 ones for testing. It is important to note that we were interested in a "pure" experiment. This means that unlike it was in many other works where these image data sets were used, no image preprocessing was performed. All images were taken as they are. This also means that the simplest possible CNNMVN was used, that is a network containing a single convolutional layer, a pooling layer, a single hidden layer and an output layer.

To convert image intensities into complex numbers located on the unit circle and use them as the CNNMVN inputs, the following transformation was applied

$$\varphi_{ij}=\frac{(x_{ij}-\min\{X\})\frac{3\pi}{2}}{\max\{X\}-\min\{X\}}, z_{ij}=e^{i\varphi_{ij}},$$

where $\max\{X\},\min\{X\}$ are the maximum and minimum intensities of input image respectively, x_{ij} is the input image pixel to be transformed into phase of the complex number.

As we mentioned above, we used here the simplest possible CNNMVN. Thus our network had the topology $I\rightarrow C\rightarrow P\rightarrow H\rightarrow O$ where I is a 28x28 input image, C is the number of 3x3 kernels in a single convolutional layer, P is the max pooling layer with the 2x2 size and stride 1, H is the number of neurons in a single hidden layer, and, O is the number of neurons in the output layer. In our case, $O=10$ because this is the number of classes in both image data sets to be recognized. This means that in the output layer, the neurons performed the binary "1 vs. all" recognition using the "winner takes it all" principle. A neuron-winner was determined by the minimal deviation of its complex-valued weighted sum from the complex-valued desired output in terms of the angular distance as it was suggested in [15].

We limited training to 15 epochs. To achieve better accuracy, we used the dropout technique [21] with the dropout rate 25%. This approach allows to avoid neural network`s overfitting due to disabling its neurons with probability t at every learning iteration. Also, as we employed soft margins [15] in the CNNMVN learning rule, the angular root mean square deviation of $\pi/2$ from the desired output was allowed. The experimental results are shown in Table 1.

[1] A software simulator and data used in this paper will be available to the research community right after publication of the paper at https://www.freewebs.com/igora/Downloads.htm

TABLE I. CLASSIFICATION RESULTS FOR MNIST AND FASHION MNIST

Dataset	Number of kernels	Number of hidden neurons	Accuracy %
MNIST	7	300	90.24
MNIST	10	300	90.03
Fashion MNIST	7	300	91.38
Fashion MNIST	10	300	90.41

It is important to mention that this stable accuracy shown by the simplest possible CNNMVN shall not be compared to the almost 100% accuracy shown, for example, for the MNIST data set by CNNs containing multiple convolutional layers or other deep learning techniques, which employed preprocessing of the data.

Moreover, the level of accuracy, which was achieved using the simplest possible convolutional neural network with multi-valued neurons containing a single convolutional layer with just a few (7 to 10) kernels and without any data preprocessing shall definitely be considered high.

V. CONCLUSION AND FUTURE WORK

Our goal was to design a convolutional neural network based on multi-valued neurons and develop its learning rule. CNNMVN was designed and fundamentals of its organization and learning algorithm were presented. Simulations using the simplest CNNMVN with a single convolutional layer show that this network may have a high potential in solving image recognition problems.

Our future work will be focused on reaching the next goal. This goal is to further develop CNNMVN by using multiple convolutional layers. This will make it possible to drastically improve the CNNMVN generalization capability. Then it will be possible to compare it, for example to CNNs employing multiple convolutional layers and presented in [3], [21], [22].

Also, our future work will include full adaptation for CNNMVN of the batch learning technique earlier used for MLMVN. It should be expected that this approach will make it possible to speed up the learning process by reducing the number of its learning iterations.

VI. ACKWNOLEDGEMENT

All software simulations were performed in the Kakos Center for Scientific Computing at the School of Science of Manhattan College.

REFERENCES

[1] Y. LeCun, F.J. Huang, and L. Bottou. "Learning methods for generic object recognition with invariance to pose and lighting", *Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (CVPR 2004), vol 2, pp. II–97-104, 2004.

[2] K. Jarrett, K. Kavukcuoglu, M. A. Ranzato, and Y. LeCun. "What is the best multi-stage architecture for object recognition?", Proceedings of the 2009 IEEE International Conference on Computer Vision, pp. 2146–2153, 2009.

[3] A. Krizhevsky, I. Sutskever,. and G.E. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks", in *Advances in Neural Information Processing Systems (Proceedings of the 25th Conference on Neural Information Processing Systems)*, F. Pereira and C. J. C. Burges and L. Bottou and K. Q. Weinberger – Eds., Curran Associates, pp. 1097-1105, 2012.

[4] O. Abdel-Hamid, A.-R. Mohamed, H. Jiang, L. Deng, G. Penn, and D. Yu, "Convolutional neural networks for speech recognition," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 22, no. 10, pp. 1533–1545, Oct. 2014.

[5] A. Hirose, *Complex-Valued Neural Networks, 2nd Edn.*, Springer, Berlin, Heidelberg, 2012.

[6] I. Aizenberg, I., *Complex-Valued Neural Networks with Multi-Valued Neurons*. Berlin: Springer-Verlag Publishers, 2011.

[7] I. Aizenberg "A Periodic Activation Function and a Modified Learning Algorithm for a Multi-Valued Neuron", *IEEE Transactions on Neural Networks*, vol. 21, No 12, pp. 1939-1949, December 2010.

[8] 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,.

[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, issue 2, pp. 169-183, Jan. 2007.

[11] J. Bruna, S. Chintala, Y. LeCun, S. Piantino, A. Szlam, M. Tygert, A mathematical motivation for complex-valued convolutional networks, published in ArXiv, available online at https://arxiv.org/abs/1503.03438

[12] N. Kuberman, On Complex Valued Convolutional Neural Networks, published in ArXiv, available online at https://arxiv.org/pdf/1602.09046.pdf, 2016.

[13] Z. Zhang; H. Wang, F.Xu; and Y. Jin, In, Complex-Valued Convolutional Neural Network and Its Application in Polarimetric SAR Image Classification, IEEE Transactions on Geoscience and Remote Sensing, vol. 55, No 12, pp. 7177-7188, Dec, 2017.

[14] J. Zhang and Y. Wu, Complex-value d unsupervise d convolutional neural networks for sleep stage classification, *Computer Methods and Programs in Biomedicine*, vol. 164, pp. 181-191, October 2018.

[15] I. Aizenberg, "MLMVN with Soft Margins Learning", *IEEE Transactions on Neural Networks and Learning Systems*, vol. 25, No 9, pp. 1632-1644, September 2014.

[16] I. Aizenberg, A. Luchetta, and S. Manetti, "A modified Learning Algorithms for the Multilayer Neural Network with Multi-Valued Neurons based on the Complex QR Decomposition", *Soft Computing*, vol. 16, pp. 563-575, April 2012.

[17] E. Aizenberg and I. Aizenberg, "Batch LLS-based Learning Algorithm for MLMVN with Soft Margins", *Proceedings of the 2014 IEEE Symposium Series of Computational Intelligence (SSCI-2014)*, pp. 48-55, December, 2014.

[18] G. H Golub and C. F. Van Loan, *Matrix Computations* (3rd ed.), Johns Hopkins University Press, Baltimore, 1996.

[19] Y. LeCun, C. Cortes, and C.J.C. Burges, The MNIST Databaase of handwritten digits, available online at http://yann.lecun.com/exdb/mnist/

[20] K. Rasul and H.Xiao, Fashion-MNIST, available online at https://research.zalando.com/welcome/mission/research-projects/fashion-mnist/ and https://github.com/zalandoresearch/fashion-mnist

[21] N. Srivastavanitish, G. Hinton, A Krizhevsky, l. Sutskever, R. Salakhutdinov, Dropout: A Simple Way to Prevent Neural Networks fromOverfitting, *Journal of Machine Learning Research*, vol. 15, pp. 1929-1958, 2014.

[22] Y. LeCun, L. Bottou, Y. Bengio and P. Haffner: Gradient-Based Learning Applied to Document Recognition, *Proceedings of the IEEE*, vol. 86, No 11, pp. 2278-2324, November 1998.

IEEE Third International Conference on Data Stream Mining & Processing
August 21-25, 2020, Lviv, Ukraine

MLMVN in Speckle Noise Filtering

Igor Aizenberg
Manhattan College
Riverdale, NY, USA
igor.aizenberg@manhattan.edu

Olivia Keohane
Manhattan College
Riverdale, NY, USA
okeohane01@manhattan.edu

Alejandro Lara
Manhattan College
Riverdale, NY, USA
alara01@manhattan.edu

***Abstract*— In this paper, we use the multilayer neural network with multi-valued neurons (MLMVN) as an intelligent tool for speckle noise filtering. MLMVN is a complex-valued feedforward neural network, which was successfully used for solving various problems including classification, prediction and additive noise filtering. Here, we apply MLMVN containing a single hidden layer for speckle noise filtering by processing overlapping patches taken from a noisy image. A resulting image is obtained by averaging the resulting intensities over all overlapping pixels. To train MLMVN, a learning set created of randomly selected patches from various images is used. A network is trained to transform a noisy patch into a clear one. To make a learning process more efficient, an incremental approach is used. A learning set is divided into batches, and a network is trained by sequentially going over all of them. In terms of PSNR, this approach significantly outperforms Lee filter – traditionally used for speckle noise filtering, and the BM3D filter – commonly recognized as one of the best nonlinear filters ever.**

Keywords*— *Complex-Valued Neural Networks, Multi-Valued Neuron, Multilayer Neural Network with Multi-Valued Neurons, MLMVN, Intelligent Filtering, Speckle Noise

I. INTRODUCTION

In this paper, we use the multilayer neural network with multi-valued neurons (MLMVN) as an intelligent tool for speckle noise filtering. MLMVN is a complex-valued feedforward neural network (CVNN). CVNNs whose observation is given, for example in [1], [2] are employed for solving a number of applied problems, such as: landmine detection [3], forecasting of wind profiles [4], medical image analysis [5], prediction of oil production from a well [6], analysis of signals in EEG-based brain-computer interfaces [7], intelligent filtering of additive noise from images [8].

MLMVN was introduced in [9]. Its derivative-free backpropagation learning was considered in [10]. MLMVN is a feedforward neural network whose generic element is the multi-valued neuron (MVN). Theoretical foundations of MVN, MLMVN, and their learning – which is based on the generalized error-correction rule, are considered in [2].

While MLMVN shown its high efficiency in solving many practical problems, the most important for us with regard to this work is its application in filtering of additive Gaussian noise from images [8]. There, MLMVN was used as an intelligent image filter. We will use it here in the same capacity, but for filtering multiplicative noise.

The most efficient learning algorithm for MLMVN, which was designed in [11] and then modified and adapted for a network with multiple output neurons in [12], employs a batch learning. The latter version of this algorithm will be used here.

Speckle noise filtering is a pressing problem in image processing. Since speckle noise is a multiplicative noise, linear filters are not capable to remove or even reduce it because it is not possible to separate it linearly from images. Speckle noise is a major issue in radar satellite images, ultrasound images, and optical coherence tomography images. It appears due to the interference of waves reflected from an uneven natural surface. When an image is taken, useful true waves that are reflected from a surface in the image may interfere with each other, as well as with waves re-reflected from an uneven surface. This interference creates a specific granular pattern on an acquired image and complicates image interpretation as well as its automatic analysis.

There are many nonlinear filters, which can be used for speckle noise reduction. The most known of them are the Lee filter [13] and the Frost filter [14]. Both these spatial domain filters were specifically designed for taking care of speckle noise. They are based on the analysis of local statistics in a small window around a pixel of interest. Weights are assigned corresponding to intensities from this window based on said analysis. A noisy intensity should be substituted a weighted estimated value found based on the minimum mean square error criterion.

The BM3D filter introduced in [15] is commonly considered as one of the best universal nonlinear filters ever. This filter is based on the frequency domain processing of 3D blocks created from statistically similar overlapping patches. It is highly efficient for additive and speckle noise removal.

It was proposed in [16]-[18] to use a deep learning technique for image filtering. Basically, it was suggested to use a classical feedforward neural network (multilayer perceptron - MLP) with several hidden layers as an intelligent filter. This approach is based on the simultaneous filtering of all pixels from an n x n patch, processing of overlapping patches with a small offset and averaging the filtering results for each pixel over all overlapping patches which to this pixel belongs. A learning process was organized in a way where a neural network should have learned how to remove noise from a noisy input patch and create a clear output patch.

This idea was a basement for the use of MLMVN as an intelligent filter to remove additive noise from images [8]. However, unlike in [16]-[18] where MLP with several hidden layers was used, it was enough to use MLMVN with a single hidden layer to obtain very competitive results, which were slightly better than or comparable to the ones obtained using the BM3D filter [8].

978-1-7281-3214-3/20/$31.00 ©2020 IEEE

There are some fairly recently published works where deep learning and, particularly, a convolutional neural network are used for speckle noise filtering [19], [20].

In this paper, we would like to use MLMVN for speckle noise reduction. We will use the same basic approach, which was employed in [8] to filter additive noise. At the same time, a learning algorithm will be modified with the use of incremental learning. The results will be compared to the ones obtained using the Lee and BM3D filters as the most efficient traditional filters used for speckle noise filtering.

II. Speckle Noise and its Models

Speckle noise is a multiplicative noise. Its mathematical model can be presented in the most general form as follows

$$A(x,y) = I(x,y) \times S(x,y) + R(x,y), \quad (1)$$

where (x,y) are coordinates of a pixel, $A(x,y)$ is an observed intensity value, $I(x,y)$ is an intensity value from an ideal (clear) image, $S(x,y)$ is a multiplicative component of speckle noise, $R(x,y)$ is an additive component of speckle noise.

Often an additive component is not presented (for example, in ultrasound imaging) and in such a case model (1) shall be transformed to

$$A(x,y) = I(x,y) \times S(x,y). \quad (2)$$

We will use in this paper model (2).

To simulate speckle noise corresponding to various image acquisition techniques, various statistical models of noise should be used accordingly. In this work, to simulate noise *S* in (2) we used the following model. A field of the same size as a corresponding image to be corrupted should be generated. This field contains evenly distributed random numbers having the following properties. Its mean as a noise mean shall be equal to that of an image to be corrupted. A variance of noise shall be measured in terms of a fraction of the variance of an image to be corrupted. We used in this work a noise variance $0.05\sigma^2$, $0.1\sigma^2$, and $0.2\sigma^2$ where σ is a standard deviation of an image to be corrupted and σ^2 is its variance, accordingly. A field, generated as it was described, shall be multiplied component-wise with a clear image *I*. As a result, this image becomes corrupted according to (2). The result shall be finally divided by the corresponding mean value.

III. Filtering using MLMVN

We use here the same approach to speckle noise filtering as the one used for additive noise filtering in [8]. Our expectation to get a good result was based on the following considerations. MLMVN was used in [8] as an intelligent filter for additive Gaussian noise and its efficiency was compared to the one of BM3D filter. Since MLMVN returned results that were slightly better than or at least comparable to what can be obtained using BM3D filter – which is highly efficient also for speckle noise filtering, we should have expected that MLMVN can also be efficient in speckle noise reduction.

Let us very briefly recall key moments of the approach presented in [8].

The most important in this approach is that a two-layer MLMVN (thus a network with a single hidden layer) takes an *n* x *n* patch from a noisy image and creates an *m* x *m* filtered patch (where $m < n$) on the output. This means that a neural network processes a smaller sub-patch in the middle of the input patch. An image is divided into *m* x *m* overlapping patches with the offset equal to 1. To get a final result after processing of all patches, it is necessary to average the results over all overlapping estimations of every single intensity value: $f(x,y) = \sum_{i=1}^{R_{xy}} f_i(x,y) / R_{xy}$. Here R_{xy} is the number of overlapping estimations of the pixel with the coordinates (x,y) and $f_i(x,y), i = 1,...,R_{xy}$ are these estimations.

Hence, to use MLMVN for filtering as it is described, it shall have the $n^2 \to H \to m^2$ topology where $n^2 = n \times n$ is the number of network inputs corresponding to the number of pixels in each input patch, H is the number of hidden neurons, and $m^2 = m \times m$ is the number of network outputs. Thus, the number of output neurons matches the number of pixels in an output patch. To take care of patches in any of the border areas, an image should be cropped by an amount of pixels matching the size of an output patch.

To organize a learning process, it is necessary to create a learning set. Each learning sample consists of $n^2 = n \times n$ inputs (intensities, which are taken from a noisy patch) and $m^2 = m \times m$ desired outputs (intensities, which are taken from a clear patch). Patches to be included in a learning set should be selected randomly from various images. The wider the variety of images used in the learning set, the better the results should be expected, in terms of the network generalization capability.

To project integer intensities on the unit circle (which is necessary because MLMVN is a complex-valued neural network), the exact same approach that was used in [8] is used here. The unit circle was divided into 288 sectors while only the first 256 of them were used. This is important in order to avoid occasional mix of black (0) and white (255) output values. If occasionally an output of the network could fall into sectors from 256 through 287, then the following procedure was applied to adjust the output

$$255 < r < 272 \to r = 255,$$
$$272 \le r < 288 \to r = 0.$$

Here, as it was already mentioned above, we used a batch learning algorithm presented in [12]. However, two modifications to this algorithm were made. Typically, any

learning process is based on the minimization of the learning error. There is a common expectation that the smaller the learning error, the more capable a neural network should be of generalization.

However, the actual goal of any learning process is to achieve the best possible generalization capability of a neural network. Taking this into account, we modified the learning algorithm presented in [12]. Since the goal of learning in our case is to reach the best filtering results, a validation error calculated for images that did not participate in the learning process and averaged over all of them, was used as a stopping criterion. Hence the goal of learning should be minimization of the validations error – not the learning error.

To speed up the learning process and avoid overfitting, we also made another modification to the learning algorithm. Basically, we applied an incremental learning technique in the following way. A learning set for solving a problem of image filtering should be fairly large [8]. The only way to make a neural filter robust is to include in a learning set many learning samples taken from many images. However, if a learning set consists of tens of thousands of samples and we try to learn the set as a whole, this may lead to overfitting or to the enlargement of the number of hidden neurons in the network. The latter may help to learn, but it negatively affects the generalization capability. This is because a network basically "memorizes" a learning set but cannot generalize as it is desired. There is also another disadvantage of this – a significantly longer time needed for learning. An alternative approach is to learn not a learning set as a whole, but to divide it into smaller batches and run a learning process batch by batch. We organized a learning process in the following way. Let us suppose our learning set contains N samples. Let us divide it into $r = n / k$ batches containing an equal number of samples. A learning process starts from random weights and only samples from the 1st out of r batches should be involved. The batch learning algorithm [12] should be applied. To stop the learning process for a given batch, a threshold value for a desirable level of the learning error (in terms of the root mean square error – RMSE) should be specified. Once the 1st batch is learned, we move to the 2nd one. It is important that while we start learning the 2nd batch from the weights resulted from the learning of the 1st batch, but samples from the 1st batch are not involved in the learning process for the 2nd batch. Upon the learning process converging for the 2nd batch, we move to the 3rd one, etc. Finally, when the r^{th} batch was learned, this completes a global learning iteration, which means that a pass over the entire learning set is complete. At this moment, we filter all images from a validation set – which did not participate in the learning, and evaluate the average testing RMSE and peak signal to noise ratio (PSNR). Our goal, as it was mentioned above is to minimize testing RMSE or (which is the same) maximize testing PSNR. If the validation error exceeds our desired targeting value, we return to the 1st batch and the next global iteration should be started from the weights resulted from the previous one. This process should continue while the validation error exceeds the desired level. Of course, the learning process can also be stopped by restricting the maximum amount of its global iterations.

IV. Simulation Results

In our experiments, we used exactly the same collection of 410 images, which was used in [8]. It contains 110 images which are available from [21] and 300 images from the authors' collection[1].

Learning sets were created from 400 images, while 10 images, which have not participated in the learning process, were used for validation. We created three learning sets corresponding to noise variance $0.05\sigma^2$, $0.1\sigma^2$, and $0.2\sigma^2$ (where σ is a standard deviation of a clear image), accordingly.

To create a learning set, 100 patches were randomly selected from each of 400 images. As a result, each learning set contains 40,000 samples.

Based on the experimental results, we found that the best size of an input patch is 7x7 and the best size of an output patch is 5x5.

The number of hidden neurons in a single hidden layer may vary. In our experiments, we used from 512 to 3072 hidden neurons. Our experiments show that a larger number of hidden neurons makes it possible to improve the network's generalization capability. At the same time, the number of hidden neurons does not affect the duration of the learning process, which is based on the incremental learning procedure described above. A smaller network needs more iterations to learn a single batch, while a larger network needs fewer iterations, but more time per each iteration. It should also be noted that we did not use parallel data processing in our simulations. They were performed using a serial software simulator.

In our experiments, we tested batches containing 50, 100, 200, and 400 learning samples. Since 100 patches were taken to a learning set from each of 400 learning images, this corresponds to batches composed from blocks taken from a single image (50 or 100), two images (200) and four images (400). We found that with a batch containing 200 learning samples it possible to get better results than with batches of other size. This optimal size does not depend on the noise standard deviation. Thus, our entire learning set containing 40,000 samples is divided into 200 batches containing 200 samples each.

Hence, a global learning iteration with a batch containing 200 samples consists of sequentially running the learning algorithm [12] for each batch. This process starts from the random weights for the first batch and continues for each following batch starting from the weights resulted from the learning of a preceding batch.

[1] All software simulations were performed using designed in MATLAB. The corresponding MATLAB functions along with the data used for simulations in this paper can be found at http://www.freewebs.com/igora/ under "Download Software Simulators and Data" right after publication of this paper.

TABLE I
SIMULATION RESULTS: FILTERING USING MLMVN VS. BM3D FILTER AND LEE FILTER
FOR NOISE WITH THE STANDARD DEVIATION $0.05\sigma^2$, $0.1\sigma^2$, $0.2\sigma^2$ AND ALL TEST IMAGES USED IN SIMULATIONS.

			TEST IMAGES	1	2	3	4	5	6	7	8	9	10	AVERAGE
NOISE $0.05\sigma^2$	**MLMVN 7X7→3072→5X5 71 ITERATIONS.**	PSNR		**32.81**	**29.53**	**30.33**	**30.79**	**29.07**	**28.95**	**34.28**	**33.12**	**35.12**	32.04	**31.60**
		RMSE		**5.82**	**8.51**	**7.76**	**7.36**	**8.98**	**9.10**	**4.92**	**5.62**	**4.47**	6.37	**6.94**
	BM3D FILTER	PSNR		32.17	29.04	29.44	30.39	28.96	28.37	32.21	31.78	34.39	**32.45**	30.83
		RMSE		6.28	9.00	8.59	7.71	9.09	9.73	6.25	6.56	4.86	**6.08**	7.48
	LEE FILTER	PSNR		26.87	26.79	26.72	28.13	27.10	27.06	29.57	28.74	30.60	27.12	27.87
		RMSE		11.12	11.66	11.77	10.00	11.24	11.31	8.47	9.32	7.52	11.23	10.30
NOISE $0.1\sigma^2$	**MLMVN 7X7→3072→5X5 10 ITERATIONS**	PSNR		**30.65**	**27.34**	**27.69**	**28.52**	26.98	**27.09**	**32.35**	**30.80**	33.02	29.86	**29.43**
		RMSE		**7.47**	**10.95**	**10.51**	**9.56**	11.41	**11.28**	**6.15**	**7.35**	5.69	8.19	**8.86**
	BM3D FILTER	PSNR		30.10	26.68	27.15	28.11	**27.01**	26.75	31.41	30.45	**33.37**	**31.64**	29.27
		RMSE		7.97	11.82	11.19	10.02	**11.37**	11.72	6.85	7.65	**5.47**	**6.67**	9.07
	LEE FILTER	PSNR		26.62	25.17	25.19	26.44	25.25	25.20	27.38	26.81	28.71	24.85	26.11
		RMSE		11.89	14.05	14.01	12.14	13.93	14.01	10.89	11.63	9.79	14.58	12.69
NOISE $0.2\sigma^2$	**MLMVN 5X5→2048→3X3 52 ITERATIONS**	PSNR		**28.12**	**25.17**	**24.98**	26.15	**24.85**	**24.76**	**30.21**	**28.35**	30.47	**28.10**	**27.12**
		RMSE		**10.01**	**14.06**	**14.38**	12.55	**14.59**	**14.73**	**7.87**	**9.74**	7.63	**10.04**	**11.56**
	BM3D FILTER	PSNR		26.62	24.50	23.83	**26.96**	24.32	23.92	29.91	27.51	**31.80**	26.29	26.57
		RMSE		11.94	15.39	16.46	**11.56**	15.50	15.50	8.77	10.76	**7.01**	12.32	12.49
	LEE FILTER	PSNR		24.07	22.66	22.46	24.09	22.84	22.76	25.08	24.47	25.80	29.31	24.35
		RMSE		15.95	18.75	19.19	15.91	13.39	18.56	14.20	15.24	13.08	8.73	15.30

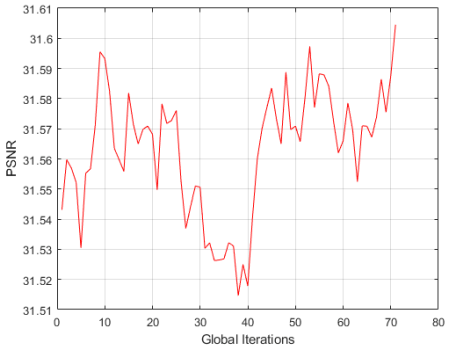


Fig. 1. Convergence curve (PSNR vs. # of global iterations) for the learning set for noise with the variance $0.05\sigma^2$

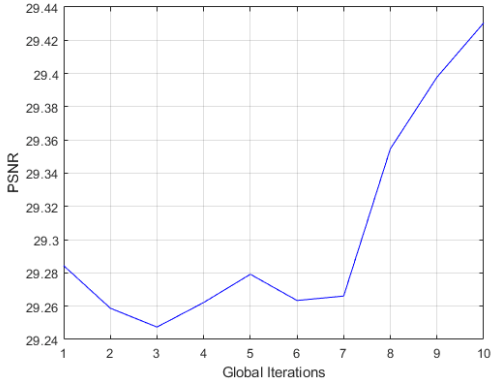


Fig. 2. Convergence curve (PSNR vs. # of global iterations) for the learning set for noise with the variance $0.1\sigma^2$

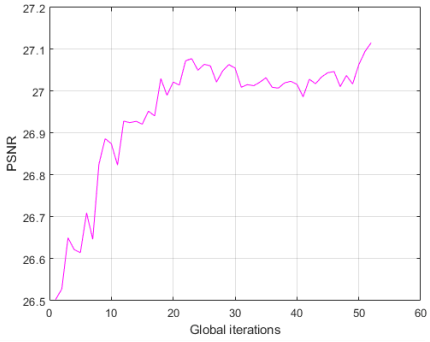


Fig. 2. Convergence curve (PSNR vs. # of global iterations) for the learning set for noise with the variance $0.2\sigma^2$

We tested three RMSE threshold values (4.0, 5.0, and 6.0) as stopping criteria for the learning process running for every single batch. Contrary to a standard, traditional logic of the learning error minimization, our experiments show that the RMSE threshold 6.0 gave better filtering results than 4.0 and 5.0. Evidently, lower thresholds lead to some kind of "memorization" of the corresponding learning samples, but we should not forget that our goal is to obtain better generalization capability and better filtering results for our validation images.

On the first global iteration, only the learning of the first batch starting from random weights requires 15-30 local iterations. To learn the rest of batches, 5-10 iterations are needed. On the next global iterations, just a few local iterations are needed to learn every batch. This speed up the learning process drastically.

Even after the very first global learning iteration, which takes just a few minutes, the method presented here shows slightly better results than BM3D filter (for noise with variance $0.05\sigma^2$, $0.1\sigma^2$) and about the same results as BM3D

filter (for noise with variance $0.2\sigma^2$). These results can be improved continuing the learning process. In Fig. 1-3, convergence curves for the learning algorithm described here are shown. In every experiment, if the results in terms of validation PSNR were improving, then the learning process would continue to run. Our final goal was to get the average PSNR over 10 test images by 0.5-0.7 higher than the one resulted from BM3D filter. Table 1 contains the results for all three noise variances. The method, which is presented here is compared to BM3D filter and Lee filter. The results are shown for all 10 validation images, which have not participated in the MLMVN learning process. The last column contains RMSE and PSNR averaged over all 10 test images. The best results are shown in bold. MLMVN outperforms the BM3D filter for 7-9 out of 10 validation images and the Lee filter for all of them.

Fig. 4-7 illustrate the results of filtering of one out of the ten test images corrupted by speckle noise with variance $0.05\sigma^2$. Fig. 8-11 illustrate the results of filtering of one out of the ten test images corrupted by speckle noise with variance $0.1\sigma^2$. While BM3D filter can remove noise completely, at the same time it also "washes" or even removes small details from an image. After MLMVN filtering some very slight noise leftovers can still be visible, but MLMVN preserves small details on an image more efficiently. This is resulted in higher PSNR obtained using MLMVN for all test images containing many small details. At the same time, the BM3D filter slightly outperforms MLMVN when working with images containing no small details.

Image processing with BM3D filter was done using the MATLAB implementation available online from the web page devoted to BM3D filter by its authors [22].

V. Conclusions

The main result presented in the paper is the experimental proof that MLMVN can be used for intelligent speckle noise filtering by processing overlapping patches. It performs slightly better or at least it is comparable to the BM3D filter. MLMVN better preserves small details in images and this is its main advantage. We also used minimization of the validation error, not of the learning error as a stopping criterion for the learning algorithm. To avoid overfitting and better adapt a large learning set, an incremental learning approach was employed to organize the learning process. As a result, the learning process can be significantly sped up.

References

[1] A. Hirose, *Complex-Valued Neural Networks, 2nd Edn.*, Springer, Berlin, Heidelberg, 2012.

[2] I. Aizenberg, I., *Complex-Valued Neural Networks with Multi-Valued Neurons*. Berlin: Springer-Verlag Publishers, 2011.

[3] Y.Nakano, 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, Issue 1, pp. 102-108, Jan. 2009.

[4] 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.

[5] 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, issue 1, pp. 41-52, Jan. 2010.

[6] Aizenberg I., Sheremetov L., Villa-Vargas L., and Martinez-Muñoz J., "Multilayer Neural Network with Multi-Valued Neurons in Time Series Forecasting of Oil Production", *Neurcomputing*, vol. 175, part B, pp. 980-989, Jan. 2016.

[7] 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, pp. 185-208, 2012.

[8] I. Aizenberg, A. Ordukhanov, and F. O'Boy, "MLMVN as an Intelligent Image Filter", *Proceedings of the 2017 IEEE International Joint Conference on Neural Networks* (IJCNN 2017), Anchorage, pp. 3106-3113, May, 2017.

[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, issue 2, pp. 169-183, Jan. 2007.

[11] I. Aizenberg, A. Luchetta, and S. Manetti, "A modified Learning Algorithms for the Multilayer Neural Network with Multi-Valued Neurons based on the Complex QR Decomposition", *Soft Computing*, vol. 16, pp. 563-575, April 2012.

[12] E. Aizenberg and I. Aizenberg, "Batch LLS-based Learning Algorithm for MLMVN with Soft Margins", *Proceedings of the 2014 IEEE Symposium Series of Computational Intelligence (SSCI-2014)*, pp. 48-55, December, 2014.

[13] J-S Lee, "Digital Image Enhancement and Noise Filtering by Use of Local Statistics", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol.. 2, No. 2, pp. 165 – 168, March 1980.

[14] V. S. Frost, J. A. Stiles, K. S. Sshanmugan,, D Julian, C. Holtzman, "A Model for Radar Images and Its Application to Adaptive Digital Filtering of Multiplicative Noise", ", *IEEE Transactions on Pattern Analysis and Machine Intelligence*,, vol.. 4, No. 2, pp. 157 – 166, March 1982.

[15] K. Dabov, A. Foi, V. Katkovnik, and K. Egiazarian, "Image Denoising by sparse 3-D Transform-Domain Collaborative Filtering". *IEEE Transactions on Image Processing*, vol. 16, no 8, pp. 2080–2095, 2007.

[16] H. C. Burger, C. J. Schuler, and S. Harmeling, "Image denoising: Can plain Neural Networks compete with BM3D?", *Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition* (CVPR-2012), pp. 4321-4328, 2012.

[17] H. C. Burger, C. J. Schuler, and S. Harmeling, "Image Denoising with Multi-Layer Perceptrons, Part 1: Comparison with Existing Algorithms and with Bounds", available online at http://arxiv.org/pdf/1211.1544.pdf, 2012.

[18] H. C. Burger, C. J. Schuler, and S. Harmeling, "Image Denoising with Multi-Layer Perceptrons, Part 2: Training Trade-offs and Analysis of their Mechanisms, available online at http://arxiv.org/pdf/1211.1552.pdf, 2012.

[19] P. Wang, H. Zhang, and V. M. Patel, SAR Image Despeckling Using a Convolutional Neural Network, IEEE Signal Processing Letters, vol. 24, No 12, pp. 1763 – 1767, December, 2012

[20] T. L. Bobrov, F. Mahmoud, M. Inserni, and N. J. Durr, DeepLSR: a deep learning approach for laser speckle reduction, Biomedical Optics Express, vol. 10, No 6, pp. 2869-2882, June 2019.

[21] Test Images. University of Granada Image Data Base. Available Online http://decsai.ugr.es/cvg/dbimagenes/

[22] Image and video denoising by sparse 3D transform-domain collaborative filtering. Block-matching and 3D filtering (BM3D) algorithm and its extensions. Tampere University of Technology. Online http://www.cs.tut.fi/~foi/GCF-BM3D/index.html#ref_software

Fig. 4. The “Train Station” - original image, which was not used in the learning sets

Fig. 5. The “Train Station” image from Fig. 4 corrupted by speckle noise with variance $0.05\sigma^2$

Fig. 6. The image from Fig. 5 filtered with MLMVN containing 3072 hidden neurons. PSNR=32.81. Small details are mostly preserved.

Fig. 7. The image from Fig. 5 filtered using BM3D filter; PSNR=32.17. Many small details are gone after filtering.

Fig. 8. The “Sailfish” image, which was not used in the learning sets (the original image)

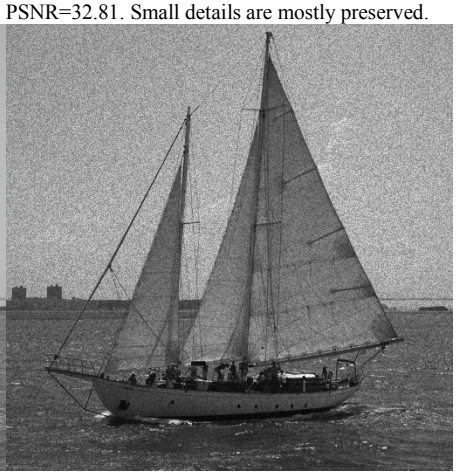


Fig. 9 The “Sailfish” image from Fig. 8 corrupted by speckle noise with variance $0.1\sigma^2$

Fig. 10. The image from Fig. 9 filtered with MLMVN MLMVN containing 3072 hidden neurons. PSNR=32.35. Small details are mostly preserved

Fig. 11. The image from Fig. 9 filtered using BM3D filter; PSNR=31.41. Many small details are gone after filtering

IEEE Third International Conference on Data Stream Mining & Processing
August 21-25, 2020, Lviv, Ukraine

Two Approaches to Machine Learning Classification of Time Series Based on Recurrence Plots

Lyudmyla Kirichenko
Department of Applied mathematics
Kharkiv National University of Radio Electronics
Kharkiv, Ukraine
lyudmyla.kirichenko@nure.ua

Tamara Radivilova
V.V. Popovskyy dept. of Infocommunication Engineering
Kharkiv National University of Radio Electronics
Kharkiv, Ukraine
tamara.radivilova@gmail.com

Vitalii Bulakh
Department of Applied mathematics
Kharkiv National University of Radio Electronics
Kharkiv, Ukraine
bulakhvitalii@gmail.com

Petro Zinchenko
Department of Applied mathematics
Kharkiv National University of Radio Electronics
Kharkiv, Ukraine
petro.zinchenko@nure.ua

Abed Saif Alghawli
Department of Computer Science
College of Sciences and Humanities
Prince Sattam Bin Abdulaziz University
Aflaj, Kingdom of Saudi Arabia
alghauly@yahoo.com

***Abstract*— The article considers the task of classifying fractal time series based on the construction of their recurrence plots. Short realizations of EEG signals were used as input data. Two classification machine learning methods were considered: in the first case, quantitative fractal and recurrent characteristics of the time series were classification features, in the second case, image recognition of recurrence plots was carried out. The results showed fairly high classification quality for both methods, and relative advantage of the image classification method.**

Keywords— fractal time series, classification of time series, recurrence plot, machine learning classification

I. INTRODUCTION

Time series analysis is found in many practical applications ranging from human activity detection to cybersecurity. In many cases, the problem of time series classification occurs. Moreover, any classification task that uses ordered data can be considered as a time series classification task.

Over the past decades, time series classification methods based on machine learning have been developed [1-5]. One of the most popular approaches to classification is to extract from time series a set of some features, which are input data for the classifier [6-9].

Nowadays methods of nonlinear dynamics, including the method of recurrence plots, are widely used for time series analysis. Recurrence analysis is based on repeatability of time series states and allows presenting a time series as a geometric structure. The topology of such geometric structure allows to reveal and analyze the characteristic features of time series dynamics of different nature [10, 11].

With the development of machine learning methods and approaches to the selection of features for classification, a number of researchers began to use quantitative recurrence characteristics of time series in the tasks of classification. Thus, for example, in [12] the classification of time series is considered by an example of video compression algorithms with the use of distance measure of cross recurrence plots; in [13], using the support vector machine, the heart rate variability is classified based on recurrence features; in [14] support vector machine algorithm and features of recurrence plots are used.

The other approach to time series classification is the visualization of recurrence plots and image recognition using deep neural networks. In [15], this approach is proposed to recognize and monitoring user behavior; in [16] and [17], the recognition of recurrence plots using convolutional neural networks was applied to different types of time series.

The classification of time series with fractal properties is of particular interest. Numerous studies have shown that time series, changing the character of dynamics, also change the fractal properties. This concerns attacked traffic [18], biomedical signals [19], financial series [20], etc. In [21–23], it was shown that change in the fractal properties of the time series leads to change in the recurrence characteristics. Thus it is possible to use recurrence quantitative characteristics as features for classification of fractal time series [23-25] or to use visualization of recurrence plots for classification [26].

The purpose of the present work is comparative analysis of two approaches to classifying fractal time series: first use fractal and recurrence quantitative characteristics as features for a classifier and second base on the classification of recurrence plot images

II. RECURRENCE PLOTS

A recurrence plot is a matrix of distances between points of a time series in pseudophase space. It is assumed that the analyzed time series $x(t)\ \ t=1,...N$ is the trajectory of some dissipative system. Using the Packard-Tackens's method [27], one can reconstruct attractor of the system in the m-dimensional phase space from this trajectory, where the phase variables are time-delay values of $x(t)$:

978-1-7281-3214-3/20/$31.00 ©2020 IEEE

$$F = \left(x(t), x(t+\tau), ..., x(t+(m-1)\tau)\right)$$

where τ is the time lag.

In this case, the recurrence plot is a square matrix $RP_{i,j}$, $i, j = 1,...N$. An element of this matrix $RP_{i,j}$ is equal to 1 if the distance between the points $x(t_i)$ and $x(t_j)$ in the *m*-dimensional phase space F does not exceed some pre-assigned value ε. Otherwise, $RP_{i,j}$ equal to 0. [28]

If put the recurrence plot elements $RP_{i,j}$ according black and white, a some topological structure of the time series is got. Qualitative and quantitative topological analysis of the recurrence structures allows to classify time series based on their recurrence l properties. Quantitative measures of recurrence structures were discussed in detail in [29,30]. Consider the main ones.

Recurrence rate (*RR*) shows the density of recurrence points, i.e., the probability of recurrence $x(t_i)$ in the *m*-dimensional phase space:

$$RR = \frac{1}{N^2}\sum_{i,j=1}^{N} RP_{i,j} .$$

Important in the topology of recurrence structures are diagonal structures, which show the time interval when one time series section comes close enough to another. Deterministic time series, for example, periodic ones, have long diagonal lines, while stochastic series are either very short or do not have them at all.

To calculate quantitative characteristics based on diagonal structures, we consider the frequency distribution of the lengths *l* of diagonal lines in a recurrence plot:

$$P\ (l) = \{l_i; i = 1...N_l\},$$

where l_i is length of i -th diagonal line, N_l is number of diagonal lines. The average length of the diagonal lines is calculated as follows:

$$L = \frac{\sum_i^{N_l} l_i \cdot P(l_i)}{\sum_i^{N} P(l_i)}.$$

The determinism (*DET*) or system predictability measure is called the following relation of recurrence points:

$$DET = \frac{\sum_i^{N_l} l_i \cdot P(l_i)}{\sum_{i,j=1}^{N} RP_{i,j}}$$

Denote the frequency distribution for the set of lengths of vertical lines as

$$P\ (v) = \{v_k; k = 1, 2, ..., K\},$$

where v_k is length of k -th vertical line, K is number of vertical lines.

The laminarity measure (*LAM*) characterizes the presence of system states when the system motion along the phase trajectory practically stops:

$$LAM = \frac{\sum_{v=v_{\min}}^{N} vP\ (v)}{\sum_{i,j}^{N} RP_{i,j}} .$$

Quantitative measures of recurrence can be used as features in the classification using machine learning [23-25].

III. Fractal Characteristics of Time Series

The fractality of stochastic processes is to preserve distribution laws 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 distribution laws as $X(t)$. The parameter H, $0 < H < 1$, is called the Hurst exponent and it is the self-similarity measure. The moments of the self-similar stochastic process satisfy the following scaling relation

$$E\left[\left|X(t)\right|^q\right] \propto t^{qH} .$$

Multifractal stochastic processes are inhomogeneous fractal ones. Scaling relation for their moment characteristics is described by

$$E\left[\left|X(t)\right|^q\right] \propto t^{qh(q)},$$

where $h(q)$ is generalized Hurst exponent. The Hurst exponent H of multifractal processes is equal to the value $h(q)$ at $q = 2$. [31]

There are many. One of the most practical approaches to estimate the fractal characteristics by time series is the method of multifractal detrended fluctuation analysis (MFDFA) [32]. MFDFA allows to calculate the fluctuation function $F(\tau)$ which has scaling relation $F_q(\tau) \propto \tau^{h(q)}$.

IV. classification Methods

As a result of a number of preliminary experiments, the simplest type of feedforward neural network (perseptron) was selected as a classifier based on quantitative features. In this work the perseptron with seven layers was used.

The ReLU (rectified linear unit) semi-line function was used as an activation function. The main advantage of this function is the multifold increase in the rate of the gradient descent convergence in comparison with traditional activation functions, which is due to the linear nature of the function [33].

After each fully connected layer of the neural network, the regularization layer was connected. The method of batch normalization was used in layer regularization to improve the efficiency of neural network learning and to solve the problem of overfitting [34].

Batch normalization reduces the covariance shift, i.e. the divergence of distribution parameters of features values in the

learning and test samples (mathematical expectation, dispersion, etc.). That is, when using normalization batches, the input data are normalized in such a way as to obtain a zero math expectation and a unit dispersion. Normalization is performed before entering each layer.

Also batch normalization has a number of advantages, including: faster convergence of models is achieved, each layer of the network is trained more independently of the others; batch normalization is a mechanism of regularization, because it brings some noise to the outputs of nodes hidden layers, etc.

The Adam (adaptive moment estimation) algorithm was used as the optimization method for the neural network [35]. In contrast to the classic stochastic gradient descent, which supports a single learning rate, Adam calculates individual adaptive learning rates for each parameter calculated on the basis of the first and second gradient moments.

The Adam algorithm has many advantages, among which are the following: it is simple to implement, computationally efficient, has small memory requirements, and is good for tasks with a large number of parameters.

The most frequently used for image classification are convolution neural networks, which are deep neural networks. Deeper neural networks should theoretically work better than smaller models. But the big problem is optimization in deep models, as they are more difficult to optimize because gradients spread from upper to lower layers. [36]

To solve this problem, a deep residual network architecture was developed, based on a residual block, which is a number of bundled layers with activations, with a shortcut connection (fast connections). The basic idea of the residual network is that instead of the usual serial connection of layers of neural network, a fast access connection is used, which allows the signal to pass almost without changes. [37,38]

In this case, if on some layer the network has already sufficiently well approximated the original function that generates the data then on further layers the optimizer in Res blocks can make weights close to zero, and the data will pass through the shortcut connection almost unchanged, which leads to a much faster convergence and better learning.

In this work, a residual neural network containing 131 layers with weights was constructed for image classification. These layers were divided into 11 separate layers and 120 combined into blocks with a quick access connection. The Adam algorithm was also used as the optimization method for this residual neural network.

V. Input Data

To conduct a comparative analysis, the well-known dataset was used [39]. It contains records of electroencephalograms (EEG) for various human conditions. It is widely known that EEG realizations have fractal properties, which vary depending on the state of a person [19, 40]. It allows us to use fractal and recurrent characteristics as features for classifier.

This date set, which contains 5 classes of EEG records, is traditionally used for binary classification: epileptic seizure data (one class) and all other cases (four classes). Each class consists of 100 files, where each file corresponds to one object. Every file contains 23 records with length of 178 values, which corresponds to 1 second. Thus, the data is divided into 2 classes, where the first contains 2300 time series, and the second - 11500.

Fig. 1 represents two EEG realizations from the first class (above) and two realizations from the second class (below).

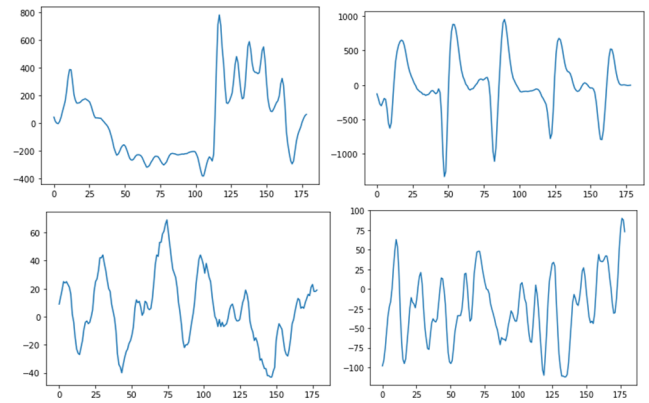


Fig. 1. EEG realizations of the first class (above) and of the second class (below).

Fig. 2 shows the recurrence plots constructed by the above time series. At the top of Fig. 2, plots of the first class are represent, and at the bottom - of the second class. These recurrence plots were the basis of inputs to neural networks. In the first case, the recurrence plot was the matrix for calculation of quantitative recurrence characteristics, and in the second case, recurrence plot was image for recognition.

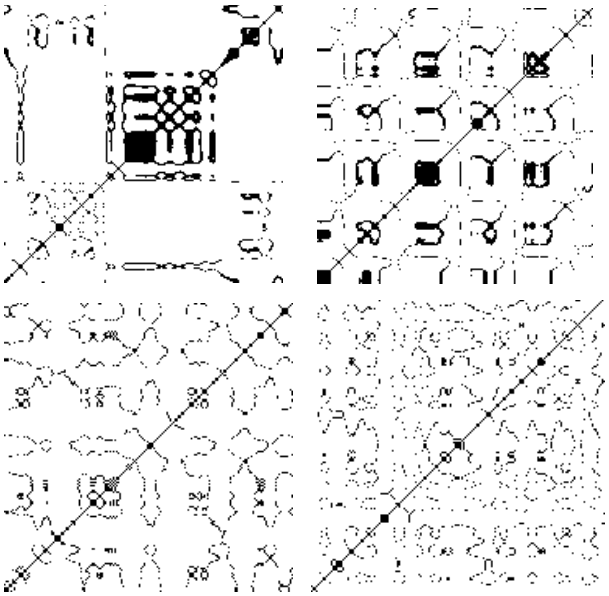


Fig. 2. Recurrence plots of the first class (above) and of the second class (below).

VI. Experiment Description

When classifying the entire data set was used (11,500 time series), where 8,500 time series were selected for training (6,800 cases with epilepsy and 1,700 without it) and 3,000 for testing (2,400 cases with epilepsy and 600 without it).

Consider the case when the classification was carried out according to quantitative characteristics. Fig. 3 shows the time realizations and the corresponding functions of the generalized Hurst exponent $h(q)$, evaluated by the MFDFA method.

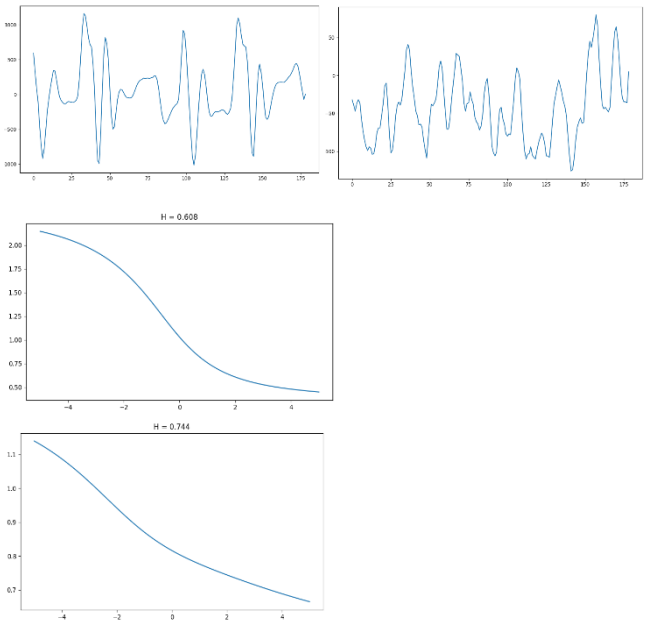


Fig. 3. EEG realizations and corresponding functions of generalized Hurst exponent.

The left part of Fig. 3 represents a typical realization and function $h(q)$ for the 1st class of EEG records, and on the right part the 2nd class realization with the corresponding function $h(q)$ is shown. Preliminary studies have demonstrated that realizations of different classes have significant differences in the range of generalized Hurst exponent $h(q)$ and Hurst exponent H values. It should be noted that the short length of the time series (178 values) causes large errors in the direct estimation of generalized Hurst exponent [5, 8]. Fig. 4 shows the recurrence plots for realization presented in Fig. 3.

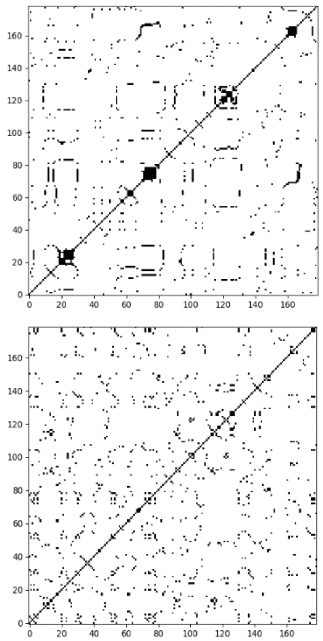


Fig. 4. Recurrence plots: first class (left) and second class (right)

For each recurrence plot, recurrence characteristics, such as determinism, average length of the diagonal lines, laminarity and others were obtained. Table 1 presents some quantitative recurrence characteristics corresponding to the plots shown in Fig. 3. We can note a significant difference between the characteristics of different classes realizations. Studies have shown that such differences are typical.

TABLE I. QUANTITATIVE RECURRENCE CHARACTERISTICS

Class	*RR*	*Det*	*L*	*LAM*
1	0,027	0.147	4.58	0.293
2	0,035	0.031	3.12	0.064

Thus, in the case of classification based on quantitative characteristics, the following features were received at the input of a fully connected perceptron:

- statistical characteristics (median, coefficient of variation, etc.) calculated from the values of the normalized EEG time series;
- fractal characteristics (the value of the Hurst exponent, the range of values of the generalized Hurst exponent and some its values, etc.) obtained from sample generalized Hurst exponent which was estimated by the MFDFA method;
- recurrent characteristics, such as *DET*, *LAM*, *L*, etc., obtained from the recurrence plot of EEG time series recurrence plot.

In the second case (classification based on recurrence plots), black-and-white images of recurrence plots of EEG time series were provided to the input of a residual neural network.

VII. RESULTS AND DISCUSSION

As a result of binary classification, in both cases confusion matrix with True Positive (TP), False Positive (FP), False Negative (FN) and True Negative (TN) values was obtained. The following metrics were selected as the classification results.

Accuracy is proportion of correct algorithm answers:

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

This metric is poorly informative in tasks with unbalanced classes. Since in our task the distribution of EEG records by classes is uneven, it is also necessary to consider other metrics.

Precision shows the proportion of objects that are called positive by the classifier and are truly positive:

$$Precision = \frac{TP}{TP + FP}$$

Recall shows what proportion of positive class objects from all objects of a positive class the algorithm found:

$$Recall = \frac{TP}{TP + FP}$$

Recall demonstrates the ability of the algorithm to detect a given class in general, and *Precision* demonstrates the ability to distinguish this class from other classes. *Precision* and *Recall* do not depend, unlike *Accuracy*, on the correlation of

classes and are therefore applicable in conditions of unbalanced samples.

One of several ways to combine *Precision* and *Recall* metrics into one aggregated criterion is to calculate the *Fmeasure,* in this case it is harmonic mean of *Precision* and *Recall*:

$$F\ measure = 2\frac{Precision * Recall}{Precision + Recall}$$

Table 2 shows the classification metrics obtained after the experiment by both methods.

TABLE II. CLASSIFICATION METRICS

	Classification by quantitative characteristics	Classification by recurrence plots
Accuracy	0.97	0.984
Precision	0.93	0.9792
Recall	0.926	0.94
F-measure	0.928	0.9592

Both methods showed a fairly high classification quality. The obvious advantage of image-based classification allows us to make the assumption that the most important features for recognition are recurrence ones. In this case, the recurrence plots themselves, presented in the form of black and white images, contain more information about the time series than the quantitative characteristics calculated from them. It can also be assumed that fractal characteristics calculated over short time series do not carry enough information to create the advantage of classification according to the quantitative characteristics of the time series.

It is worth noting that in [41], the classification results of the considered dataset using the machine learning algorithms Artificial Neural Networks, Naive Bayesian, k-Nearest Neighbor, Support Vector Machines and k-Means are presented. These results do not exceed the results presented in this paper.

VIII. CONCLUSION

A comparative analysis of the machine learning classification methods of fractal time series based on the use of recurrence plots have been carried out. The data for the binary classification were the EEG records. One class was represented by EEG realizations with epileptic seizure, another - without a seizure.

Two different classification methods have been used. In the first, classification on the basis of quantitative characteristics of the time series, such as fractal and recurrent was carried out. In the second, images of recurrence plots were classified.

In both approaches, neural networks were used as classifiers. In the case of classification based on quantitative characteristics, fully connected perceptron with regularization layers was chosen. In the case of classification based on recurrence plots images, the residual neural network was a classifier.

Both methods showed a fairly high classification quality. Higher accuracy was shown by the method of image recognition. The results can be used to classify biomedical signals with fractal properties.

Future research will be aimed at improving the network architecture for recognizing images of recurrence plots and development of methods for obtaining more “contrasting” recurrence plots in the sense of image classification.

REFERENCES

[1] J. C. B. Gamboa. (2017). “Deep learning for time-series analysis,” *arXiv preprint*. Available: https://arxiv.org/pdf/1701.01887.pdf [April, 04, 2020].

[2] P. Esling and C. Agon. “Time-series data mining.” *ACM Computing Surveys (CSUR),* vol. 45(1), pp. 1-34, 2012.

[3] H. Ismail Fawaz, G. Forestier, J. Weber, L. Idoumghar and P. A. Muller. “Deep learning for time series classification: a review.” *Data Mining and Knowledge Discovery,* vol.33(4), pp.917-963, 2019. doi: https://doi.org/10.1007/s10618-019-00619-1

[4] K. Buza, “Time series classification and its applications,” In Proceedings of the 8th International Conference on Web Intelligence, Mining and Semantics, June 2018, pp. 1-4. doi: https://doi.org/10.1145/3227609.3227690

[5] L. Kirichenko, T. Radivilova and V. Bulakh. “Machine Learning in Classification Time Series with Fractal Properties.” *Data,* vol.4(1) 5, pp.1-13, 2019. doi: https://doi.org/10.3390/data4010005

[6] B. D. Fulcher. “Feature-Based Time-Series Analysis.” *Feature Engineering for Machine Learning and Data Analytics,* 2018, pp.87-113.

[7] M. A. Trovero and M. J. Leonard. “SAS2020-2018 Time Series Feature Extraction.” *Paper SAS Institute Inc*., pp.1-18, 2018. Available: https://www.sas.com/content/dam/SAS/support/en/sas-global-forum-proceedings/2018/2020-2018.pdf [April, 13, 2020].

[8] V. Bulakh, L. Kirichenko and T. Radivilova, "Time Series Classification Based on Fractal Properties," *2018 IEEE Second International Conference on Data Stream Mining & Processing (DSMP)*, Lviv, 2018, pp. 198-201. doi: 10.1109/DSMP.2018.8478532

[9] M. Faraggi and K. Sayadi. (2019). “Time series features extraction using Fourier and Wavelet transforms on ECG data.” [Online]. Available: https://blog.octo.com/en/time-series-features-extraction-using-fourier-and-wavelet-transforms-on-ecg-data/ [April, 05, 2020].

[10] N. Marwan, M. C. Romano, M. Thiel and J. Kurths. “Recurrence plots for the analysis of complex systems.” *Physics reports*, vol. 438, iss.5-6, pp.237-329, 2007. doi: https://doi.org/10.1016/j.physrep.2006.11.001

[11] N. Marwan. “A historical review of recurrence plots.” *The European Physical Journal Special Topics*, vol.164, iss. 1, pp.3-12, 2008. doi: https://doi.org/10.1140/epjst/e2008-00829-1

[12] T. Michael, S. Spiegel and S. Albayrak. “Time series classification using compressed recurrence plots,” *In Proceedings of ECML-PKDD*, 2015.

[13] M. Mohebbi, H. Ghassemian and B. M. Asl. “Structures of the recurrence plot of heart rate variability signal as a tool for predicting the onset of paroxysmal atrial fibrillation.” *Journal of medical signals and sensors*, vol. 1(2), pp.113-121, 2011.

[14] V. M. A. Souza, D. F. Silva and G. E. Batista, “Extracting Texture Features for Time Series Classification,” *2014 22nd International Conference on Pattern Recognition*, Stockholm, 2014, pp. 1425-1430. doi: 10.1109/ICPR.2014.254

[15] E. Garcia-Ceja, M. Z. Uddin and J. Torresen. “Classification of recurrence plots’ distance matrices with a convolutional neural network for activity recognition.” *Procedia computer science*, vol. 130, pp.157-163, 2018.

[16] N. Hatami, Y. Gavet and J. Debayle. “Bag of recurrence patterns representation for time-series classification.” *Pattern Analysis and Applications*, vol. 22, iss.3, pp.877-887, 2019. doi: https://doi.org/10.1007/s10044-018-0703-6

[17] N. Hatami, Y. Gavet and J. Debayle, “Classification of time-series images using deep convolutional neural networks,” *In Tenth International Conference on Machine Vision* (ICMV 2017) (Vol.

10696, p. 106960Y). International Society for Optics and Photonics. April, 2018.

[18] R. Deka and D. Bhattacharyya, "Self-similarity based DDoS attack detection using Hurst parameter." *Security and Communication Networks*, vol. 9, № 17, pp. 4468-4481, 2016. doi: https://doi.org/10.1002/sec.1639

[19] A. Alghawli and L. Kirichenko, "Multifractal Properties of Bioelectric Signals under Various Physiological States", *Information Content & Processing International Journal*, vol. 2(2), pp. 138-163, 2015.

[20] E. E. Peters. "Fractal Market Analysis: applying chaos theory to investment and economics". Wiley, 2003

[21] L. Kirichenko, Yu. Kobitskaya and A. Habacheva. "Comparative Analysis of the Complexity of Chaotic and Stochastic Time Series." *Radio Electronics, Computer Science, Control,* vol. 2(31), pp. 126-134, 2014.

[22] L. Kirichenko, O. Baranovskyi and Yu. Kobitskaya. "Recurrence analysis of self-similar and multifractal time series." *Information Content and Processing*, vol. 3(1), pp. 16-37, 2016.

[23] L. Kirichenko, T. Radivilova and V. Bulakh, "Classification of Fractal Time Series Using Recurrence Plots," *2018 International Scientific-Practical Conference Problems of Infocommunications. Science and Technology (PIC S&T)*, Kharkiv, Ukraine, 2018, pp. 719-724. doi: 10.1109/INFOCOMMST.2018.8632010

[24] L. Kirichenko, T. Radivilova and V.Bulakh. "Binary Classification of Fractal Time Series by Machine Learning Methods." In: Lytvynenko V., Babichev S., Wójcik W., Vynokurova O., Vyshemyrskaya S., Radetskaya S. (eds) Lecture Notes in Computational Intelligence and Decision Making. ISDMCI 2019. *Advances in Intelligent Systems and Computing,* 1020, pp. 701-711, 2020. doi: https://doi.org/10.1007/978-3-030-26474-1_49

[25] T. Radivilova, L. Kirichenko, D. Ageiev and V. Bulakh, "Classification Methods of Machine Learning to Detect DDoS Attacks," *2019 10th IEEE International Conference on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications (IDAACS)*, Metz, France, 2019, pp. 207-210. doi: 10.1109/IDAACS.2019.8924406

[26] L. Kirichenko, P. Zinchenko, T. Radivilova and M. Tavalbeh, "Machine Learning Detection of DDoS Attacks Based on Visualization of Recurrence Plots," *Proceedings of the International Workshop on Conflict Management in Global Information Networks* (CMiGIN 2019) co-located with 1st International Conference on Cyber Hygiene and Conflict Management in Global Information Networks (CyberConf 2019), Lviv, Ukraine, November 29, 2019. http://ceur-ws.org/Vol-2588/paper3.pdf

[27] F. Takens, "Detecting strange attractors in turbulence." In: Rand D., Young L.S. (eds) Dynamical Systems and Turbulence, Warwick 1980. *Lecture Notes in Mathematics,* vol.898, pp.366-381, 1981. doi: https://doi.org/10.1007/BFb0091924

[28] J. P. Eckmann, S. O. Kamphorst and D. Ruelle. "Recurrence plots of dynamical systems." *Europhysics Letters*, vol. 4(9), pp.973-977, 1987.

[29] J.P. Zbilut, J.-M. Zaldivar-Comenges and F. Stozzi. "Recurrence quantification based Liapunov exponent for monitoring divergence in experimental data." *Physics Letters A*, vol.297(3–4), pp.173-181, 2002. doi: https://doi.org/10.1016/S0375-9601(02)00436-X

[30] E. J. Ngamga, A. Nandi, R. Ramaswamy, M.C. Romano, M. Thiel and J. Kurths. "Recurrence analysis of strange nonchaotic dynamics." *Physical Review E*, vol. 75(3), pp.036222, 2007. doi:10.1103/PhysRevE.75.036222

[31] R.H. Riedi. Multifractal Processes.. Multifractal processes (No. RU-ECE-TR-99-06). Rice Univ Houston Tx Dept Of Electrical And Computer Engineering, 1999.

[32] J. Kantelhardt, S. Zschiegner, E. Koscielny-Bunde, S. Havlin, A. Bunde and H. Stanley. "Multifractal detrended fluctuation analysis of nonstationary time series." P*hysica A: Statistical Mechanics and its Applications*, vol. 316, № 1-4, pp. 87-114, 2002. doi: https://doi.org/10.1016/S0378-4371(02)01383-3

[33] J. Brownlee (2019, January 9). "A Gentle Introduction to the Rectified Linear Unit (ReLU)." [Online]. *Machine learning mastery*. Available: https://machinelearningmastery.com/rectified-linear-activation-function-for-deep-learning-neural-networks/

[34] S. Ioffe and C. Szegedy, "Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift," *In Proceedings of the 32nd International Conference on Machine Learning*, Lille, France, PMLR, vol. 37, pp.448-456, 2015.

[35] D. P. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization," *In Proceedings of the 3rd International Conference on Learning Representations* (ICLR), San Diego, USA, 2015.

[36] D. C. Ciresan, U. Meier, J. Masci, L. M. Gambardella and J. Schmidhuber, "Flexible, High Performance Convolutional Neural Networks for Image Classification," *Proceedings of the Twenty-Second International Joint Conference on Artificial Intelligence*, vol.2, pp. 1237–1242, 2011.

[37] K. He, X. Zhang, S. Ren and J. Sun, "Deep Residual Learning for Image Recognition," *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, NV, 2016, pp. 770-778. doi: 10.1109/CVPR.2016.90

[38] V. Fung. (2017, July) "An overview of resnet and its variants." [Online]. *Towards data science.* Available: https://towardsdatascience.com/an-overview-of-resnet-and-its-variants-5281e2f56035

[39] Q. Wu, E. Fokoue, "Epileptic Seizure Recognition Data Set." [Online]. *DataSet.* Available: https://archive.ics.uci.edu/ml/datasets/Epileptic+Seizure+Recognition

[40] , R. G. Andrzejak, K. Lehnertz, F. Mormann, C. Rieke, P. David and C. E. Elger. "Indications of nonlinear deterministic and finite-dimensional structures in time series of brain electrical activity: Dependence on recording region and brain state." *Physical Review E,* vol. 64(6), pp.061907, 2001.

[41] Supriya and Harun-Ur-Rashid. (2018). "Machine Learning Algorithms for Epileptic Seizures." [Online]. Dataset. Available: https://www.kaggle.com/harunshimanto/machine-learning-algorithms-for-epileptic-seizures.

IEEE Third International Conference on Data Stream Mining & Processing
August 21-25, 2020, Lviv, Ukraine

Student Performance Score Prediction Using Artificial Neural Network with the Support of Exploratory Factor Analysis and Clustering

Selçuk Öğütcü
Informatics Department
Kocaeli University
Kocaeli, Turkey
selcuk.ogutcu@gmail.com

***Abstract*— This study proposes a sequence of data analysis techniques for prediction purposes which consists of the use of exploratory factor analysis for classification, clustering of variables in a factor for focusing to a prediction goal and prediction of missing data via an artificial neural network using performances of engineering students' data. Average end-year scores of courses have been used to represent performance of a student. Prediction goal is the forecasting of a course score before the examination. Factor analysis has been used to group courses under individual factors. The factor in which course score to be predicted exists is clustered to group students with similar scores from various courses in that factor, excluding the course score to be predicted. An artificial neural network is trained for course scores in the cluster and target course score of a student is simulated in the artificial neural network for prediction after completion of training with an acceptable training error. The study showed a significant accuracy for prediction. In addition, the realized work has shown that it is possible to re-evaluate an education program, guide a student for better academic performance and provide support for employment via a combination of data mining methods. Python has been used to carry out the analysis.**

Keywords—factor analysis, clustering, neural network

I. INTRODUCTION

Artificial neural networks are designed to mimic the human brain. For an artificial neural network used for prediction, sensory inputs are processed via hidden layers consisting of artificial neurons to provide desired outputs. For this purpose, an artificial neural network is trained with inputs for desired outputs and this is accomplished by feeding back output errors due to comparisons with expected outputs and internal weight adjustments for connection between neurons. For processing large amounts of data, artificial neural networks may require more neurons and computational power. This leads to consumption of more time and resources. For some big data solutions, it becomes a necessity to support artificial neural networks with data mining methods.

Data mining with huge amount of data is a trending topic in healthcare [1]. Likewise, students attending education programs prodive huge data. There are uses of data mining techniques under certain titles such as pattern tracking, classification, clustering, regression, association and prediction. These techniques can be applied directly or in conjunction depending on goal of desired information to be extracted. Prediction and estimation are substantial components of data analysis goals. In order to extract significant outcomes and create strategies based on extracted information, it may be required to simplify multivariate data to its components (i.e., core variables) instead of processing it. When data is too big and/or complicated, an approach like factor analysis can be used to reduce set of variables in the dataset. Hence, dataset can be prepared for further processing to help information retrieval or extraction. In the paper, four years of performance data regarding courses students have taken for engineering education is used to analyze education structure, to cluster similar students and training a neural network (NN) with the corresponding cluster, for prediction.

In this study, 1499 students and 50 compulsory courses they attended have been processed. Engineering education consists of compulsory courses and elective courses. Graduated students are IT employee candidates. IT has various fields requiring different specialties. Education at universities is structured to cover all different fields. However, in general, update of traditional education programs in universities are slow to follow latest trends in IT sector. This enforces universities to review and update their courses and education programs accordingly. Therefore, a multi-disciplinary approach for decision making is a necessity.

In order to support decisions regarding education programs and guidance for students, an assessment and evaluation approach should be developed based on student performances.

II. COMBINED METHODOLOGY

A. Factor Analysis Method for Education Program Dataset

Exploratory factor Analysis (EFA) applied on multi-dimensional data depends on assumption of causality. When there are satisfactory evidences of cause and effect in data subject to information extraction, each step and feature of EFA becomes representation of underlying facts (latent variables) of a cause or in other words, base hypothesis. Thus, methods applied in EFA such as rotation becomes crucial. EFA extracts factors with the help of correlation. However, causality and correlation do not always overlap.

Student performances in the study that are based on year-end averages have been processed via EFA. For this purpose, 50 fundamental courses have been chosen. These courses are compulsory for all engineering education students in their four years of education. While some courses are offered by the social sciences, most of the courses are related to engineering,

978-1-7281-3214-3/20/$31.00 ©2020 IEEE

IT technologies and courses that require mathematics background.

Typically, there are five steps in a factor analysis [2];

- Determining any variable to be analyzed
- Testing the variable, the measurement could be done by considering the score of Kaiser Meyer Oikin (KMO), Bartlett Test of Sphericity as well as Measure Sampling Adequency (MSA)
- Conducting main process in factor analysis
- Conducting factor rotation process or rotating the formed factor
- Determining the group

In the dataset, column variables are courses and rows are performance scores with student ID indexes. All courses are subject to analyze.

Analysis of data should allow rejection of null hypothesis for this study. Therefore, relationships found between variables will depend on facts or in other words, evidence. To that end, Kaiser Meyer Oikin (KMO) and Bartlett Test of Sphericity tests have been applied to dataset. For a large dataset that includes multiple sub-relations, the entire dataset may not fit the model properly. This factor may cause a high chi-squared test result.

Main process of EFA requires determining of number of factors. As the first step, the correlation matrix for the dataset was calculated using Pandas DataFrame correlation function with Pearson method as in [3]. Using the 50x50 square-shaped correlation matrix, eigen values were obtained. Generally accepted method to determine number of factors is using the number of eigenvalues in the vector that are greater than 1. In the eigen values vector, 14 values were greater than 1 and so, this value has been used to define the number of factors.

Following the preliminary calculations and evidence collected to continue with the EFA, varimax rotation [4] has been chosen. When goal is to extract uncorrelated factors out of a dataset, in general, varimax rotation satisfies the desired results. There are rotation methods using different approaches. Two of these are varimax, and promax. These rotation methods may output different factor structures that each may be required for various purposes.

Unlike principal component analysis (PCA) which models the variance in the data in its entirety while treating noise and signal equally, factor analysis is used to identify common factors shared between variables but does not respond to independent noise or factors that belong to only one variable [5]. Thus, humanly interpretable factors can be acquired after the process. Processing data for assessment and evaluation of a subject like education where conclusions are subject to discussion, underlying factors found in a dataset should be compatible with human decision making.

For fitting the factor analysis model, maximum-likelihood, principal axis or minimal residual (minres) can be used. Among these fitting methods to use, minres has been chosen for the study because it shows monotonically decreasing of residual norms [6].

B. Clustering of an Extracted Factor

Following the factor analysis, distinguishing factors in the engineering education program have been retrieved. For prediction of a student success rate for a course, corresponding factor which the course is in should be determined. Next step is to cluster students in the factor in question using their performance values. Using the courses in the factor, a new dataset is created with students and their scores for courses. To predict a particular course for a student, the course to be predicted has been excluded from the students' dataset in the factor. After applying K-Means clustering, the student was moved in a relevant cluster.

The students in the cluster have been used to create a new dataset, after excluding the student with the course subject to prediction. Thus, another important step for neural network training and prediction has been accomplished. For clustering process, in order to fit dataset, an arbitrary number of iterations value has been used. Trials have shown that a maximum of 1000 iterations are adequate to create centroids with desired accuracy. Also, number of centroids has been set to 16 arbitrarily, based on trials. Fitted estimator and centroids have been gathered following the main process of K-Means algorithm. The main disadvantage of this algorithm is its sensitivity to the initial centroids and its convergence to local optima [7].

C. Neural Network Training for Prediction

With the support of EFA and K-Means clustering, the neural network had a significantly simplified data for training. Nonetheless, for a better training of neural network, trials shown that it is required to purify dataset from the noise before clustering process.

In the study, to predict a student's success rate with higher precision for a course in such dataset, outliers should have been filtered out. For each student, the average vector distance of seven input courses from dataset has been calculated. The same calculation has been made for target course. In an ideal case, these values should match. The student removal threshold has been set to 30 for the sub-dataset consists of these eight values. A higher value has shown negative impact on NN training while a lower value filtered out tolarable student data.

The NN is created accordingly to number of inputs required by number of courses in the factory with the course that is the prediction goal. The NN consists of five hidden layers and an output in addition to auto-modified input layer.

Trials for the study has shown that choosing gradient descent with momentum is more suitable for the dataset in question. It was not necessary to change the default value for momentum. Gradient descent with momentum is specially suited for noisy gradients [8]. Its convergence performance with accuracy has been better compared to Broyden–Fletcher–Goldfarb–Shanno (BFGS) algorithm.

III. Analysis Results and Discussion

A. Factor Analysis

Tests to be performed prior to further analysis can be seen in Table I. In Table II and Table III, sub-matrixes regarding corresponding factors derived from the main correlation matrix have been shown.

TABLE I. Pre-Analysis

Tests	Results	
	Subtest	***Value***
Kaiser-Meyer-Olkin Measure of Sampling Adequacy.	MSA	0.859
Bartlett's Test of Sphericity	Chi-Square	17827.175

Tests	Results	
	Subtest	*Value*
	Df	1498
	Sig.	0.000

TABLE II. SUB-CORRELATION MATRIX FOR FACTOR I

Correlation matrix for Factor I	Courses in Factor I								
	M1	***M2***	***M3***	***M4***	***M5***	***M6***	***E1***	***M7***	***E2***
M1	1.000	0.314	0.167	0.232	0.148	0.298	0.198	0.201	0.152
M2	0.314	1.000	0.393	0.294	0.267	0.457	0.277	0.377	0.251
M3	0.167	0.393	1.000	0.305	0.266	0.317	0.298	0.317	0.295
M4	0.232	0.294	0.305	1.000	0.143	0.319	0.213	0.357	0.321
M5	0.148	0.267	0.266	0.143	1.000	0.290	0.301	0.211	0.164
M6	0.298	0.457	0.317	0.319	0.290	1.000	0.319	0.324	0.292
E1	0.198	0.277	0.298	0.213	0.301	0.319	1.000	0.264	0.267
M7	0.201	0.377	0.317	0.357	0.211	0.324	0.264	1.000	0.402
E2	0.152	0.251	0.295	0.321	0.164	0.292	0.267	0.402	1.000

TABLE III. SUB-CORRELATION MATRIX FOR FACTOR III

Correlation matrix for Factor III	Courses in Factor III							
	E3	***M4***	***E4***	***C1***	***C2***	***C3***	***M8***	***E2***
E3	1.000	0.307	0.330	0.244	0.227	0.361	0.332	0.229
M4	0.307	1.000	0.404	0.359	0.286	0.446	0.403	0.321
E4	0.330	0.404	1.000	0.330	0.322	0.416	0.401	0.314
C1	0.244	0.359	0.330	1.000	0.371	0.388	0.290	0.316
C2	0.227	0.286	0.322	0.371	1.000	0.441	0.382	0.424
C3	0.361	0.446	0.416	0.388	0.441	1.000	0.488	0.396
M8	0.332	0.403	0.401	0.290	0.382	0.488	1.000	0.412
E2	0.229	0.321	0.314	0.316	0.424	0.396	0.412	1.000

Results in Table I show that chi-square value is high and p value equals to zero. This is expected due to size and complexity of the dataset. By the reason of measure of sampling adequacy (MSA) for the KMO test is around 0.86, overall MSA that is more than 0.5 suggests that there is enough correlation between the variables for a dimension reduction [9].

In the study, correlation threshold for factors has been set to 0.322 that is the lowest maximum value of factor rows. This allowed at least one factor with a single variable in list of factors providing the highest possible threshold for all factors. A part of correlation matrix specific to first of 14 factors is in Table II. Course names from M1 to M7 represent fundamental courses of math and physics. Courses E1 and E2 are field-specific engineering courses. Although at first glance it may be thought that E1 and E2 courses should be in different categories, it was confirmed that the content of the courses was closer to the courses in this factor compared to other engineering courses. Therefore, this factor can be considered as mathematical basics of the education program.

In Table III, courses C1, C2 and C3 are information technology lessons. There are three different types of courses in the table. E3 course includes sections as introductions from all other courses in the factor. For this reason, this factor can be considered as the technical fundamentals of the education program. M4 and E3 courses appear in both tables. This points to latent features of the four years education program. The other factors grouped similar courses more obviously. Coding, web design, language, social, career courses and such have been put in the relevant factors. When one or more course performances are subject to prediction, these factors hint on courses to be compared.

B. K-Means Clustering

After completion of EFA, course determined for prediction goal was searched in factors. This allowed a focused clustering via K-Means. For a dataset, a number of clusters of objects close to each other can be obtained [10]. In a factor consists of 9 variables (courses), distance calculations still depend on euclidean distance approach.

The expected outcome from the clustering process was clustering of the most similar-performance students. This is for ensuring that NN to be trained for prediction will have an adequately correlated and relatively small data compared to the main dataset. It can be considered as extraction of factors having entire dataset given without summarizing.

Dataset has been filtered by vector distances and split into two parts. The first part is the input for NN and the second part is the prediction goal. Each row's vector distance of NN inputs

from the entire dataset has been calculated in addition to prediction goals' absolute distance from all prediction goals in dataset. Rows with average absolute difference between these values exceeded an arbitrary threshold, such as 30, have been removed from the dataset.

Due to complexity of dataset correlation matrix as seen in Fig. 1., filtering and clustering focusing on a specific part of the dataset is required after EFA.

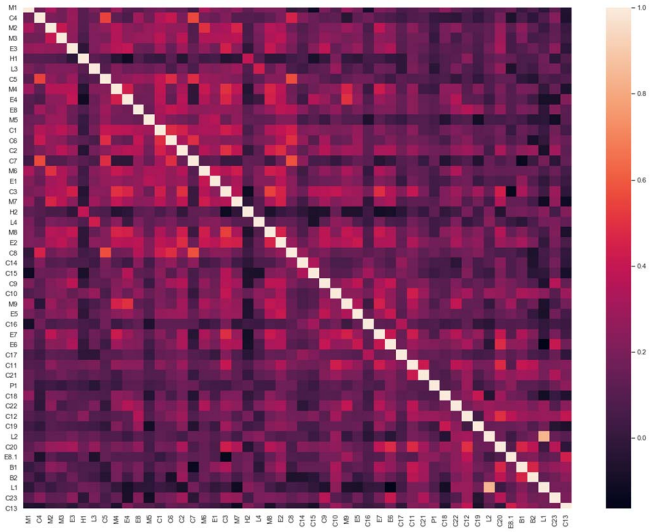


Fig. 1. Correlation matrix heatmap

Courses displayed in Fig. 1. have course-content relevant prefixes. These are;

- M for courses in mathematics or physics branches
- C for IT-related courses
- E for technical engineering courses
- L for language courses
- B for business sector related courses
- H for history courses
- P for projects

In the study, 50 students that have taken courses in Table III, except IT-related C3 course have been chosen. Course C3 has also been excluded from factor III dataset for creation of a new dataset. For ten students, min-max normalization and K-Means clustering have been applied to this dataset after appending corresponding student to the dataset, with a unique index. The cluster that included this specific index after clustering progress has been determined. Student has been excluded from the cluster and remaining relevant students' indexes have been used to create a new cluster with complete performance scores, from the original dataset. Thus, a cluster with the most relevant scores to the prediction goal has been created.

For clustering, trials with dataset with eight specific courses which includes 800 students has shown that 16 clusters were adequate for predictions. A future work has been noted for automating determination of number of clusters for obtaining the best results based on sub-dataset features. Fig. 2. shows the Python clustering code used for pre-NN training. This code snippet requires Scikit-learn and Numpy (np) libraries. The sub-dataset has been represented by Pandas library data frame object (df).

```
num_of_clusters=16
kmeans = KMeans(n_clusters=num_of_clusters,max_iter=1000).fit(df)
centroids = kmeans.cluster_centers_
labels = kmeans.predict(df)
```

Fig. 2. Python code used for clustering performance scores

C. Neural Network Training

The final phase of proposed analysis is the training of a NN with focused on cluster. To this phase, EFA has been used to summarize the entire dataset and K-Means clustering has shown the most relevant clusters of students for predictions. The NN was created with seven inputs and an output. Five hidden layers have been included in NN.

The NN created with Python's Neurolab (nl) in Fig. 3. has been trained with student performance scores in the created cluster. 10 iterations have been run and results have been put in list of predictions for each student. Goal was to provide a range of predictions for students.

```
nn_hiddenlayer_and_output_features=[32,16,8,4,2,1]
nn_input_features=[[-1,1],[-1,1],[-1,1],[-1,1],[-1,1],[-1,1],[-1,1]]
net = nl.net.newff(nn_input_features,nn_hiddenlayer_and_output_features)
net.trainf=nl.train.train_gdm
net.out_minmax=np.array([[-1.,1.]])
net.init()
```

Fig. 3. Properties of NN used for prediction

Table IV shows 10 of 50 prediction results. Column for course C3 shows current performance scores. Columns min and max show minimum and maximum predictions based on inputs courses. The last column shows whether the prediction is in the range of minimum and maximum prediction values. Four courses were in range while six were out of range.

Four of the current performance scores of column C3 were in prediction range while three of them were close to their ranges. Prediction results for student data that were not in dataset have pointed to the fact that there are more factors to be included into prediction. On the other hand, this study has shown that while working with large datasets, extraction of latent variables and clustering of relevant data supports understanding of features of cases.

The study has successfully determined 24 of the 50 students' performance scores according to their prediction ranges. For 26 performance scores out of their ranges;

- Maximum distance from predicted range is 28.19
- Minimum distance from predicted range is 0.31
- Average distance from predicted ranges is 12.1

The study has decreased the load of the NN by focusing to a smaller but the most relevant part of a large dataset, instead of working on the entire dataset. During trials with eight courses dataset that has 800 students' data, clustering has limited maximum number of students to be trained by NN around 80. A NN training iteration has been terminated if minimum and maximum predictions have not been changed above 0.05, three times. A loop which has been run for 10 times for each student has collected variety of these values.

TABLE IV. NEURAL NETWORK RESULTS

	Neural Network Inputs and Outputs										
	Students' performance scores								***Prediction***		
	E3	***C2***	***E2***	***M4***	***E4***	***M8***	***C1***	***C3 (T)***	***min***	***max***	***stat***
S1	58	52	76	68	72	73	62	44	59	72	false
S2	63	65	50	65	60	59	63	62	59	65	true

S3	59	24	54	52	48	36	48	48	26	38	false
S4	66	69	48	67	91	78	67	62	63	70	false
S5	62	62	64	55	47	43	71	58	53	60	true
S6	66	51	48	68	31	55	47	45	31	52	true
S7	70	43	40	80	82	73	62	82	66	77	false
S8	71	55	70	71	63	65	68	66	62	74	true
S9	71	54	57	54	73	61	56	44	54	62	false
S10	86	65	75	85	77	70	84	88	74	88	false

D. Evaluation of Methodology

This study is expected to give way to further and in-depth analysis of student data. Thus, it will be possible to form student profiles to be used for selection of projects and internships during their education and employment after graduation. Likewise, with the support of topic models and similar methods combined with the methods used in this paper, it is possible to improve academic guidance to match students' expectations in the light of their skills evaluated.

Outlier data has its importance as much as data appropriate for NN training when the analysis is about a subject like evaluation and summarization of an education program. Data filtered out that has irrelevant vector distances compared to the dataset points to problematic performance scores. The remaining noisy data has been clarified with NN training progress, with performance scores relatively far to min-max ranges provided, during prediction trials with more students.

Other than the human-ways of criticism, EFA shows the effective part of an education program by simplifying its contents (reducing dimensions), by performance scores analysis. Comparisons between educational departments of different disciplines and evaluations of differences between relevant departments of different educational institutions may reveal problems and clarify methods to be followed for a better education.

Predictions that are out of range are hints of the need for more attributes to be processed in addition to performance scores, for more precision regarding academic success. Some of the additional attributes have been discussed in detail in [11]. It is considered that students' performance rates should have a more priority in employability for more productivity. Especially under extraordinary conditions, some of the attributes directly affecting productivity become more important, reminding the importance of algorithmic fairness in [12].

At the prediction phase, in a loop that ran 10 times, each of the 50 students were clustered, removed from their clusters and the remaining students in their clusters with their complete performance scores were used in NN training. Following the NN training, performance scores of each of 50 students were applied to NN inputs. Thus, 10 NN outputs per each student have been acquired. Different clustering and NN training results have been obtained in each loop. The minimum and the maximum NN outputs per student have been recorded and considered as prediction range.

Fig. 4. displays NN predictions for course C3. 50 students' performances regarding seven courses used as NN inputs. Course C3 scores have been predicted between minpred and maxpred ranges. Some of the predictions significantly appeared out of range. Fig. 5. represents the algorithm of the study.

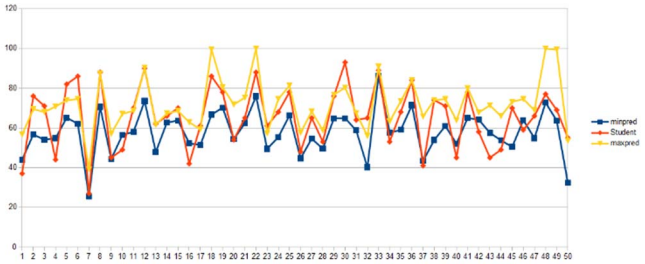


Fig. 4. Performance scores of 50 students with prediction ranges

```
Step 1.  fetch dataset
Step 2.  corrm=correlation_matrix(dataset)
Step 3.  eigen_values=eigvals(corrm)
Step 4.  number_of_factors=sum(eigen_values>1)
Step 5.  factors=factor-analyzer(dataset,number_of_factors)
Step 6.  rel_threshold=max(min(factors.rows))
Step 7.  factors=(factors>rel_threshold)
Step 8.  input target
Step 9.  factor=find(target,factors)
Step 10. sdataset=fetch(dataset[factor.columns])
Step 11. inputs=input students with n-1 course scores in factor for prediction target (nth score)
Step 12. define filter_threshold
Step 13. for row in sdataset
             compute row's n-1 scores' vector distances from sdataset (d1)
             compute row's nth value's differences from sdataset (d2)
             set row[avg_dist_diff] to average(abs(d1-abs(d2)))
Step 14. sdataset.droprow(sdataset[avg_dist_diff]>filter_threshold)
Step 15. initialize prediction_ranges as list
Step 16. initialize nn_properties as list
Step 17. for student in inputs
             initialize results as list
             initialize NN as neural_network with nn_properties
             for i in range(10)
                 clusters=cluster student[id] with sdataset vectors' n-1 scores
                 find student[id] in clusters
                 remove student from cluster
                 train NN for 100 epochs with sdataset's n-1 vectors for known nth value
                 result=NN.simulate(student)
                 append min(result) and max(result) to results
             append student[id], min(results[min_col]), max(results[max_col]) to prediction_ranges
```
Fig. 5. The algorithm used in the study

The importance of performance scores, or grades, is often a topic open to a continuous discussion[13]. During extraction of correlations between courses, some significant issues of interest have been surfaced. For example, if a student has a goal of getting employed in a job that requires data science knowledge, they should have the necessary knowledge regarding linear algebra. On the other hand, the relevance between data science related courses and mathematics courses including linear algebra does not satisfy the theoretical expectation. This fact has been noted for future work and this study has been performed for creating a measurable facts extraction for decision making.

IV. Conclusion

The current study has shown the importance of factor analysis in the evaluation and assessment of an education program. The factors represent groups of courses which are determinative on a computer engineer's qualities based on the actual education program. Using different EFA rotation methods on dataset has shown that it is possible to extract more information by enriching the dataset with more student and course features. This not only helps maximize student performances, but can also provide guidance for the selection of elective courses, the selection of projects to be involved in, the selection of companies for internships, and career options for post-graduate employment.

Clustering after EFA ensures focusing to a specific portion of dataset to meet desired outcomes such as prediction. NNs are known for their performance in information processing. Using multi layered NNs, it is possible to process complex information. But in some cases, it causes consumption of more resources and more time for meeting a goal. In order to overcome known issues of NNs such as overfitting, large dataset size or bad dataset segmentation, preparing dataset

appropriately before NN training is important as this study suggests.

Processing data with multilayered NNs provides a more direct method for prediction. In general, before training a NN, data is prepared in the favor of training performance. This study proposes an approach that provides options for deep analysis decisions with a data preparation process before NN training. Finally, trained NN gives the valuable implicit information that is useful for further decisions.

REFERENCES

[1] D. Imamovic, E. Babovic and N. Bijedic, "Prediction of mortality in patients with cardiovascular disease using data mining methods," 2020 19th International Symposium INFOTEH-JAHORINA (INFOTEH), East Sarajevo, Bosnia and Herzegovina, 2020, pp. 1-4.

[2] P. K. Sari, N. Nurshabrina and Candiwan, "Factor analysis on information security management in higher education institutions," 2016 4th International Conference on Cyber and IT Service Management, Bandung, 2016, pp. 1-5.

[3] Dharaneeshwaran, S. Nithya, A. Srinivasan and M. Senthilkumar, "Calculating the user-item similarity using Pearson's and cosine correlation," 2017 International Conference on Trends in Electronics and Informatics (ICEI), Tirunelveli, 2017, pp. 1000-1004.

[4] H. F. Kaiser, "The varimax criterion for analytic rotation in factor analysis", Psychometrika, vol. 23, pp. 187-200, 1958.

[5] D. Charles and C. Fyfe, "Kernel factor analysis with Varimax rotation," Proceedings of the IEEE-INNS-ENNS International Joint Conference on Neural Networks. IJCNN 2000. Neural Computing: New Challenges and Perspectives for the New Millennium, Como, Italy, 2000, pp. 381-386 vol.3.

[6] H. Kanayama, M. Ogino, S. Sugimoto and S. Terada, "Large-Scale Magnetostatic Domain Decomposition Analysis Based on the MINRES Method," in IEEE Transactions on Magnetics, vol. 49, no. 5, pp. 1565-1568, May 2013

[7] S. Banerjee, A. Choudhary and S. Pal, "Empirical evaluation of K-Means, Bisecting K-Means, Fuzzy C-Means and Genetic K-Means clustering algorithms," 2015 IEEE International WIE Conference on Electrical and Computer Engineering (WIECON-ECE), Dhaka, 2015, pp. 168-172.

[8] L. Zhang, P. Zhang, J. Yang, J. Li and Z. Gui, "Aperture Shape Generation Based on Gradient Descent With Momentum," in IEEE Access, vol. 7, pp. 157623-157632, 2019.

[9] T. Oladunni and S. Sharma, "Hedonic Housing Theory — A Machine Learning Investigation," 2016 15th IEEE International Conference on Machine Learning and Applications (ICMLA), Anaheim, CA, 2016, pp. 522-527.

[10] T. Zhang, W. Dong, H. Shi, R. Liu and H. Sun, "Design and implementation of teaching analysis system based on data mining," 2019 Chinese Control And Decision Conference (CCDC), Nanchang, China, 2019, pp. 4400-4405.

[11] C. D. Casuat and E. D. Festijo, "Identifying the Most Predictive Attributes Among Employability Signals of Undergraduate Students," 2020 16th IEEE International Colloquium on Signal Processing & Its Applications (CSPA), Langkawi, Malaysia, 2020, pp. 203-206.

[12] S. Verma and J. Rubin, "Fairness Definitions Explained," 2018 IEEE/ACM International Workshop on Software Fairness (FairWare), Gothenburg, 2018, pp. 1-7.

[13] G. M. Lundberg, B. R. Krogstie and J. Krogstie, "Becoming Fully Operational: Employability and the Need for Training of Computer Science Graduates," 2020 IEEE Global Engineering Education Conference (EDUCON), Porto, Portugal, 2020, pp. 644-651, doi: 10.1109/EDUCON45650.2020.9125188.

IEEE Third International Conference on Data Stream Mining & Processing
August 21-25, 2020, Lviv, Ukraine

Adaptive Matrix Model for a Crime Forecasting Task

Dmytro Uzlov
National Academy
of the National Guard of Ukraine
Kharkiv, Ukraine
poputcik@i.ua

Sergiy Popov
Control Systems Reasearch Laboratory
Kharkiv National University
of Radio Electronics
Kharkiv, Ukraine
serhii.popov@nure.ua

Oleksii Vlasov
Information and Analytical Support
Department of the Direktorate of
National Police of Ukraine
in Kharkiv region
Kharkiv, Ukraine
moonreactor@gmail.com

Yevgeniy Bodyanskiy
Control Systems Reasearch Laboratory
Kharkiv National University
of Radio Electronics
Kharkiv, Ukraine
yevgeniy.bodyanskiy@nure.ua

***Abstract*—In the police activities of any country, there are two major directions: identifying signs of crime preparation and preventing its commission (prediction), as well as crime prevention by eliminating the conditions for its commission (prevention). At the same time, various theories are explaining that the place and time of the crime occur at random, and when certain conditions are met for its commission, which depend on the type of crime and various factors of the objective side of the crime (place, time, mechanism of commission) and object of encroachment. Currently, a huge number of criminal events has been accumulated in the databases of law enforcement agencies over a long time (more than 20 years). These events occurred at a specific place (geolocation) and at a specific date and time. Considering a certain residential area as a closed system with processes taking place according to some laws, it can be assumed that criminal events also occur according to some hidden patterns. In modern science, some technologies enable identifying such hidden patterns in large data arrays – Data Mining. Identification of hidden patterns allows performing the function of predicting the commission of new criminal events in space and time. This article considers one of the technology approaches – Adaptive Matrix Model. As a result, we obtained a fairly simple and small model for crime prediction that produced promising experimental results. Ways to further improve accuracy are suggested.**

Keywords—crime prediction; spatiotemporal data mining; adaptive matrix model

I. INTRODUCTION

The task of predicting the commission of crimes in space and time is relevant and currently many software products solve this problem to varying degrees. Predictive policing is a means of forecasting and predicting crime using sophisticated techniques that require computer analysis. Police analytics software maps crimes and carries out analyses of crime-related data. Law enforcement officials utilize police analytics or crime analytics software to predict and forecast potential crimes. They then use those forecasts to properly distribute policing resources, target investigations and dispatch law enforcement officials. Police analytics software mines and sorts data, helping police departments uncover crime patterns and then use that information to drive policing strategy and resource distribution. These solutions import statistical, spatial, and temporal information, conduct its analysis and generate detailed reports.

Police analytics software is typically used in conjunction with other police software, including records management and dispatch functionality. To qualify for inclusion in the Police Analytics category, a product must:

- Import crime-related data from multiple sources;
- Produce analytical crime reports;
- Perform predictive crime analysis.

Examples of products are:

- Accurint Crime Analysis provides crime pattern analysis, predictive analytics, crime mapping and reporting to crime investigations and intelligence analysis by assisting you in pattern identification, data mining, analysis and prediction of serial events and more [1];
- PredPol uses a machine-learning algorithm to make its predictions. Historical event datasets are used to train the algorithm for each new city (ideally 2 to 5 years of data). PredPol then updates the algorithm each day with new events as they are received from the department [2];
- Verisk Crime Analytics uses data aggregation and management, analytical tools, and training to help customers predict, measure, deter, and respond to crime risks [3];
- RICAS – Realtime Intelligence Crime Analytics System is the intelligent system of criminal data analysis, which brings together in one view-space the most advanced techniques and methods of criminal analysis and analytical research in real-time, which can significantly improve the efficiency and effectiveness of crime detection in hot pursuit and previously unsolved crimes [4].

978-1-7281-3214-3/20/$31.00 ©2020 IEEE

Software products are based on technologies that can be expressed in the Table I below [5]. It summarizes predictive policing methods related to predicting crimes, i.e. identifying places and times that correspond to an increased risk of crime. As the table shows, conventional approaches start with mapping crime locations and determining (using human judgment) where crimes are concentrated (the hot spots). These approaches might include making bar graphs showing when crimes have occurred (time of day or day of the week) to identify "hot times." The corresponding predictive analytics methods start, at the most basic level, with regression analyses similar to what one would learn in an introductory statistics class and extend to cutting-edge mathematical models that are the subjects of active research. Some methods also attempt to identify the factors driving crime risk.

TABLE I. LAW ENFORCEMENT USE OF PREDICTIVE TECHNOLOGIES: PREDICTING CRIMES

Problem	Conventional Crime Analysis (low to moderate data demand and complexity)	Predictive Analytics (large data demand and high complexity)
Identify areas at increased risk Using historical crime data	Crime mapping (hot spot identification)	Advanced hot spot identification models; risk terrain analysis
Using a range of additional data (e.g., 911 calls, economics)	Basic regression models created in a spreadsheet program	Regression, classification, and clustering models
Accounting for increased risk from a recent crime	Assumption of increased risk in areas immediately surrounding a recent crime	Near-repeat modeling
Determine when areas will be most at risk of crime	Graphing/mapping the frequency of crimes in a given area by time/date (or specific events)	Spatiotemporal analysis methods
Identify geographic features that increase the risk of crime	Finding locations with the greatest frequency of crime incidents and drawing inferences	Risk terrain analysis

The key points in the search for hidden patterns of committed criminal events are:

- the presence of a certain closed system, namely, a certain territory on which certain number of people live, part of whom make these events;
- a long period of data accumulation in the format: event class, date and time, geo-coordinates;
- a long period of accumulation of metadata about the date.

An important part of the search for hidden patterns in the commission of crimes and their further prediction is the preparation of data. It is very important to have a uniform, noise-free data belonging to the same class of events. For this purpose, a preliminary classification of all events is carried out according to predefined signs. Such signs include circumstances and the mechanism for committing a crime. The commit mechanism may have a multi-level detail system. Accordingly, a class of homogeneous events can be sub-classified. Let's look at an example.

Take only one prepared class of events – theft with penetration into apartments or houses. The events were taken in a city with a population of more than 1 million people permanently residing in the territory. Maps of the distribution of events by geo-coordinates and temperature maps of the concentration of events by year (Fig. 1) show the existence of stable events concentration areas.

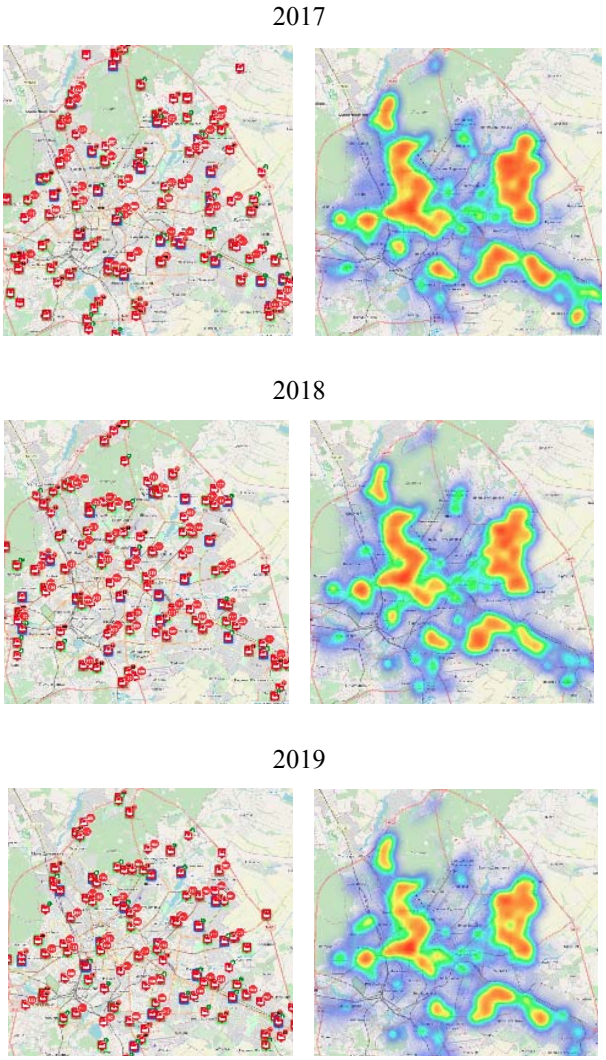


Fig. 1. Maps of events distribution by geo-coordinates and temperature maps of the events concentration by year

Then we divide the existing map of events by a grid with a cell of 300×300 m, in our case it will be a 64×64 matrix (Fig. 2). One can take any grid and accordingly a larger or smaller matrix will be formed. The number of events that occurred in the territory covered by the corresponding cell

will be assigned to each element of the matrix. There will be no undefined values, some elements will be zero.

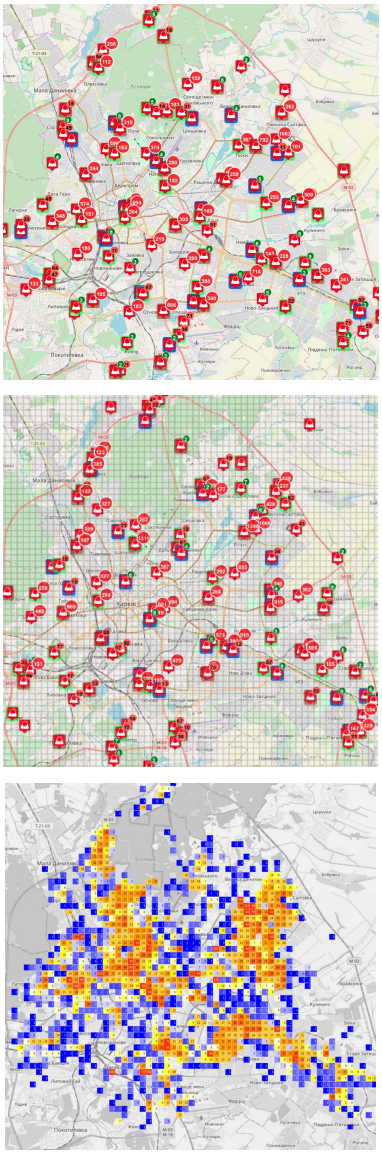


Fig. 2. Map of events distribution, same map with a grid, and color-coded matrix representation of the same map

Considering a single cell, we can display the diagrams of absolute values and the dynamics of events in the cell for different months and days of week.

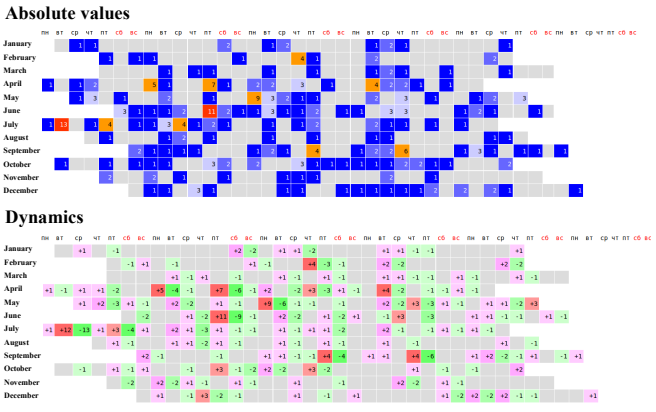


Fig. 3. Absolute values of events distribution over days of week and its dynamics

In this study, we are going to use the 64×64 matrix representation of the events distribution data which we call "observation fields". In general, rectangular fields of any size (not just square ones) can be considered.

II. Adaptive Matrix Model Approach

State matrices will be used as a mathematical description of the sequence of observation fields

$$X_n = \left(x_{ij,n}\right),\ i = 1,2,\cdots,M,\ j = 1,2,\cdots,N,$$

where M and N – number of rows and columns of matrix X_n respectively, $n = 0,1,2,\ldots$ – current discrete time.

As a hypothesis about the mechanism of a sequence X_n generation, the operation of linear transformation of matrix X into matrix Z of the same size was introduced in [6] as

$$Z = A * X = \left(\sum_{j=1}^{N} A_{1j} X_j \;\middle|\; \sum_{j=1}^{N} A_{2j} X_j \;\middle|\; \cdots \;\middle|\; \sum_{j=1}^{N} A_{Nj} X_j \right), \quad (1)$$

where X_j – j^{th} column of matrix X, A_{ij} – $M \times M$ transformation matrices, unknown in a general case.

Transformation (1) using the operations of cyclic rearrangement and transplantation [6] or vectorization and devectorization [7] can be reduced to the standard vector-matrix form. Vectorizing matrix X in the form $\overset{+}{X}{}^T = \left(X_1^T, X_2^T, \cdots, X_N^T\right)$, introducing the block transformation matrix

$$A = \left(\begin{array}{c|c|c|c} A_{11} & A_{12} & \cdots & A_{1N} \\ \hline A_{21} & A_{22} & \cdots & A_{2N} \\ \hline \vdots & \vdots & \ddots & \vdots \\ \hline A_{N1} & A_{N2} & \cdots & A_{NN} \end{array} \right)$$

(here A_{ij} – $M \times M$ transformation matrices in (1)) and the product

$$\left(\overset{+}{A}\,\overset{+}{X}\right)^T = \left(\left(\sum_{j=1}^{N} A_{1j} X_j \right)^T, \left(\sum_{j=1}^{N} A_{2j} X_j \right)^T, \cdots, \left(\sum_{j=1}^{N} A_{Nj} X_j \right)^T \right),$$

after devectorizing we get

$$Z = \left(\overset{+}{A}\,\overset{+}{X}\right) = A * X, \quad (2)$$

where $\overset{+}{(\bullet)}$ denotes vectorization, $\underset{+}{(\bullet)}$ – devectorization.

Since real observation fields are almost always distorted by various kinds of disturbances, a matrix analogue of a first-order Markov sequence was introduced in [6] as

$$X_{n+1} = A * X_n + W_{n+1}$$

(here $W_n = \left(w_{ij,n}\right)$ – discrete matrix white noise), which, taking into account (2), can be represented in the form

$$X_{n+1} = \left(A \overset{+}{X}_n \right)^{+} + W_{n+1}. \quad (3)$$

If the parameters of the transformation A are unknown, we can solve the problem of estimating them in real time using some adaptive identification procedures. Matching (3) with a tuned filter

$$\hat{X}_{n+1} = \left(A_n \overset{+}{X}_n \right)^{+}, \quad (4)$$

it's easy to write the adaptive Widrow-Hoff algorithm [8]

$$A_{n+1} = A_n + \alpha \frac{\overset{+}{X}_{n+1} - A_n \overset{+}{X}_n}{\overset{+}{X}_n^{T} \overset{+}{X}_n} \overset{+}{X}_n^{T}, \quad (5)$$

where A_n – current estimates of transformation A parameters, $0<\alpha<2$ – a scalar parameter of the algorithm.

It should be noted that in practical applications this approach is ineffective due to the large number of unknown parameters. For example, in our simple case with $M = N = 64$, it is necessary to estimate in real time $(MN)^2 = 16777216$ unknown parameters of matrix A. This circumstance motivates to look for alternative approaches to solving the problem of observation fields forecasting. For that purpose we introduce a simplified transformation

$$Z = A * X = \left(\sum_{j=1}^{N} b_{j1} A X_j \;\middle|\; \sum_{j=1}^{N} b_{j2} A X_j \;\middle|\; \cdots \;\middle|\; \sum_{j=1}^{N} b_{jN} A X_j \right), \quad (6)$$

corresponding to a two-dimensional field description in the form

$$X_{n+1} = A X_n B + W_{n+1}, \quad (7)$$

where A and B are ($M\times M$) and ($N\times N$) transformation matrices. Description (7) contains $M^2 + N^2$ parameters, and for $M> 2$, $N> 2$ it is obvious that $M^2 + N^2 < (MN)^2$, and the gain in the number of estimated parameters is greater with larger M and N. For example, for the considered case of $M = N = 64$ the number of transformation parameters in (7) is 8192, which is significantly less than in model (4).

Algorithm (5) due to a specific structure of (7) cannot be used for direct calculations and requires significant modification. In principle, description (7) can be converted to a matrix-vector form

$$\overset{+}{X}_{n+1} = (B^T \otimes A) \overset{+}{X}_n + \overset{+}{W}_{n+1}$$

(here $\otimes$ is the tensor product symbol), however, the number of parameters in the matrix $B^T \otimes A$, as in case (3), is $(MN)^2$.

Hence we come to a problem of synthesizing algorithms for estimating the unknown parameters of matrices A and B in (7) and forecasting of observation fields using the obtained estimates.

Let us match (7) with a custom filter of the form

$$\hat{X}_{n+1} = A_n X_n B_n, \quad (8)$$

introduce three types of errors that occur during the tuning process

$$\begin{cases} V_{n+1} = X_{n+1} - A_n X_n B_n, \\ V^A_{n+1} = X_{n+1} - A_{n+1} X_n B_n, \\ V^B_{n+1} = X_{n+1} - A_{n+1} X_n B_{n+1}, \end{cases}$$

and two criteria characterizing the forecasting quality

$$\begin{cases} J^A_{n+1} = Tr\, V_{n+1} V^T_{n+1} = \left\| V_{n+1} \right\|^2, \\ J^B_{n+1} = Tr\, V^A_{n+1} V^{A\,T}_{n+1} = \left\| V^A_{n+1} \right\|^2, \end{cases} \quad (9)$$

where Tr is the matrix trace symbol.

Each step of the estimation process consists of two sub-steps: tuning of matrix A_{n+1} based on error V_{n+1} and criterion J^A_{n+1} and subsequent tuning of matrix B_{n+1} based on error V^A_{n+1} and criterion J^B_{n+1}. A gradient tuning algorithm can be written as

$$\begin{cases} A_{n+1} = A_n - \gamma^A_{n+1} \nabla^A_{n+1} = A_n - \gamma^A_{n+1} V_{n+1} B^T_n X^T_n, \\ B_{n+1} = B_n - \gamma^B_{n+1} \nabla^B_{n+1} = B_n - \gamma^B_{n+1} X^T_n A^T_{n+1} V^A_{n+1}, \end{cases} \quad (10)$$

where $\gamma^A_{n+1}, \gamma^B_{n+1}$ – scalar algorithm gain coefficients, $\nabla^A_{n+1} = V_{n+1} B^T_n X^T_n$, $\nabla^B_{n+1} = X^T_n A^T_{n+1} V^A_{n+1}$ – ($M\times M$) and ($N\times N$) matrices formed by derivatives of criteria J^A_{n+1} and J^B_{n+1} with respect to the tuned parameters.

It can be shown that the maximum parameter tuning rate is ensured by a special choice of gain coefficients $\gamma^A_{n+1}, \gamma^B_{n+1}$ where (10) takes the form

$$\begin{cases} A_{n+1} = A_n + \dfrac{\left\|\nabla^A_{n+1}\right\|^2}{Tr\nabla^A_{n+1}X_nB_nB_n^TX_n^T\nabla^{A\ T}_{n+1}}\nabla^A_{n+1}, \\ B_{n+1} = B_n + \dfrac{\left\|\nabla^B_{n+1}\right\|^2}{Tr\,A_{n+1}X_n\nabla^B_{n+1}\nabla^{B\ T}_{n+1}X_n^TA_{n+1}^T}\nabla^B_{n+1}, \end{cases} \quad (11)$$

which is a generalization of the optimal (α=1) Widrow-Hoff algorithm to the case of a two-dimensional field.

Evolution of forecasting errors during the operation of algorithm (11) is described by difference equations

$$\begin{cases} V^A_{n+1} = V_{n+1} - \dfrac{\left\|\nabla^A_{n+1}\right\|^2 X_nB_n\nabla^A_{n+1}}{Tr\nabla^A_{n+1}X_nB_nB_n^TX_n^T\nabla^{A\ T}_{n+1}}, \\ V^B_{n+1} = V^A_{n+1} - \dfrac{\left\|\nabla^B_{n+1}\right\|^2 A_{n+1}X_n\nabla^B_{n+1}}{Tr\,A_{n+1}X_n\nabla^B_{n+1}\nabla^{B\ T}_{n+1}X_n^TA_{n+1}^T}, \end{cases} \quad (12)$$

in this case it can be shown that $\left\|V^B_{n+1}\right\|^2 \le \left\|V^A_{n+1}\right\|^2 \le \left\|V_{n+1}\right\|^2$, i.e. algorithm (11) provides a monotonous decrease in errors in the process of parameters tuning.

The practical use of algorithm (11) can be complicated by the fact that in the vicinity of optimal modes when the errors V_{n+1} and V^A_{n+1} are small, the values of gain coefficients γ^A_{n+1} and γ^B_{n+1} increase sharply which can lead to computational instability. This can be avoided by introducing the following modifications to algorithm (10):

- additive

$$\gamma^{A\ A}_{n+1} = \frac{\left\|\nabla^A_{n+1}\right\|^2}{\gamma^A + Tr\nabla^A_{n+1}X_nB_nB_n^TX_n^T\nabla^{A\ T}_{n+1}},$$
$$\gamma^{B\ A}_{n+1} = \frac{\left\|\nabla^B_{n+1}\right\|^2}{\gamma^B + Tr\,A_{n+1}X_n\nabla^B_{n+1}\nabla^{B\ T}_{n+1}X_n^TA_{n+1}^T}$$

- multiplicative

$$\gamma^{A\ M}_{n+1} = \frac{\gamma^A_M}{\left\|X_nB_n\right\|^2}, \quad \gamma^{B\ M}_{n+1} = \frac{\gamma^B_M}{\left\|A_{n+1}X_n\right\|^2}.$$

Here γ^A, γ^B are regularizing terms, $0 < \gamma^A_M < 2, \quad 0 < \gamma^B_M < 2$.

The process of errors decrease when using algorithm (11) is limited from below to a value determined by the characteristics of interference W_{n+1}. If $M\{W_{n+1}\} = 0, \quad M\left\{\overset{+}{W}_{n+1}\overset{+\ T}{W}_{n+1}\right\} = P_W < \infty I$ (here $M\{\bullet\}$ is the mathematical expectation symbol), then the error is limited by the value

$$\left\|V^{\min}\right\|^2 = Tr(B^T \otimes A + I)P_W(B^T \otimes A + I)^T.$$

A matrix analogue of the autoregression equation can also be considered

$$X_{n+1} = \sum_{h=0}^{r-1} A^h * \left(q^{-h}X_n\right) + W_{n+1}, \quad (13)$$

where q^{-h} – backward shift operator defined as $q^{-h}X_n = X_{n-h}$, $A^h, h = 0,1,\cdots,r-1$ – matrix convolution operators of type (1). The total number of parameters of such a model is $(MN)^2 r$.

The number of unknown parameters of the forecasting model can be significantly reduced by using a simplified transformation instead of structure (13). In this case either this structure can be used

$$X_{n+1} = \tilde{A}\left(X_n \mid X_{n-1} \mid \cdots \mid X_{n-r+1}\right)\tilde{B} + W_{n+1}, \quad (14)$$

containing $M^2 + rN^2$ parameters (here $\tilde{A}$ and $\tilde{B}$ – ($M\times M$) and ($rN\times N$) transformation matrices) or

$$X_{n+1} = \sum_{h=0}^{r-1} A^{h+1}X_{n-h}B^{h+1} + W_{n+1}, \quad (15)$$

containing $\left(M^2 + N^2\right)r$ parameters.

Introducing matrices

$$\tilde{A} = \left(A^1 \mid A^2 \mid \cdots \mid A^r\right),$$
$$\tilde{B} = \begin{pmatrix} B^1 \\ \hline B^2 \\ \hline \vdots \\ \hline B^r \end{pmatrix}, \quad \tilde{X}_n = \left(\begin{array}{c|c|c|c} X_n & 0 & \cdots & 0 \\ \hline 0 & X_{n-1} & \cdots & 0 \\ \hline \vdots & \vdots & \ddots & \vdots \\ \hline 0 & 0 & \cdots & X_{n-r+1} \end{array}\right),$$

(15) can be rewritten in a compact form

$$X_{n+1} = \tilde{A}\tilde{X}_n\tilde{B} + W_{n+1}. \quad (16)$$

Then we match (16) with the tuned predictor

$$\hat{X}_{n+1} = \tilde{A}_n\tilde{X}_n\tilde{B}_n,$$

which allows to reduce the problem to the one already considered above and solve it using the matrix analogue of the Widrow-Hoff algorithm.

III. EXPERIMENT

To check operability of the proposed approach, we try to forecast the described above 64×64 observation fields sequence corresponding to events distribution maps. Considering daily and weekly patterns in crime distribution

data, we apply autoregressive model of type (16) with the following block matrix structure

$$\tilde{A} = \left(A^1 \mid A^2 \mid A^3 \right),$$

$$\tilde{B} = \begin{pmatrix} B^1 \\ \hline B^2 \\ \hline B^3 \end{pmatrix}, \quad \tilde{X}_n = \left(\begin{array}{c|c|c} X_n & 0 & 0 \\ \hline 0 & X_{n-1} & 0 \\ \hline 0 & 0 & X_{n-6} \end{array}\right).$$

This model has $\left(64^2 + 64^2\right) * 3 = 24576$ tuned parameters. We use 1 year (365 fields) of historical data. As the proposed adaptive approach is basically of the online type (parameters tuning is performed after each forecasting step), we don't need to divide data into training and test sets. As we can see from the errors graph (fig. 4), algorithm (11) ensures quite fast convergence of the tuned weights.

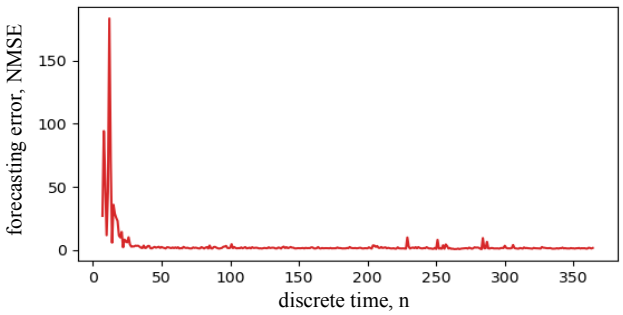


Fig. 4. Errors evolution during the model tuning

Fig. 5 shows an example forecast. It can be seen that despite not all points' locations coincide, general patterns are forecasted quite accurately. This allows adjusting police patrols distribution to match the predicted patterns, which increases probability of crime prevention or quicker reaction to ones that would actually happen in the "hot" areas.

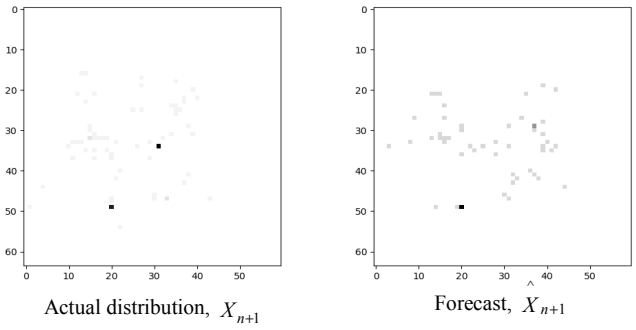


Fig. 5. Forecasting example

IV. Conclusions

In this paper, we introduced an adaptive matrix model approach to crime events forecasting. Experiments show promising results and suggest ways of improving forecasting accuracy by e.g. taking into account local patterns and external events.

References

[1] "Accurint® Crime Analysis", https://risk.lexisnexis.com/products/accurint-crime-analysis

[2] "PredPol", https://www.predpol.com/

[3] "Verisk Crime Analytics and NICB Assist in Criminal Investigation", https://www.verisk.com/archived/2015/july/verisk-crime-analytics-and-nicb-assist-in-criminal-investigation/

[4] "RICAS - Realtime intelligence crime analytics system", https://ricas.org/

[5] W. L. Perry, B. McInnis, C. C. Price, S. C. Smith, J. S. Hollywood, Predictive policing the role of crime forecasting in law enforcement operations, RAND Corporation, 2013.

[6] V.M. Kuntsevich "On the solution of two-dimentional discrete filtering problem", in Avtomatika i telemekhanika, N. 6, 1987, pp 68-78.

[7] S. Mashtalir, V. Mashtalir "Spatio-Temporal Video Segmentation. in Advances", in Spatio-Temporal Segmentation of Visual Data. Editors: V. Mashtalir, I. Ruban, V. Levashenko, Springer, 2020, pp. 161-210.

[8] S. Haykin, Neural networks and learning machines (3rd edition), Prentice Hall, 2009.

IEEE Third International Conference on Data Stream Mining & Processing
August 21-25, 2020, Lviv, Ukraine

Intelligent Forecasting System for NPP's Energy Production

Olena Chornovol
Intelligent Information Systems Department
Petro Mohyla Black Sea National University
Mykolaiv, Ukraine
lenchornovol@gmail.com

Galyna Kondratenko
Intelligent Information Systems Department
Petro Mohyla Black Sea National University
Mykolaiv, Ukraine
halyna.kondratenko@chmnu.edu.ua

Ievgen Sidenko
Intelligent Information Systems Department
Petro Mohyla Black Sea National University
Mykolaiv, Ukraine
ievgen.sidenko@chmnu.edu.ua

Yuriy Kondratenko
Intelligent Information Systems Department
Petro Mohyla Black Sea National University
Mykolaiv, Ukraine
yuriy.kondratenko@chmnu.edu.ua

***Abstract*— This paper examines existing intelligent forecasting approaches, in particular, for solving the problem of evaluating Nuclear Power Plant's (NPP's) energy production. Artificial intelligence (AI) and machine learning (ML) techniques contribute to energy consumption forecasting models. Such models considerably improve the accuracy, reliability, and precision contributing to conventional time series forecasting tools. NPPs do physically produce more or less of the demanded amount of energy. The process of calculating imbalances requires a comparison of electricity purchased or sold under contract with the results of commercial metering of physical production and consumption. This study orients to choose the rational approach by reviewing different ML models for energy prediction. Advanced analytics and AI-enabled algorithms can help identify off-line behaviors to increase efficiency and help balance supply and demand. Better short-term forecasts can improve power planning by enabling operators to both reduce their dependence on polluting stations and activate an increase in the number of alternating sources. Adjusting the amount of energy produced will lead to energy conservation and an improved environmental situation.**

Keywords*— *artificial intelligence, machine learning, energy consumption, nuclear energy efficiency, deep learning, artificial neural network

I. INTRODUCTION

There are three main energy development trends called decentralization, decarbonization, and digitalization (also known as Energy 4.0). Decentralization intends moving away from monopolistic energy markets to improve energy efficiency, implement energy saving measures, and introduce modern methods of energy management and audit [1-3]. Regional grid operators generate energy demand forecasts based on historical information and environmental factors (such as temperature). They then compare forecasts with available resources (e.g. coal, natural gas, nuclear, solar, wind, hydro). Electricity generation technologies highly depend on environmental conditions and subject to routine and unscheduled maintenance [4].

Since energy comes from three different resources like fossil fuels, renewable and nuclear resources [5], it is required to keep tracking of energy consumption in terms of types in different areas in order to make energy consumption plans, based on a specific usage and area. For all energy types mentioned above, estimating the usage is essential for agile energy management in NPP. Optimized ML-based forecasts of the energy demand will lead to energy savings and improve the environmental situation [6].

Nowadays the energy sector requires more attention for smart energy consumption tracking. Using energy predictive models, authors can predict the amount of energy, which is consumed in different areas and make plans for energy efficiency events. Forecasting the energy usage both in short-term and long-term manner helps to understand the sustainable way of energy use. Having multiple factors, such as water, wind, temperature, predicting energy consumption requires model to be accurate and span multiple horizons in time and space to better quantify support uncertainty use cases. [7]. Nowadays ML models are widely applied to multiple industries since they behavior as a function which best map the input data to output [8-10]. ML models deploy prediction tasks for energy consumption with high accuracy [11].

II. RELATED WORKS AND PROBLEM STATEMENT

ML and artificial neural network (ANN) techniques have recently been very helpful in promoting energy consumption forecasting models [12-14]. Future ML algorithms will require meaningful inclusion of specific domain data [2, 3, 15-18]. In practice, traders can buy more or less electricity than they sold, and manufacturing companies can physically produce more or less than use in fact. The process of calculating imbalances requires comparison of purchased or sold electricity under contract with the results of commercial metering of physical production and consumption volumes. One of the tasks is to determine the price and adjust the amount of energy produced, which in turn will lead to energy conservation and an improved environmental situation.

978-1-7281-3214-3/20/$31.00 ©2020 IEEE

Traditionally, an autoregression method is used to forecast electricity consumption: a linear algorithm based on the forecast of the future to the immediate past [19]. It is possible to build separate autoregressive models for different types of days. However, all other factors affecting energy consumption are used indirectly because of their influence on past values of electricity consumption. This method allows getting good forecast results in stable situations [20]. However, with an unexpected sharp change in external parameters, the application of this approach does not allow to predict the situation correctly. For example, in the event of a severe cold, even if the approximate weather forecast is known, the method does not allow it to be used.

The purpose of the study is to predict energy efficiency using real-time atomic electricity using ML methods [6]. The subject of the study is the analysis of selected machine models for energy prediction: multiple linear regression (MLR), Support vector regression (SVR), Random Forest and Extreme gradient boosting (XGBoost). Scientific novelty is based on the study of non-traditional machine learning models using hyper parameterization.

Authors calculate energy efficiency as a time series problem. For this reason, there have been implemented time series methods such as ARIMA from *statsmodels* and Seasonal Decomposition using the Prophet library. Prophet is an open source library published by Facebook based on a time series model (trend + seasonality + holidays). It allows you to make time series forecasts with good precision using simple intuitive parameters and supports the inclusion of special seasonality and holidays.

When the forecasting model is not working as planned, authors want to be able to adjust the method parameters to suit the specific problem. Setting up these methods requires a deep understanding of how basic time series models work [21]. For example, the first input parameters for automated ARIMA are maximum resolutions, autoregressive components, and average components. A typical analyst will not know how to adjust these orders to avoid behavior, and this is an experience that is difficult to acquire and expand. The Prophet Library offers intuitive, easy-to-configure options.

Using time as a regressor, Prophet tries to set several linear and nonlinear functions of time as components [22]. Seasonality modeling is the same approach used by exponential smoothing in the Holt-Winters technique [23]. In fact, authors set the prediction problem as an exercise in setting curves, rather than looking closely at the dependence of each time series observation. The trend is modeled by fitting a detailed linear curve over the trend or a non-periodic part of the time series. The linearity of the series shows the minimal impact of data change.

III. Classification of the Integrated ML Models

Support Vector Regression (SVR) is a ML algorithm that can be used to solve both classification and regression problems [24]. However, it is mostly used in classification problems. In this algorithm, each data element is constructed as a point in an n-dimensional space (where n is the number of functions), and the value of each function is the value of a specific coordinate. Then authors classify by finding a hyperplane that divides the two classes very well (Fig. 1). Support vectors are simply the coordinates of an individual observation. The vector support machine is the boundary that best separates the two classes (hyperplane/line).

The SVR problem statement is often derived from a geometric point of view. The approximate function of continuous value can be written as in (1). For multidimensional data, you increase x by one and include b in the w-vector for simple mathematical notation and get multivariate regression in (2).

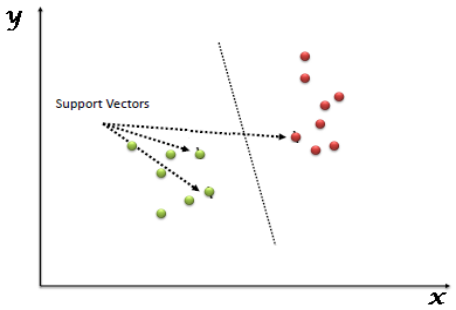


Fig. 1. Classification of data by SVR

The SVR formulates the problem of approximating this function as an optimization problem that tries to find the narrowest tube centered on the surface, minimizing the prediction error, that is, the distance between the predicted and desired results:

$$y = f(x) = \langle w, x \rangle + b = \sum_{j=1}^{M} w_j x_i + b,\ y, b \in R,\ x, w \in R^M. \quad (1)$$

$$f(x) = \begin{bmatrix} w \\ b \end{bmatrix}^{\tau} \begin{bmatrix} x \\ l \end{bmatrix} = w^{\tau} x + b x,\ w \in R^{M+1}. \quad (2)$$

The former condition produces an objective function $\min_{w} \frac{1}{2}\|w\|^2$, where $\|w\|$ is the value of the normal vector close to the surface.

To imagine how the magnitude of the weights can be interpreted as a measure of the plane, let us consider the following example:

$$f(x, w) = \sum_{j=1}^{M} w_j x^i,\ x, w \in R^M, \quad (3)$$

where M is the polynomial order used to approximate the function. As the magnitude of the vector w increases, more w_j is nonzero, which leads to higher order solutions.

Currently, *Random Forest* model is very common and versatile for most application tasks [25]. With its help it is possible to solve both the problem of division of objects into two or more classes (the task of classification), and to predict the material response for each object (the task of regression). By completing a composition of several build trees, combining the answers of each of them, you can get a stable and much better solution than many other algorithms.

Consider the labeled sample of objects $\{(x_i, y_i)\}_{i=1}^{N}$, where $x_i \in R^2$ is a description of an object in two-dimensional space, and $y_i \in \{0,1\}$ is a class label (Fig. 2).

Despite the fact the objects of different classes are strongly mixed in the middle, using a decision tree with such a sample is quite convenient to work, at each step, it is necessary to select the sign and threshold values at which the optimal by the given criterion of partition occurs [13]. The following criteria are often used when solving application problems.

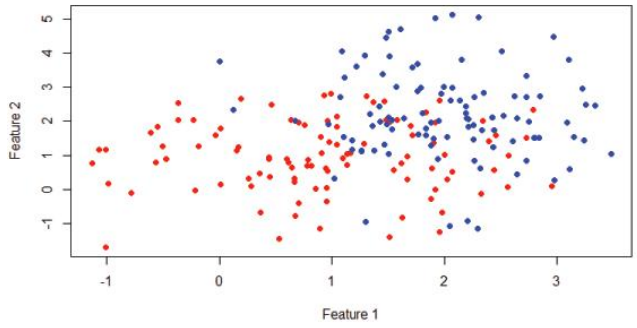


Fig. 2. The output of the classification task

For *iGain* classification tasks:

$$\sum_{v\in\{L,R\}}\frac{|S_v|}{|S|}H(S_v),\ \mathrm{H(S)}=-\sum_{c\in C}p_c\log_2(p_c),, \tag{4}$$

where C is the set of classes of this problem, and p_c is the probability of class c for the set of objects S.

For regression problems, a similar *iGain* criterion using variance:

$$iGain(S)=|S|Var(S)-\sum_{v\in\{L,R\}}|S_v|Var(S_v), \tag{5}$$

where *Var (S)* is the variance of the responses of objects from the set S.

At each division, all objects are divided into two smaller groups. Specifying the maximum number of objects sets one of the possible stopping criteria for the algorithm at the top-leaf of a tree.

The tree resulting from the algorithm can be represented in several ways. Thus, it is possible to fairly qualify the considered sample of objects with only one decision tree, if as the answer for the test object, which got into the cell, to give the number most common in this cell of the class.

Multiple Linear Regression (MLR) is an extension of ordinary least squares (OLS) regression that includes more than one explanatory variable [26, 27]. Authors use the Standard Scaler for standardization for linear regression. Simple least squares regression (SLR) is a statistical analysis method that evaluates the relationship between one or more independent variables and a dependent variable; the method estimates the ratio by minimizing the sum of squares in the difference between the observed and predicted values of the dependent variable, configured as a straight line. In this introduction, OLS regression will be discussed in the context of a bivariate model, that is, a model in which there is only one independent variable (X) that implies a dependent variable (Y). However, OLS regression logic easily extends to a multivariate model that has two or more independent variables. The formula for multiple linear regression is:

$$y_i=\beta_0+\beta_1x_{i1}+\beta_2x_{i2}+\ldots+\beta_px_{ip}+\varepsilon, \tag{6}$$

where y_i is the dependent variable, x_{ij} is the expansion variable, β_0 is the constant term, β_p is the slope coefficients for each expansion variable, and ε is the model error term (also known as residuals).

Simple linear regression allows analysts to make predictions about one variable based on information that is known about the other variable. The multiple regression model is used to validate the data, for training the authors use SVR models, Random Forest, and XGBoost. In addition to multiple regression, authors validate the data with the k-fold cross and random split libraries to predict the model's suitability for the hypothetical test suite in the training step.

Extreme Gradient Boosting Method (XGBoost) is one of the most popular and effective implementations of the gradient busting algorithm on trees. It originally started as a terminal application that could be configured using the libsvm configuration file [5]. After winning the Higgs Machine Learning Challenge, the model became well known in ML racing circles.

XGBoost is based on the algorithm of gradient boosting of decision trees. The boosting method is a ML technique for classification and regression problems that builds a prediction model in the form of an ensemble of weak prophetic models, usually decision trees [21]. Training of the ensemble is carried out consistently unlike, for example, from bagging [28]. At each iteration, the deviations of the predictions of the already trained ensemble in the training sample are calculated.

The behavior of the model at one point in the abstract linear regression problem is shown on Fig. 3 [29]. Suppose that the first model of ensemble F always produces a sample mean of f_0 value. This forecast is rather rough, meaning the standard deviation at the selected point will be quite large. Correcting this by teaching the model Δ_1 will "correct" the prediction of the previous ensemble F_0. In this way, authors get the ensemble F_0, where prediction will be summed up by model predictions f_0 and Δ_1. Continuing this sequence, come to the ensemble F_4 of the prediction, which is summed up from the forecasts f_0, Δ_1, Δ_2, Δ_3, Δ_4 and predicts exactly the value of the given target.

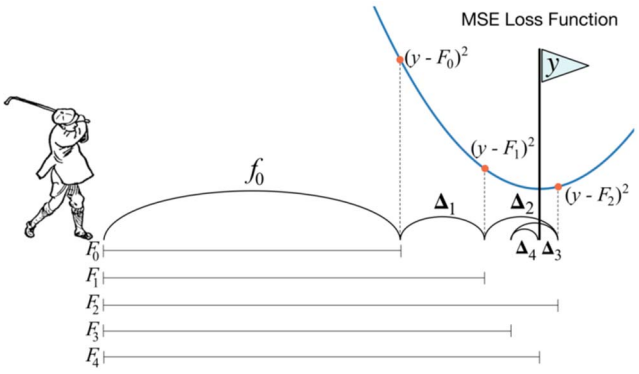


Fig. 3. Illustration of Extreme Gradient Boosting Method [29]

The most important factor in the success of XGBoost is its scalability in all scenarios. The system runs more than ten

times faster than existing popular single-machine solutions, scaling billions of examples in distributed or limited storage settings. The scalability of XGBoost is driven by several important systems and algorithmic optimizations. These innovations include a new tree-learning algorithm designed to handle liquid data, a theoretically sound weighted quantile sketch procedure allows handling the weights of specimens in a close examination of the tree. Parallel and distributed computing make learning faster, allowing you to learn the model faster.

XGBoost supports all possible libraries such as scikit-learn with the ability to add regularization. There are three supported forms of gradient boosting: standard gradient booster with shoe speed capability (training speed); stochastic gradient boosting [14] with the possibility of sampling by rows and columns of the data set; regulated gradient boosting [30] with L1 and L2 regularization.

Function for optimization of gradient busting:

$$L^{(t)} = \sum_{i=1}^{n} l(y_i, \hat{y}_i^{(t-1)} + f_t(x_i)) + \Omega(f_t), , \quad (7)$$

where l is the loss function, $y_i, \hat{y}_i^{(t-1)}$ are values of the i-th element of the training sample and the sum of the predictions of the first t trees respectively, x_i is a set of features of the i-th element of the training sample, f_t is a function (in our case a tree) that authors want to learn in step t, $f_t(x_i)$ is a prediction on the i-th element of the training sample, $\Omega(f_t)$ is a regularization of the function f.

IV. Implementation of ML Models for Predicting Energy Production

Formation of a dataset for the study of the energy consumption forecast. The dataset contains observations over a 5-month period: January 11, 2016 to May 27, 2016. The dataset is a combination of humidity and temperature measurements for households and weather data in Mykolaiv Region (Fig. 4).

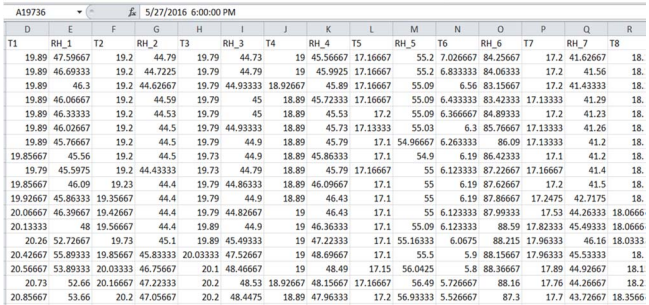


Fig. 4. Dataset for the energy consumption forecast

Calculation of energy efficiency (production) is seen as a time series problem. For this reason, authors have implemented time series methods such as ARIMA from *statsmodels* and Seasonal Decomposition using the *Prophet* library.

The Exploratory Data Analysis (EDA) method was used for data analysis and processing. Authors analyzed data from *rp5.ua* in combination with experimental datasets from *archive.ics.uci.edu*. The *sunrise-and-sunset.com* resource was used to calculate the day length in the Mykolayiv region. In this way, an experimental dataset *energydata.csv* was formed (Table 1).

The following conclusions for the used dataset: most temperatures are highly correlated; the outside temperature (building and airport) is also highly correlated; transformations of humidity indicators do not affect linearity. The distribution of external humidity (measured from the airport side) also has negative vectors. The same result is true for RH6 (humidity outside the building). Therefore, the highest probability of humidity is above average.

TABLE I. The Experimental Dataset of *energydata.csv*

Variables	Definition
date	Date and time
Appliances	Energy use, Wh
lights	Energy use of lamps in the house, Wh
T1, T2, T3, T4, T5, T6, T7, T8, T9	Celsius temperature, °C
RH_1, RH_2, RH_3, RH_4, RH_5, RH_6, RH_7, RH_8, RH_9	Humidity, %
T_out	Temperature outside the weather station (WS), °C
Press_mm_hg	Pressure (from Mykolaiv WS), mm of mercury
RH_out	Humidity outside WS, %
Windspeed	Wind speed (from Mykolaiv WS), m/s
Visibility	Visibility (from Mykolaiv WS), km
Tdewpoint	Tdewpoint (from Mykolaiv WS), °C
rv1, rv2	Random variables

Using the Prophet library authors specified the parameters of changepoint_prior_scale = 0.1, weekly_seasonality = True, daily_seasonality = True, we get energy consumption trends from January to May (Fig. 5), as well as weekly and daily consumption trends (Fig. 6).

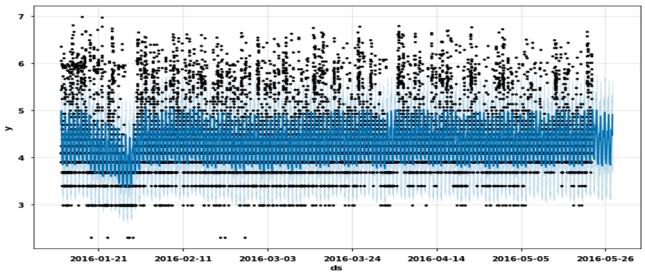


Fig. 5. Forecasting energy consumption by six months

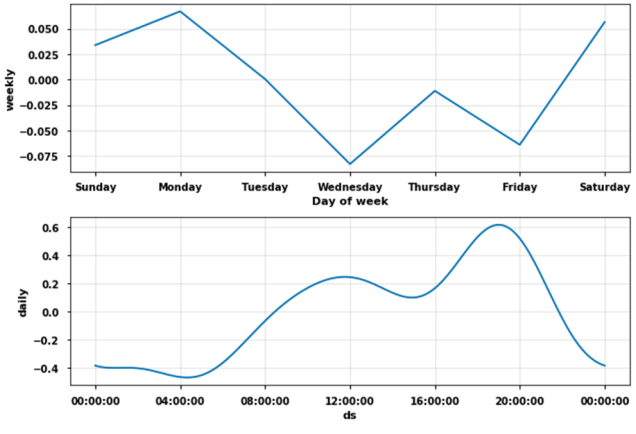


Fig. 6. Weekly and daily forecasting energy consumption

The observations of the dataset relate to a period of less than one year, so seasonality can be curbed. This seasonality can be expressed by dummy variables that explain the dependent variable (consumption).

After adjusting the hyperparameters during validation setting, authors make the first estimate using MLR model. There is used the Standard Scaler for standardization for linear regression tests. The multiple linear regressions help to obtain indicators of the forecasting model. This model is used for data validation, for training authors use models of SVR, Random Forest, and XGBoost.

For each of the models, authors perform 3 tests with changes in hyperparametics. The first test was optimized using the GridSearchCV library. GridSearchCV takes a dictionary that describes the options you can try on a model to train it. The parameter grid is defined as a dictionary, where the keys are the parameters and the values are the parameters to be checked. The GridSearchCV process will build and evaluate one model for each combination of parameters. Cross-validation is used to evaluate each individual model, and 3-fold cross-validation is used by default, although this can be changed by specifying the cv argument for the GridSearchCV constructor.

Authors implemented the models described earlier and compared the results of their work using Root Mean Square Error (RMSE) to see how much the data differs from the typical from the arithmetic mean. The results are presented in the Table 2.

TABLE II. RESULTS AND PARAMETERS COMPARISON OF ML MODELS

Model name	Parameters	Values	RMSE
MLR	-	-	0.569
SVR	regressor_gamma	[0.001, 0.01, 0.1]	0.434
	regressor_C	[50, 60, 70]	
Random Forest	n_estimators	[100, 110, 130, 140]	0.407
	max_depth	[0.001, 0.01, 0.1]	
	max_features	[50, 60, 70]	
XGBoost	n_estimators	[4, 5, 6]	0.443
	max_depth	[100, 110, 130, 140]	
	Learning_rate	[0.1, 0.5, 0.7]	

From these results we conclude that the Random Forest model gives the best result for solving energy consumption forecasting problems due to the lowest RMSE deviation. The Random Forest model was selected for further training of the dataset.

To analyze the resulting dataset of forecasted energy consumption and power generation data at South-Ukraine NPP, authors create an interface (Fig. 7) using the Predix development package using the Remote Monitoring & Diagnostics (RMD) Reference application. The Predix platform allows creating interfaces that focus on the Industrial Internet to manage and scale data analytics projects.

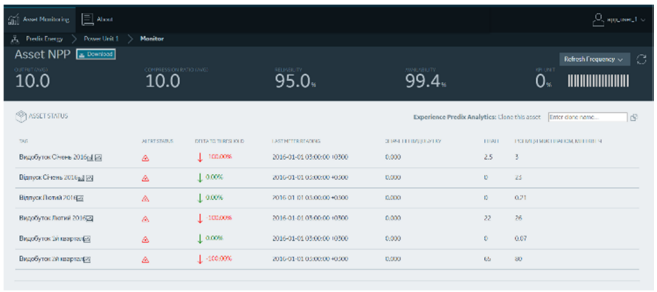


Fig. 7. Web-interface for downloading and analyzing the resulting dataset of forecasted energy consumption and power generation data at South-Ukraine NPP

CONCLUSIONS

The paper describes the general characteristics of 4 ML models used in the task of forecasting energy production: Multiple Linear Regression (MLR), Support vector regression (SVR), Random Forest and Extreme gradient boosting (XGBoost). The Random Forest model has shown the best results for the energy demand forecasting task. The Python programming language as well as the PySpark library were used to create and model it.

As part of the research, GridSearchCV is used to train algorithms to optimize model parameters. After examining the process of adjusting parameters to determine the optimal values for these models, authors concluded that the Random Forest model is based on the highest performance on the set values of the parameters.

REFERENCES

[1] N. I. Nwulu and O. P. Agboola, "Modelling and predicting electricity consumption using artificial neural networks," 11th International Conference on Environment and Electrical Engineering, Venice, 2012, pp. 1059–1063.

[2] N. Karimtabar, S. Pasban, and S. Alipour, "Analysis and predicting electricity energy consumption using data mining techniques — A case study I.R. Iran — Mazandaran province," 2nd International Conference on Pattern Recognition and Image Analysis (IPRIA), Rasht, 2015, pp. 1–6.

[3] W. Li-Yao and Y. Feng-Mao, "Using seasonal time series analysis to predict China's demand of electricity," International Conference on Computational and Information Sciences, Shiyang, 2013, pp. 76–79.

[4] C. Szepesv´ari, Algorithms for reinforcement learning. Synthesis lectures on artificial intelligence and machine learning. California, USA: Morgan & Claypool Publishers, 2009.

[5] J. Friedman, "Greedy function approximation: a gradient boosting machine," in Ann. Stat., vol. 29, no. 5, 2001, pp. 1189–1232.

[6] N. M. Ball and R. J. Brunner, "Data mining and machine learning in astronomy," in International Journal of Modern Physics D, vol. 19, no. 7, 2010, pp. 1049–1106.

[7] K. P. Burnham and D. R. Anderson, Model selection and multimodel inference: a practical information-theoretic approach. New York: Springer-Verlag, 2002.

[8] P. K. Mallapragada, R. Jin, A. K. Jain, and Y. Liu, "Semiboost: boosting for semi-supervised learning," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 31, no. 11, 2009, pp. 2000–2014.

[9] I. Sidenko, G. Kondratenko, P. Kushneryk, and Y. Kondratenko, "Peculiarities of human machine interaction for synthesis of the intelligent dialogue chatbot," 10th IEEE International Conference on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications (IDAACS), Metz, France, 2019, pp. 1056–1061.

[10] Y. P. Kondratenko, O. V. Kozlov, G. V. Kondratenko, and I. P. Atamanyuk, "Mathematical model and parametrical identification of ecopyrogenesis plant based on soft computing techniques," in Complex Systems: Solutions and Challenges in Economics, Management and Engineering. Studies in Systems, Decision and Control, vol. 125, C.

Berger-Vachon, A. Gil Lafuente, J. Kacprzyk, Y. Kondratenko, J. Merigó, and C. Morabito, Eds. Cham: Springer, 2017, pp. 201–233.

[11] F. J. Ardakani and M. M. Ardehali, "Novel effects of demand side management data on accuracy of electrical energy consumption modeling and long-term forecasting," in Energy Convers. Manag., vol. 78, 2014, pp. 745–752.

[12] E. Yu. Shchetinin, "Cluster-based energy consumption forecasting in smart grids," Springer Communications in Computer and Information Science (CCIS), 2018, pp. 446–456.

[13] M. J. Kane, N. Price, M. Scotch M., and P. Rabinowitz, "Comparison of ARIMA and Random Forest time series models for prediction of avian influenza H5N1 outbreaks," in BMC Bioinformatics, vol. 15, 2014, pp. 276–302.

[14] R. K. Jain, K. M. Smith, P. J. Culligan, and J. E. Taylor, "Forecasting energy consumption of multi-family residential buildings using support vector regression: investigating the impact of temporal and spatial monitoring granularity on performance accuracy," in Appl. Energy, vol. 123, 2014, pp. 168–178.

[15] Y. Xie and M. Li, "Application of gray forecasting model optimized by genetic algorithm in electricity demand forecasting," Second International Conference on Computer Modeling and Simulation, Sanya, Hainan, 2010, pp. 275–277.

[16] I. E. Kafazi, R. Bannari, and A. Abouabdellah, "Modeling and forecasting energy demand," International Renewable and Sustainable Energy Conference (IRSEC), Marrakech, 2016, pp. 746–750.

[17] I. Atamanyuk, V. Kondratenko, Y. Kondratenko, V. Shebanin, M. Solesvik, "Models and algorithms for prediction of electrical energy consumption based on canonical expansions of random sequences," in Green IT Engineering: Social, Business and Industrial Applications. Studies in Systems, Decision and Control, vol. 171, V. Kharchenko, Y. Kondratenko, and J. Kacprzyk, Eds. Cham: Springer, 2019, pp. 397–421.

[18] I. Atamanyuk, V. Shebanin, Y. Volosyuk, and Y. Kondratenko, "Generalized method for prediction of the electronic devices and information systems' state," 14th International Conference on Perspective Technologies and Methods in MEMS Design, MEMSTECH, Lviv, Ukraine, 2018, pp. 91–95.

[19] F. Recknagel, "Application of macine learning to ecological modelling," in Ecological Modelling, vol. 146, no. 1, 2001, pp. 303–310.

[20] E. Yu. Shchetinin, V. S. Melezhik, and L. A. Sevastyanov, "Improving the energy efficiency of the smart buildings with the boosting algorithms," Proceedings of the Selected Papers of the 12th International Workshop on Applied Problems in Theory of Probabilities and Mathematical Statistics in the framework of the Conference on Information and Telecommunication Technologies and Mathematical Modeling of High-Tech Systems (APTP+MS'2018), CEUR Workshop Proceedings, vol. 2332, 2018, pp. 69–78.

[21] L. Yang, Y. Li and C. Di, "Application of XGBoost in Identification of Power Quality Disturbance Source of Steady-state Disturbance Events," IEEE 9th International Conference on Electronics Information and Emergency Communication (ICEIEC), Beijing, China, 2019, pp. 1–6.

[22] X. Zhu, Semi-supervised learning literature survey. Comp. Sci. Tech. Rep. University of Wisconsin-Madison, 2008.

[23] J. Taylor, "Short-term electricity demand forecasting using double seasonal exponential smoothing," in Journal of Operational Research Society, vol. 54, 2003, pp. 799–805.

[24] C. Clancy, J. Hecker, E. Stuntebeck, and T. O. Shea, "Applications of machine learning to cognitive radio networks," in IEEE Wireless Communications, vol. 14, no. 4, 2007, pp. 47–52.

[25] A. Liaw and M. Wiener, "Classification and regression by random forest," in R News, vol. 2, no. 3, 2002, pp. 18–22.

[26] A. K. Jain, M. N. Murty, and P. J. Flynn, "Data clustering: a review," in ACM Computing Surveys, vol. 31, no. 3, 1999, pp. 264–323.

[27] P. Kushneryk, Y. Kondratenko, and I. Sidenko, "Intelligent dialogue system based on deep learning technology," 15th International Conference on ICT in Education, Research, and Industrial Applications: PhD Symposium (ICTERI 2019: PhD Symposium), vol. 2403, Kherson, Ukraine, 2019, pp. 53–62.

[28] R. J. Hyndman and Y. Khandakar, "Automatic time series forecasting: the forecast package for R," in Journal of Statistical Software, vol. 26, no. 3, 2008, pp. 1–22.

[29] Y. Freund and R. E. Schapire, "A short introduction to boosting," International Joint Conference on Artificial Intelligence, 1999, pp. 1401–1406.

[30] Z. H. Zhou, "Ensemble learning," in Encycl. Biom., 2015, pp. 411–416.

IEEE Third International Conference on Data Stream Mining & Processing
August 21-25, 2020, Lviv, Ukraine

Hybrid Artificial Intelligence Systems for Complex Neural Network Analysis of Abnormal Neurological Movements with Multiple Cognitive Signal Nodes

Ivan Mudryk
Department of Software Engineering
Ternopil Ivan Pulu'j National Technical University
Ternopil, Ukraine
i1mudryk@ukr.net

Mykhaylo Petryk
Department of Software Engineering
Ternopil Ivan Pulu'j National Technical University
Ternopil, Ukraine
mykhaylo_petryk@tu.edu.te.ua

***Abstract* — Research related to feedback-neuro-biosystems regarding the analysis of the status and behavior of T-objects under the cognitive influence of neural nodes. They focus on the study of parameters relative to normal states and behavior (normal wave movements of certain parts of the body), which used classical Fourier transform digital processing techniques and the cognitive connections were estimated approximately using neural network techniques and software technologies. In order to decompose the complex abnormal movement path, a scheme of multicomponent decomposition of segments of the trajectory of trajectories is used to further formulate the mathematical model.**

Keywords — software system, cognitive signal, tremor, modeling of objects and processes, multi-parameter identification, neural network analysis

I. INTRODUCTION

Use of the latest high-tech artificial intelligence, new high-performance computing solutions taking into account modern computer systems software architectures (parallel multi-step calculations, new multi-parameter identification algorithms) allows to provide a systematic approach to the design of high-tech methods of digital diagnostics and health monitoring, which requires gradual changes in the scientific-engineering methods. In the context of a number of European Union programs, including Horizon 2020 and Horizon Europe, critical millennium illnesses have been characterized by enormous and abnormal neurological tremors associated with human limb friction, rapid progression of which causes Alzheimer's disease, Parkinson's disease, cause paralysis of the nervous system. Early diagnosis and treatment will save hundreds of thousands of lives. According to WHO, more than 100 million people in the world today are suffering from these diseases. [1]

II. COMPREHENSIVE METHODOLOGY AND ANALYSIS TOOLS FOR THE DIAGNOSIS OF NEUROLOGICAL CONDITIONS

The authors propose a method of analysis for the diagnosis of neurological conditions of T-objects focused primarily on determining the parameters of abnormal movements of patients with tremor-signs caused by the negative effects of a certain set of neural nodes of the cerebral cortex.

Model elements are integrated into the system by using appropriate interface conditions. Based on hybrid integral transformations (Fourier, Bessel, Hilbert), obtained a rapid analytic solution of the model as a vector function that describes the 3D elements of the trajectories at each ANM segment. The basis of the proposed hybrid transformation provides an integral vector solution of the model. The main element of the solution is the adaptive matrix (response) that determines the state parameters of the action of certain groups of brain neuro- cortex.

The technique is based on a hybrid model of the neuro-system (cortex nodes and tremor-object) based on the waveform propagation and the behavior of T-objects, namely, segmental description of 3D elements of trajectories of abnormal neurological movements of the studied T-object (limb of the hand) taking into account the matrix of cognitive influences of neuro-nodal groups of the cerebral cortex. Including the construction of hybrid spectral function system with all segments of ANM, [2, 5].

In order to decompose complex ANM motions into simpler elements, the number of chosen partitions arbitrarily, depending on the complexity of the ANM images. In this technique of analyzing the data of the spiral test, it is extremely important and up-to-date to obtain the frequency response using hybrid Fourier transform and methods of digital signal processing on hybrid spectral functions and spectral values [6, 10-11]. The mathematical model assumes quantitative characteristics of tremor.

A known method of identifying tremor on the plane using a picture recognition Archimedean spiral that can be performed on the graphical pen tablet.

The basis for the implementation of the hardware solution is the method of continuous determination of the position of the electronic pen with respect to any control coordinate [4]. To perform empirical studies, a touch pen (Wacom Cintiq 12WX digitizer on Fig.1) with a sampling rate of 133 Hz and an accuracy of ± 0.25 mm was used. The pattern looks like an Archimedes spiral with several turns for or counter-clockwise, with an inter-loop of 9 mm. This template is located on the screen of an interactive tablet, with the possibility of drawing a patient by an electronic pen. The e-pen is used to identify handwritten input (numbers, text information, template drawings) or to record and digitize arbitrary movements of the limb.

978-1-7281-3214-3/20/$31.00 ©2020 IEEE

Analysis of forms of tremor is of great importance in neurological practice, in particular, in the diagnosis of Parkinson's syndrome, when increased visible naked eye tremor is an indicator of the pathological condition of the patient.

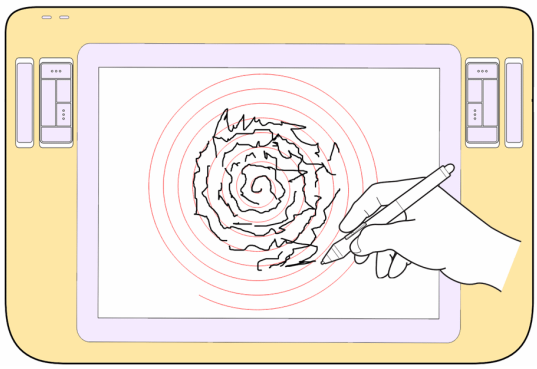


Fig. 1. Example how to use the software with the Archimedes spiral pattern on a Wacom graphic tablet.

We have proposed a graphical digital pen device with a built-in 3D micro-accelerometer for diagnostic testing. The microcontroller reads and processes information from a three-axis acceleration sensor (micro-accelerometer). According to the proposed formulas, the instantaneous coordinates of the accelerometer position in space are determined [3-6]. In parallel flow information is received about moving the electronic pen on the plane of the graphic tablet.

Data of pen movement in the form of a 3D model of the ANM T-object is formed in a graphical window with the possibility of decomposing complex 3D motions into 3 projections and further analysis of each of them and choosing the most decisive one for the identification and complex estimation of the parameters of the ANM.

III. Hybrid Mathematical Model of ANM T-Object Analysis Based on Feedbacks

A. Basic physical assumptions

Formulation and method of solving the direct inhomogeneous boundary value problem of ANM analysis based on cognitive feedbacks. According to the program of experimental research of ANM T-objects, the data of one of the determining projections of motion in the form of a spiral are used to obtain qualitative formulation of the problem and to construct the mathematical model of ANM, which is easily transformed into a Cartesian graph (relative to the z-axis, Fig. 2).

This trajectory is related to the cognitive feedbacks of a certain set of cortex neural nodes, which send signals to control this oscillatory neurological movement and determine the overall dynamics of the ANM of the T-object under study.

In order to decompose the complex ANM path, a scheme of its multicomponent decomposition of segments of the trajectory of trajectories is used to further formulate the mathematical model.

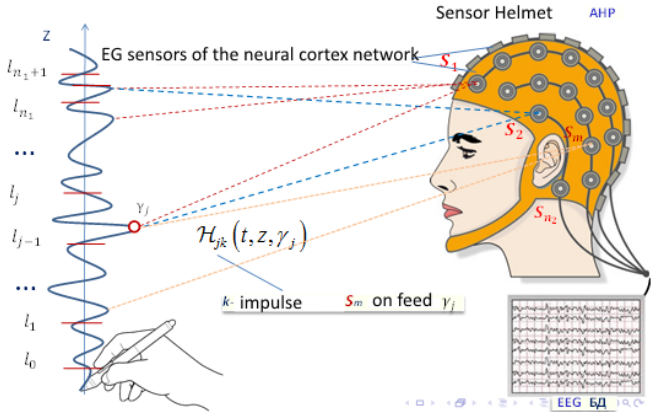


Fig. 2. T-Object Neuro-Nodes Interaction System. Component decomposition of complex ANR path into arbitrary of simple elements of motion.

Accordingly, the trend distributions of the EEG signals of the neuronodes controlling the oscillatory neurological movement and, in general, determine the dynamics of the ANM for each j-th segment of the route are correlated. $j=\overline{1,n_1+1}$, where n_1 is the number of breaking points of the ANR route. The partitioning can be set automatically arbitrarily, with any subtraction the number of segments whose lengths can also be different depending on the level of detail of the sections of motion and the choice of acceptable basis functions and their construction on the basis of acceptable dependences of their approximation. [5-8] One of the criteria for determining the lengths of the elements of the partition may be the amplitude characteristics of the individual trends of the oscillatory ANR motions, etc. [7].

B. Mathematical formulation of the problem

Based on the stated physical assumptions of this subject area of neurological analysis, the direct inhomogeneous initial-boundary-value problem of determining the ANM parameters of a T-object can be described as a system of equations:

$$\frac{\partial^2 u_j(t,z)}{\partial t^2}=b_j^2\frac{\partial^2 u_j}{\partial z^2}+S_j^*(t,z),\quad z\in(l_{j-1},l_j),\ j=\overline{1,n_1+1} \tag{1}$$

with homogeneous initial conditions:

$$u_j(t,z)\big|_{t=0}=0,\qquad \frac{\partial u_j}{\partial t}\bigg|_{t=0}=0\qquad ,j=\overline{1,n_1+1} \tag{2}$$

with homogeneous boundary conditions and system interface conditions:

$$\frac{\partial}{\partial z}u_1(t,z)_{z=0}=0,\quad \frac{\partial}{\partial z}u_n(t,z)_{z=l}=0, \tag{3}$$

$$\left[u_j(t,z)-u_j(t,z)\right]\Big|_{z=l_j}=0,$$

$$\left(b_j^2\frac{\partial}{\partial z}u_k(t,z)-b_{j+1}^2\frac{\partial}{\partial z}u_{j+1}(t,z)\right)\Big|_{z=l_j}=0,\ j=\overline{1,n_1} \tag{4}$$

in a multi-component area:

$$D_{n_1}^{+}=\left\{(t,z):t\in(0;\mathrm{T}),z\in I_{n_1}=\bigcup_{j=1}^{n_1+1}\left(l_{j-1},l_j\right);\ l_0=0,\ l_{n_1+1}\equiv l<\infty\right\}$$

Here (1) is a system of wave equations describing the ANR-trajectory of tremor at each j-th segment of the trajectory $j=\overline{1,n_1+1}$. Depending on the resultant action of the set of signals $S_j^*(t,z)$, originating from EEG sensors for a given set of cortex neuronodes, that control the behavior of the investigated T-object, $b_j,\ j=\overline{1,n_1+1}$ - components of the phase propagation speed of the ANM waves, which are the amplitude characteristics of the wave tremor motion:

$$S_j^*(\tau,\xi)=\sum_{i=1}^{n_2}\alpha_{ji}S_i(\tau,\xi),\ \left[\alpha_{ji}\right],\ j=\overline{1,n_1},\ i=\overline{1,n_2}$$

- is an adaptive matrix that determines the relationships and feedback effects of specific cortex neurosignals on selected small segments of the ANM path. The matrix element α_{ji} is a weighting factor (0 to 1) that determines the integral influence of the i-th neuron node S_i on the *j-th* segment of movement (determined by data mining machine learning methods [9]). The interface conditions (3), (4) ensure the continuity and integrity of the solution of the problem for the entire multicomponent domain of its definition.

C. *Construction of an analytical solution to the ANM boundary value problem*

To construct the analytic solution of a direct inhomogeneous problem (1)-(4), we apply a hybrid Fourier transform, defined in [13-14]. The basis of the conversion is the hybrid integral operators written in matrix form:

- direct action:

$$F_{n_1}[...]=\begin{bmatrix}\int_{l_0}^{l_1}...V_1(z,\beta_m)\sigma_1 dz\int_{l_1}^{l_2}...V_2(z,\beta_m)\sigma_2 dz...\\ ...\int_{l_{n_1-1}}^{l_{n_1}}...V_{n_1}(z,\beta_m)\sigma_{n_1}dz\int_{l_{n_1}}^{l_{n_1+1}}...V_{n_1+1}(z,\beta_m)\sigma_{n_1+1}dz\end{bmatrix} \tag{5}$$

- reverse action:

$$F_{n_1}^{-1}[...]=\begin{bmatrix}\sum_{m=1}^{\infty}...V_1(z,\beta_m)\left(\|V(z,\beta_m)\|^2\right)^{-1}\\ \sum_{m=1}^{\infty}...V_2(z,\beta_m)\left(\|V(z,\beta_m)\|^2\right)^{-1}\\ .....\\ \sum_{m=1}^{\infty}...V_{n_1+1}(z,\beta_m)\left(\|V(z,\beta_m)\|^2\right)^{-1}\end{bmatrix} \tag{6}$$

Here, $\left[V_k(z,\beta_m)\right]_{,k=\overline{1,n_1+1}}$ the vector of the hybrid spectral function defined:

$$\begin{bmatrix}V_1(z,\beta_m)\\ ...\\ V_k(z,\beta_m)\\ ...\\ V_{n_1+1}(z,\beta_m)\end{bmatrix}=\begin{bmatrix}\prod_{i=1}^{n_1}\xi_{i+1}\frac{\beta_m}{b_{i+1}}\left(\omega_0^2(\beta_m)\vartheta_1^{11}\left(\frac{\beta_m}{b_1}z\right)-\omega_0^1(\beta_m)\vartheta_1^{21}\left(\frac{\beta_m}{b_1}z\right)\right)\\ ...\\ \prod_{i=k}^{n_1}\xi_{i+1}\frac{\beta_m}{b_{i+1}}\left(\omega_{k-1}^2(\beta_m)\vartheta_k^{11}\left(\frac{\beta_m}{b_k}z\right)-\omega_{k-1}^1(\beta_m)\vartheta_k^{21}\left(\frac{\beta_m}{b_k}z\right),\ _{k=\overline{2,n_1}}\right)\\ ...\\ \omega_{n_1}^2(\beta_m)\vartheta_{n_1+1}^{11}\left(\frac{\beta_m}{b_{n_1+1}}z\right)-\omega_{n_1}^1(\beta_m)\vartheta_{n_1+1}^{21}\left(\frac{\beta_m}{b_{n_1+1}}z\right)\end{bmatrix} \tag{7}$$

$\{\beta_m\}_{m=0}^{\infty}$ - set of spectral values of hybrid integral Fourier transform, which are the roots of the transcendental equation:

$$\omega_{n_1}^2(\beta)\vartheta_{n+1}^{11}\left(\frac{\beta}{b_{n_1+1}}l_{n_1+1}\right)-\omega_{n_1}^1(\beta)\vartheta_{n_1+1}^{21}\left(\frac{\beta}{b_{n_1+1}}l_{n_1+1}\right)=0 \tag{8}$$

As a result of the application of hybrid integral operators (1)-(4) to the system of equations (1)-(4), the analytic solution of the transformation model is obtained and the uniform solution of the inhomogeneous boundary value problem ANM (1)-(4) obtained:

$$u_j(t,z)=\sum_{k=1}^{n_1+1}\int_0^t\int_{l_{k-1}}^{l_k}\mathcal{H}_{jk}(t-\tau,z,\xi)S_k^*(\tau,\xi)\sigma_k d\xi d\tau,\quad j=\overline{1,n_1+1} \tag{9}$$

Here, the influence matrix is the response of the ANM system to the influence of the *k-th* segment of the resultant action of the signals of a certain set of cortex neuro-nodes on the j-segment of the ANM route.

$$\mathcal{H}_{jk}(t,z,\xi)=\sum_{m=1}^{\infty}\frac{\sin\beta_m t}{\beta_m}\frac{V_j(z,\beta_m)V_k(\xi,\beta_m)}{\|V(z,\beta_m)\|^2};\ j,k=1,\overline{n+1} \tag{10}$$

IV. Spatial Results Visualization of a Digital Analysis of the T-object's ANM Trajectory

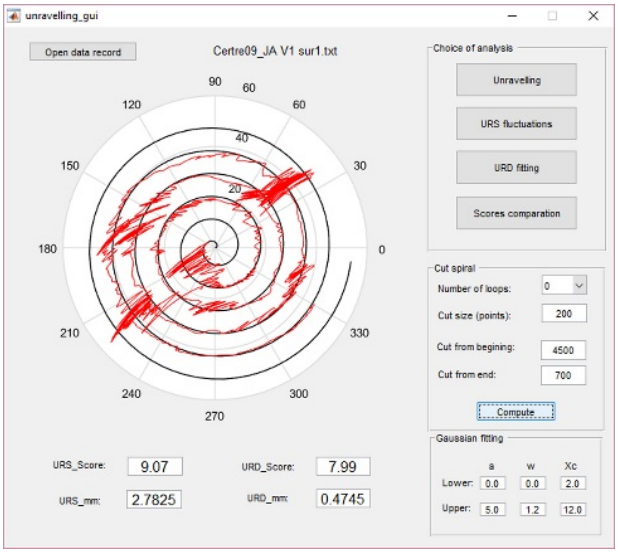


Fig 3. Results of digital analysis of highly oscillated ANM movements performed by the limb of an electronic pen on an electronic tablet by the Archimedes spiral circumference of a patient's arm with severe signs of tremor.

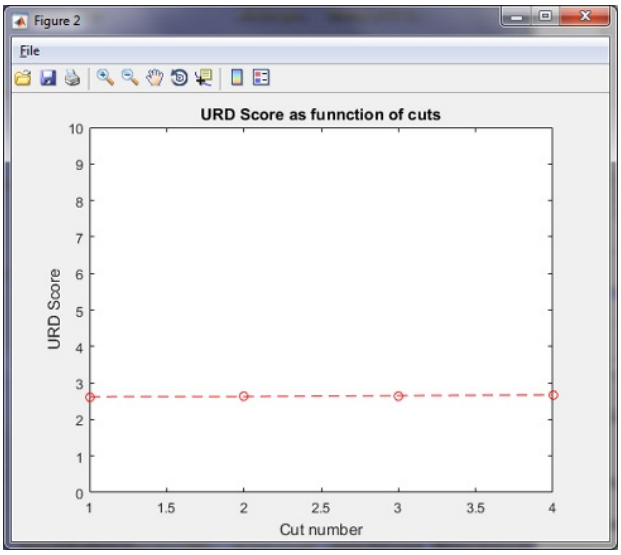


Fig 4. Results of digital analysis score of highly oscillated ANM movements performed

Figure 3 presents the results of digital analysis of abnormal neurologically highly oscillatory movements performed by the extremity of an electronic pen on an electronic tablet by the circumference of a test pattern (Archimedes spiral) by the patient's hand with strongly expressed signs of tremor (T-object).

A mathematical method for analyzing input information and converting data are described in the sources score Figure 4 [10-14].

Such sections of ANM movements can be studied more detailed, breaking them into separate segments over the studied time interval, establishing the dependence of their real amplitude and frequency characteristics on the integral temporal distributions of the cognitive signals of cerebral cortex nodes.

CONCLUSIONS

The analysis of forms of tremor is of great importance in neurological practice, in particular, in the diagnosis of Parkinson's syndrome, when the raised visible to the naked eye tremor is an indicator of the pathological condition of the patient.

Hybrid model of a neuro feedback system developed, describes the state and behavior of tremor (T) objects - segment description of 3D elements of trajectories of normal neurological movements (ANM) through the cognitive impact matrices of groups of neuro-nodes of the cerebral cortex.

An effective way of analyzing the results obtained is the ability to perform cyclic calculations based on the proportional reduction of the analyzed data sets. Obtained estimates are compared for each iteration of the limitation of the analyzed data. Presented results as frequency and amplitude characteristics curve, form the basis of the evaluation of the patient by computerized diagnostics.

REFERENCES

[1] J.L. Yang, R.S.Chang, F.P. Chen, C.M. Chern, J.H. Chiu. Detection of hand tremor in patients with Parkinson's disease using a non-invasive laser line triangulation measurement method / Measurement 79 (2016) 20–28

[2] A.- P. Legrand, I. Rivals, A. Richard, E. Apartis, E. Roze, M. Vidailhet, S. Meunier, E. Hainque. New insight in spiral drawing analysis methods – Application to action tremor quantification. J Clinical Neurophysiology (2017). Volume 128, Issue 10, Pages 1823–1834

[3] Mykhalyk D ., Mudryk I., Hoi A., Petryk M. Modern hardware and software solution for identification of abnormal neurological movements of patients with essential tremor. IEEE. Proceeding of 2019 9th International Conference on Advanced Computer Information Technologies (ACIT, Budejovice, Czech Republic), 183-186 (2019)

[4] Press, W.H., Teukolsky, S.A., Vetterling, W.T., Flannery, B.P., Numerical Recipes in C: The Art of Scientific Computing, 2nd Edition. Cambridge University Press, 1992

[5] I.V. Sergienko, V.S. Deineka, Optimal Control of Distributed Systems with Conjugation Conditions, New York: Kluwer Academic Publishers 2005.

[6] Haubenberger D, Kalowitz D, Nahab F B, Toro C, Ippolito D, Luckenbaugh DA, Wittevrongel L, Hallett M. Validation of Digital Spiral Analysis as Outcome Parameter for Clinical Trials in Essential Tremor. Movement Disorders 2011;26(II):2073-2080

[7] Pullman, S L. Spiral analysis: a new technique for measuring tremor with digitizing tablet. Movement Disorders 1998;3:85-89.

[8] Erasmus L-P, Sarno S, Albrecht H, Schwecht M, Pöllmann W, König N. Measurement of ataxic symptoms with a graphic tablet: standard values in controls and validity in multiple sclerosis patients. J Neurosci Methods 2001;108:25–37

[9] Petryk M., Khimitch A., Petryk M.M., Fraissard J. Experimental and computer simulation studies of dehydration on microporous adsorbent of natural gas used as motor fuel. Fuel. Vol. 239, 1324–1330 (2019)

[10] Leclerc S., Petryk M., Canet D., Fraissard J. Competitive diffusion of gases in a zeolite using proton NMR and a slice selection procedure. Catalysis Today, Elsevier B.V. (2012) Volume 187, Issue 1, 104-107

[11] Viviani P, Burkhard PR, Chiuvé SC, dell'Acqua CC, Vindras P. Velocity control in Parkinson's disease: a quantitative analysis of isochrony in scribbling movements. Exp Brain 2009;194:259–83.

[12] Petryk M., Mykhalyk D., Mudryk I. A method of digital measurement of parameters of abnormal neurological movements of the upper limbs in patients with manifestations of a tremor // Patent of Ukraine № 130247. 26.12.2018.

[13] Petryk M., Vorobiev E. Numerical and Analytical Modeling of Solid-Liquid Expression from Soft Plant Materials. AIChE Journal. Wiley USA. (2013) Volume 59, Issue 12, 4762–4771.

[14] Xie H, Wang Z. Mean frequency derived via Huang-Hilbert transform with application to fatigue EMG signal analysis. Comput Meth Progr Biomed 2006;82:114–20

[15] A. Salarian, H. Russmann, C. Wider, P. R. Burkhard, F. J. Vingerhoets, and K. Aminian, "Quantification of tremor and bradykinesia in Parkinson's disease using a novel ambulatory monitoring system," Biomedical Engineering, IEEE Transactions on, 54. Jg., Nr. 2, pp. 313-322, 2007.

[16] G. LO, A. R. Suresh, L. Stocco, S. González-Valenzuela, and V. C. Leung, "A wireless sensor system for motion analysis of Parkinson's disease patients," in: Pervasive Computing and Communications Workshops (PERCOM Workshops), 2011 IEEE International Conference on. IEEE, pp. 372-375, 2011

IEEE Third International Conference on Data Stream Mining & Processing
August 21-25, 2020, Lviv, Ukraine

A Hybrid EMD - Neuro-Fuzzy Model for Financial Time Series Analysis

Alexander Vlasenko
Department of Artificial Intelligence
Kharkiv National University of Radio Electronics
Kharkiv, Ukraine
alexander.vlasenko86@gmail.com

Olena Vynokurova
GeoGuard
Kharkiv, Ukraine
Kharkiv National University of Radio Electronics
Kharkiv, Ukraine
vynokurova@gmail.com

Dmytro Peleshko
GeoGuard
Kharkiv/Lviv, Ukraine
dpeleshko@gmail.com

Nataliia Vlasenko
Department of Informatics and Computer Engineering
Simon Kuznets Kharkiv National University of Economics
Kharkiv, Ukraine
gorohovatskaja@gmail.com

Yuriy Rashkevych
National University "Lvivska politechnika"
Lviv, Ukraine
rashkevyuriy@gmail.co>

***Abstract*— Time series in finance are distinguished by their highly complex nonlinear dynamics, which makes them hard to analyze and predict. This paper proposes a novel model based on combination of the empirical mode decomposition and multidimensional Gaussian neuro-fuzzy model in order to achieve better accuracy. The experimental results show significant improvement in comparison with the original model and its competitors.**

Keywords*— time series, neuro-fuzzy, empirical mode decomposition, prediction

I. INTRODUCTION

Financial time series arise in stock, derivatives, currency and other markets and their technical analysis is important but highly challenging due to their noisy, nonlinear, non-stationary and high-frequency nature.

Among various artificial intelligence techniques applied to the financial analysis neuro-fuzzy models have acquired special attention due to their inherent approximation capabilities. Work [1] reviews many successful applications of the neuro-fuzzy systems in business.

The neuro-fuzzy model introduced in [2] achieved notable performance through utilizing the representational power of the multidimensional Gaussian function and stochastic gradient descent optimization procedure. But its accuracy is still behind what a modification of that model displayed on the synthetic MG times series with added noise [3]. That fact implies applying signal decomposition and denoising procedure may bring significant benefits.

Hilbert-Huang transformation proposed in [4] has shown impressive results in various domains as a technique to process nonlinear and nonstationary time series. It comprises empirical mode decomposition (EMD), also known as Huang transform, which is used for signal decomposition into a set of simple intrinsic functions and Hilbert spectral transform. Validity and robustness of EMD was proved in [5]. Work [6] proves that it can be highly competitive as a denoising mechanism in financial data.

EMD has been effectively utilized as a decomposition and denoising tool and combined with various artificial intelligence models. In combination with deep long short-term memory (LSTM) networks EMD was used for Taiwan's CSR index forecasting [7]. Ensemble EMD (EEMD) and other decomposition techniques were used with artificial neural networks and support vector machines for stock market prediction [8]. In [9] complete ensemble empirical mode decomposition with adaptive noise (CEEMDAN) was combined with LSTM and displayed superior performance in comparison with regular LSTM networks and other models. Other examples of the mutually beneficial hybrid models are EEMD - independent component analysis (ICA) in the task of factor analysis [10], EMD - principal component analysis and BP neural network in analysis of Asia Pacific stock markets [11] and many other.

II. MODEL ARCHITECTURE AND TRAINING

The proposed model comprises three high level steps and depicted on figure 1.

978-1-7281-3214-3/20/$31.00 ©2020 IEEE

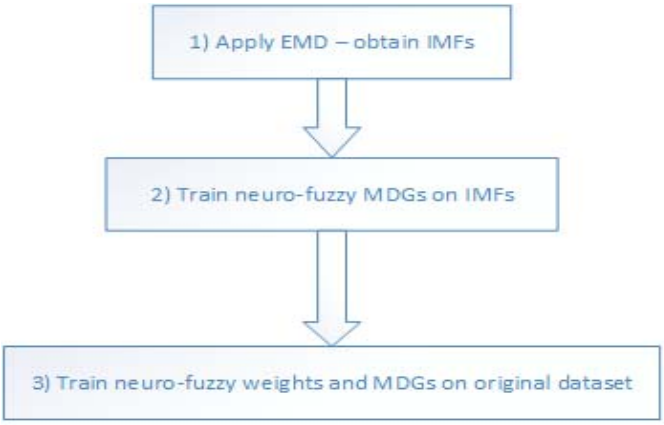


Fig. 1. The high-level structure of the proposed model.

On the first step, we apply EMD transform and obtain the intrinsic mode functions (IMF) which represent simple harmonic motion components of the original data. The IMF extraction procedure is called sifting and comprises the following actions:

- Identify all local extrema;
- Produce upper envelope $e_u(t)$ from all local maxima by a cubic spline;
- Repeat the process for the local minima in order to obtain the lower envelope $e_l(t)$.

Standard deviation criterion is used to determine the number of sifting steps [4]:

$$SD_k = \sum_{t=1}^{T} \frac{|\,\mathrm{h}_{k-1}(\mathrm{t}) - \mathrm{h}_k(\mathrm{t})\,|^2}{\mathrm{h}_k^2(\mathrm{t})} \qquad (1)$$

An example of such decomposition is shown on figure 2.

On the second step we use the obtained IMFs to train the forth layer receptive field functions one by one as in an example displayed on figure 3.

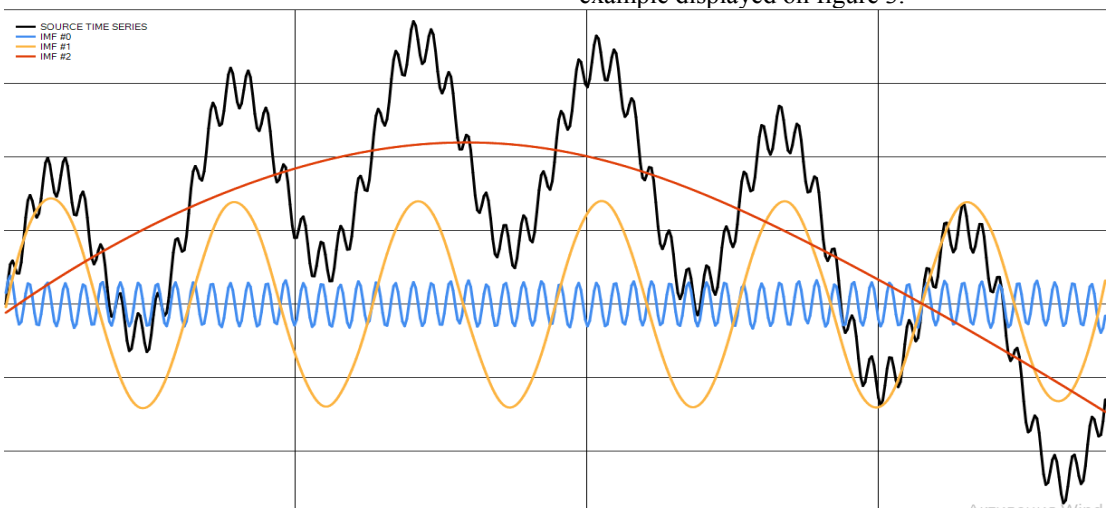


Fig. 2. The example of time series decomposition into IMFs.

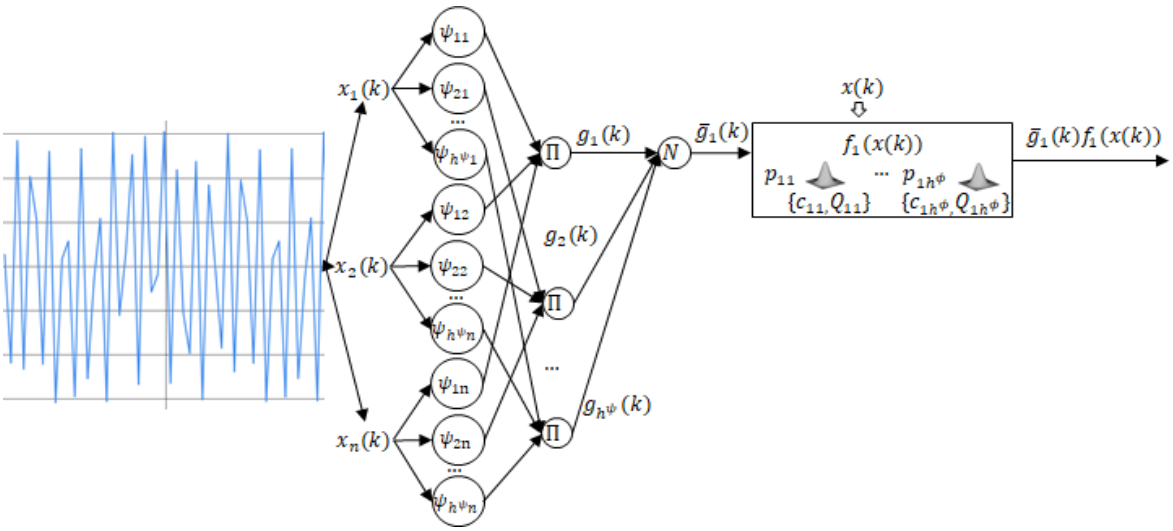


Fig. 3. The illustration of the forth layer learning.

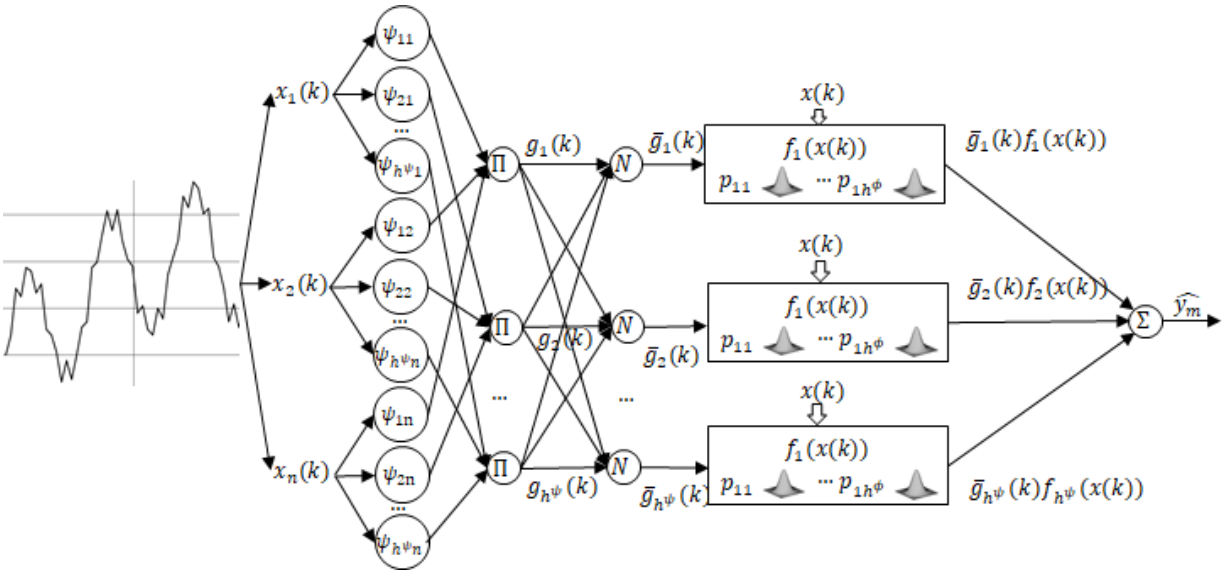


Fig. 4. The general neuro-fuzzy model architecture on the third step.

The learning of the centers c_{jl}^{φ} and the matrices Q_{jl}^{-1} performed in the following way as in [2]:

$$\begin{cases} c_{jl}^{\phi}(k+1) = c_{jl}^{\phi}(k) + \lambda_c \dfrac{\tau_{jl}^{c}(k)e(k)}{\eta_c(k)} \\ \eta_c(k+1) = \beta_c \eta_c(k) + \tau_{jl}^{c\,T}\tau_{jl}^{c} \\ Q_{jl}(k+1) = Q_{jl}(k) + \lambda_Q \dfrac{\tau_{jl}^{Q}(k)e(k)}{\eta_Q(k)} \\ \eta_Q(k+1) = \beta_Q \eta_Q(k) + Tr\left(\tau_{jl}^{QT}\tau_{jl}^{Q}\right) \end{cases} \quad (2)$$

where λ_c and λ_Q repesent learning steps β_c and β_Q is a momentum hyperparameters, τ_{jl}^{Q} is a matrix and τ_{jl}^{c} is a vector of backpropagated error values for each multidimensional Gaussian. The difference with learning in [2] that we run the learning procedure separately for each $f_j(x(k))$ and use the corresponding IMF as a training set. This part of the learning algorithm can be implemented in parallel.

Then we train the whole model on the real original training set. Weights learning is achieved through the Kaczmarz iterative method [2]:

$$p(k+1) = p(k) + \frac{y(k) - p^T f(x(k))}{f^T(x(k)) f(x(k))} f(x(k)) \quad (3)$$

where $p(k)$ is a weights vector, $y(k)$ -reference signal value, $p^T f(x(k))$ - output value.

III. Experimental results

The proposed hybrid model displayed performance improvement in comparison to original neuro-fuzzy model in stock market prediction task on the daily log returns of Alcoa and American Express datasets with 2528 records. Learning process plot is presented on figure 5.

The neuro-fuzzy model is implemented on Microsoft .Net Framework. Math.NET Numerics package [12] was used for the linear algebra operations and Accord.NET package [13] - for the reference neural networks implementation.

The computational experiments were carried out on a PC with 16 GB of RAM and two-core Intel Core i5 processor.

In order to estimate prediction accuracy Symmetric Mean Absolute Percent Error (SMAPE) and Root Mean Square Error (RMSE) criteria were used:

$$SMAPE = \frac{2}{N} \sum_{k=1}^{N} \frac{|\hat{y}(k) - y(k)|}{\hat{y}(k) + y(k)}$$
$$RMSE = \sqrt{\frac{\sum_{k=1}^{N} (y(k) - \hat{y}(k))^2}{N}} \quad (4)$$

where $y(k)$ represents reference signal value, $\hat{y}(k)$ is a prediction, and N is a length of the training set.

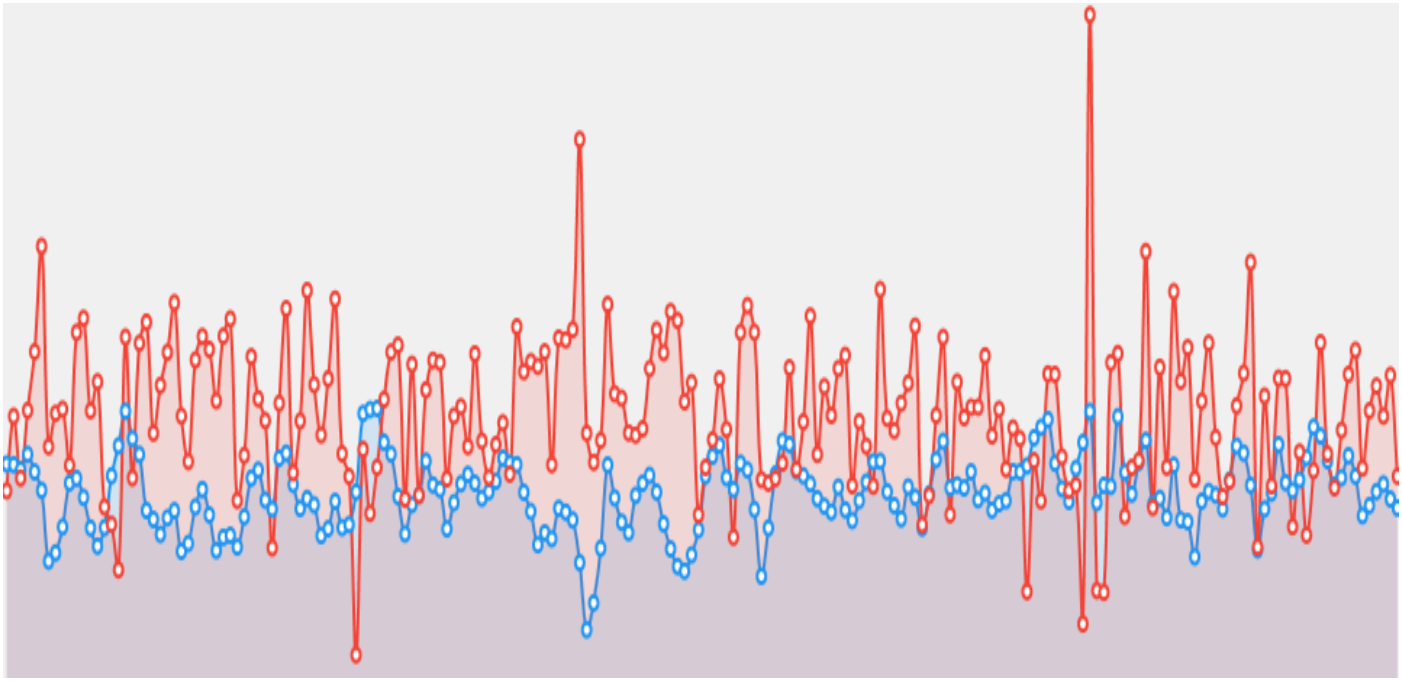


Fig. 5. Experiments on Alcoa and American Express daily returns plot

The optimal results for the proposed model were achieved with $h^{\phi} = 1$, $\lambda_c = 0.85$, $\beta_c = 0.9$, $\lambda_Q = 0.85$, $\beta_Q = 0.95$. Training of the competing bipolar sigmoid model was performed with alpha value 0.37, learning rate 0.6 and 100 epochs.

Results presented in table below show that introduced model has better prediction accuracy than the original neuro-fuzzy model but is more resource consuming.

Model	**Daily log returns dataset results**		
	Execution time, ms	***RMSE, %***	***SMAPE, %***
Proposed model	889	2.98	3.13
Simple MDG based neuro-fuzy model	450	3.17	3.45
Bipolar Sigmoid Network	2512	3.31	3.65

IV. Conclusion

This paper introduces a novel hybrid model for financial time series analysis which combines the decompositional and denoising power of empirical mode decomposition and learning and representational capabilities of the multidimensional Gaussian neuro-fuzzy model. Such solution handles complex relations in financial data, which is proven on the real life time-series forecasting tasks.

References

[1] Rajab S, Sharma V. A review on the applications of neuro-fuzzy systems in business. Artificial Intelligence Review. 2018 Apr 1;49(4):481-510.

[2] A. Vlasenko, O. Vynokurova, N. Vlasenko, and M. Peleshko, "A Hybrid Neuro-Fuzzy Model for Stock Market Time-Series Prediction," 2018 IEEE Second International Conference on Data Stream Mining & Processing (DSMP), 2018.

[3] A. Vlasenko, N. Vlasenko, O. Vynokurova, and Y. Bodyanskiy, "An Enhancement of a Learning Procedure in Neuro-Fuzzy Model," 2018 IEEE First International Conference on System Analysis & Intelligent Computing (SAIC), 2018.

[4] N. E. Huang, Z. Shen, S. R. Long, M. C. Wu, H. H. Shih, Q. Zheng, N.-C. Yen, C. C. Tung, and H. H. Liu, "The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis," Proceedings of the Royal Society of London. Series A: Mathematical, Physical and Engineering Sciences, vol. 454, no. 1971, pp. 903–995, Aug. 1998.

[5] H. Ge, G. Chen, H. Yu, H. Chen, and F. An, "Theoretical Analysis of Empirical Mode Decomposition," *Symmetry*, vol. 10, no. 11, p. 623, Oct. 2018.

[6] Y. Li, H. Han, and Y. Li, "A New HHT-Based Denoising Algorithm for Financial Time Series Data Mining," 2019 IEEE 8th Joint International Information Technology and Artificial Intelligence Conference (ITAIC), 2019.

[7] S.-L. Lin and H.-W. Huang, "Improving Deep Learning for Forecasting Accuracy in Financial Data," *Discrete Dynamics in Nature and Society*, vol. 2020, pp. 1–12, 2020.

[8] D. Jothimani and A. Başar, "Stock Index Forecasting Using Time Series Decomposition-Based and Machine Learning Models," Lecture Notes in Computer Science Artificial Intelligence XXXVI, pp. 283–292, 2019.

[9] J. Cao, Z. Li, and J. Li, "Financial time series forecasting model based on CEEMDAN and LSTM," *Physica A: Statistical Mechanics and its Applications*, vol. 519, pp. 127–139, 2019.

[10] L. Xian, K. He, C. Wang, and K. K. Lai, "Factor analysis of financial time series using EEMD-ICA based approach," Sustainable Futures, vol. 2, p. 100003, 2020.

[11] Z. Chengzhao, P. Heping, M. Yu, and H. Xun, "Analysis of Asia Pacific stock markets with a novel multiscale model," Physica A: Statistical Mechanics and its Applications, vol. 534, p. 120939, 2019.

[12] Math.NET Numerics. Available online: https://numerics.mathdotnet.com (accessed on 1 July 2019)

[13] Souza, C.R. The Accord.NET Framework. Available online: http://accord-framework.net (accessed on 1 July 2019).

IEEE Third International Conference on Data Stream Mining & Processing
August 21-25, 2020, Lviv, Ukraine

Turkish Text Detection System from Videos Using Machine Learning and Deep Learning Techniques

Jawad Rasheed
Department of Computer Engineering
Istanbul Sabahattin Zaim University
Istanbul, Turkey
0000-0003-3761-1641

Akhtar Jamil
Department of Computer Engineering
Istanbul Sabahattin Zaim University
Istanbul, Turkey
0000-0002-2592-1039

Hasibe Busra Dogru
Department of Computer Engineering
Istanbul Sabahattin Zaim University
Istanbul, Turkey
hasibe.dogru@std.izu.edu.tr

***Abstract*—With the advancement in smart devices and high-speed internet, a continual increase in videos demands an efficient and automatic video indexing and retrieval system. To accomplish it, content-based video indexing is an optimal solution by detecting text in videos. In this study, we proposed a text detection system based on machine learning approaches. We compared conventional machine learning approaches with deep learning method. For deep learning, we implemented Convolutional Neural Network (CNN), while Logistic Regression (LR) and Support Vector Machine (SVM) are employed as conventional machine learning techniques to predict the outcome as text or non-text data. We evaluated the proposed systems on our own dataset obtained from various Turkish videos. LR obtained an overall accuracy of 95.0%, whereas SVM achieved 98.7% while CNN secured 99.8% accuracy. The experimental results show that CNN (deep learning approach) was more effective for our Turkish text dataset as compared to LR and SVM.**

Keywords—text detection, CNN, SVM, LR

I. INTRODUCTION

The progression of world towards 5G data communication system triggers more multimedia data formation on hourly bases. Besides lightning-fast internet, the evolution of smart devices drastically contributed in generation of unstructured and structured audios, videos and images data that eventually requires good indexing and retrieval systems. Business platforms like Cincopa, free video hosting websites like Vimeo, Dailymotion, and Youtube™ or social-media networks like Facebook provide opportunity to publish and share videos among closed groups or public. The user has the hectic responsibility to manually annotate the description of video while uploading it to these web-platforms. The manual annotation sometimes may lead to unaligned description with respect to content present in video. Likewise, it also restricts the searching efficacy, thus strongly demands content-based video indexing and retrieval scheme.

Apart from audio content along video, text appearing within video frames provides important information about visual content that can be adopted for automatic video indexing. These textual content is categorized as scene text or graphical text. Graphical text is the supplementary insertion of textual content at the time of editing, well-known as artificial text or caption text. On the other hand, scene text occurs naturally at the time of recording the video or capturing the image. The text present inside video frames or images varies in fonts, sizes, and style with different positing and alignments. Scene text detection requires more complex job due to less readability as text appears with inconsistent orientation, irregular contrast and uncertain resolution because of capturing conditions and constraints. However, as graphical text is artificially embedded, the editor generally uniformly aligns the text in horizontal or vertical orientation, thus makes detection and extraction relatively easier.

As textual content detection and extraction from videos, images, or documents is an old problem, researchers presented various methods and techniques. These proposed research approaches are normally classified into unsupervised methods like [1] and supervised techniques such as presented in [2]. Unsupervised methods are generally based on connected components approaches like Stroke Width Transformation (SWT) or Maximally Stable Extremal Regions (MSERs). Whereas, supervised models usually based on sliding window that extracts contextual features like color, texture and edges to train the classifier either based on conventional machine learning algorithms like LR, SVM, or deep learning such as CNN.

Scientists designed distinct approaches for text detection with slight variations based on unsupervised methods, like in [1] we performed morphological procedure with some heuristic and geometric constraints to detect and extract Turkish text from news and sports videos. Similarly, authors in [3] calculated local entropy to exploit statistical features of image by horizontal projection analysis and computed gradient difference to get horizontally aligned artificial text in Urdu videos. For English text detection, author in [4] used another unsupervised approach based on Laplacian operator by labeling the candidate text region obtained through maximum gradient difference using k-means clustering. Later Sobel edge maps helped in extracting the boundaries of text fields.

For multilingual text detection researchers suggested different supervised approach like in [2], we presented a CNN model (consists of 3 convolutional layers) as feature extraction approach for Turkish text detection and achieved an overall F-measure of 99.95% on our own dataset created from Turkish videos. Another author proposed a modified version of CNN in [5] called Multi-scale Spatial Partition Network (MSP-Net) to do block-level image classification as text or not-text. It spatially divided feature maps into various block sizes and finally classified those by three fully connected layers with 94.6% F-measure score on TextDis benchmark.

An English and Chinese text detection method introduced in [6], named as Efficient and Accuracy Scene Text (EAST) detector, used Fully Convolutional Network (FCN) and Non-Maximum Suppression (NMS) to predict outcome using

978-1-7281-3214-3/20/$31.00 ©2020 IEEE

multiple channels pixel-level text score maps. Authors tested EAST network on various dataset and obtained F-measure score of 80% on ICDAR2015 and 76.08 on COCO-Text. While a slight variation of one-step NMS with FCN in [7] performed convolutional feature extraction, multi-level feature fusion and multi-task learning and passed the resultant to Recalled-NMS in order to remove quadrilaterals within text spaces that also maintains low confidence text regions. It achieved F-measure of 81%, 74%, and 86% on ICDAR2015, MSRA-TD500 dataset and ICDAR2013 respectively.

Moreover, author in [8] presented six text line detection models based on standard VGG-16 network with some modified convolutional layers that learns from weakly annotated data to predict textual regions at multiple scales with maximum F-measure score of 86.9% on ICDAR2013 dataset. An end-to-end hybrid model named as CRNN in [9] used Deep CNN for feature extraction and Recurrent Neural Network (RNN) for prediction, and achieved 89% F-measure.

In the same way, author in [10] detected Arabic text by training the CNN classifier and Bidirectional Long Short Memory (BLSTM) network with text images of five different orientations. In [11], author employed Convolutional Auto Encoder (CAE) for feature extraction and SVM for candidate region classification as Arabic text or non-text with an F-measure score of 84%.

For Turkish text detection, mostly unsupervised techniques have been employed except in [2]. In [12] author employed connected component and histogram analysis to extract the sliding text, then recognized characters and words by Transformation Based Learning (TBL) that achieved 92% words recognition accuracy and 99% character recognition accuracy. Later author enhance [12] by incorporating Hidden Markov, n-gram language and Glyph models in [13] that did semi-supervised training and accomplished 1% WER while recognizing. In addition, in [14] authors proposed Turkish News videos retrieval system with semantic annotation by segmenting news based on silence periods and performed text detection by vertical and horizontal histogram analysis and connected component.

Literature review depicts that Turkish text detection and recognition is in fancy state, therefore, in this study, we compared three distinct learning-based methods for text detection on Turkish Text dataset used in [2]. These methods includes LR and SVM as conventional machine learning techniques while CNN as deep learning approach. Our aim is to analyze the performance and efficiency of deep learning and traditional machine learning approaches for horizontally aligned artificial text detection in Turkish Videos.

The paper enlists three more sections. The next section presents dataset (materials) used for this study, while section 3 explains learning-based methods performed to detect text. Section 4 summarizes the experimental results. Lastly, section 5 outlines concluding remarks.

TABLE I. DATASET

Labels	Dataset Type		
	Total Samples	***Training Set***	***Testing Set***
Non-text	54680	32808	21872
Text	61131	36679	24452
Total	115811	69487	46324

II. MATERIALS

As text detection for Turkish videos is in fancy state, no benchmark dataset is available publically as per our knowledge. Thus we prepared our own dataset of horizontally aligned Turkish graphical text obtained from various Turkish news, sports and business channels. For information about dataset compilation see Section 3 of [2].

The dataset consists of 115811 grayscale image samples, each of size 32 x 128. Among these, 61131 images belongs to text regions while 54680 instances are non-text images, as summarized in Table 1. The dataset is divided randomly into training and testing sets. 60% of dataset is used to train the models while 40% is used for evaluation purposes. Fig. 1 depicts few samples from dataset.

III. METHODS

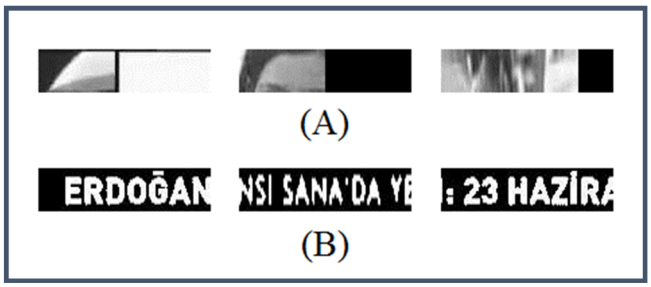


Fig. 1. Samples from dataset, each of size 32 x 128, (A) displays 3 non-text images samples, (B) shows 3 text images samples

A. Logistic Regression (LR)

Logistic Regression is a conventional machine learning algorithm used to classify categorical targets by performing regression analysis with help of sigmoid function as core operation as shown in Fig 2.

It squashes and maps the input values x to range between 0 and 1, then assigns weight w and bias (b_0 and b_1) coefficients to predict the output y as in (1).

$$y = \frac{e^{b0+b1*x}}{1+e^{b0+b1*x}} \quad (1)$$

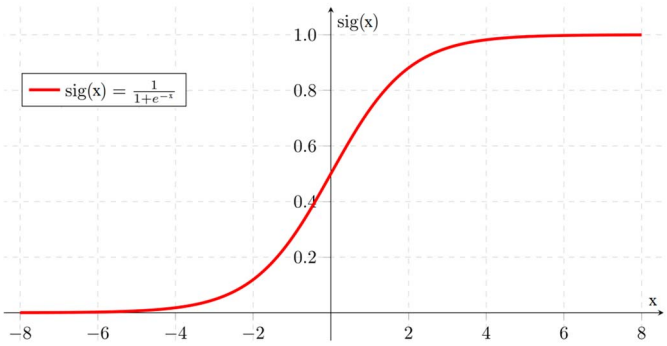


Fig. 2. LR - core function (sigmoid).

B. Support Vector Machine (SVM)

The SVM is a supervised machine learning arsenal preferred as discriminative classifier for various data

classification problems. SVMs are practiced for classification as well as regression tasks by finding the optimal separating hyperplane that distinctly classifies new instances. The hyperplane divides N-dimensional space (N is the number of features) into appropriate labels by finding group of data points (support vectors) that resides on the edge of class descriptors. SVM objective is to discover the best possible

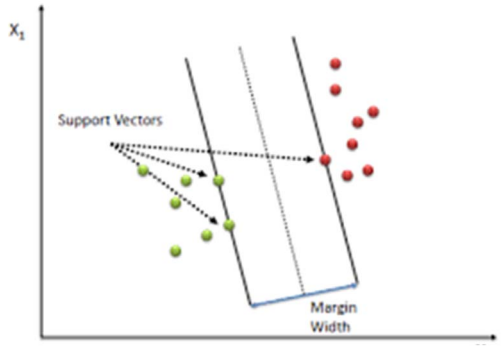


Fig. 4. Hyperplane with support vectors in SVM

hyperplane with maximum margin (distance between data points of classes) as illustrated in Fig. 3.

It requires less computation power as it only retains support vectors (data points that lies on border) while discarding rest of training instances. Initially, algorithm was formulated for data that is linearly separable, but with addition of kernel techniques [15], it can be exercised for non-linear cases by mapping non-linear to linear space. Different kernels have been evolved, but for our analysis, we used Radial Basis Function (RBF). RBF kernel is determined by (2) with as gamma (learnable parameter).

$$K(x_i + x_j) = \exp(-\gamma|x_i - x_j|^2) \qquad (2)$$

C. Convolutional Neural Network (CNN)

CNN is a simple deep learning based algorithm that converts the input data to useful representation to do complex image classification. It usually consists of convolutional layers, pooling layers and fully connected layers (known as dense layers) as depicted in Fig. 4.

Convolutional layers learn the feature representation of input image by assigning appropriate weights and biases to each neuron connecting with some neurons of preceding layer. The resultant is passed to activation function such as Rectified Linear Unit (ReLU). Pooling layer plays role to reduce features dimensionality. In the end, a fully connected layer combines all the learned features of previous layers to formulate the prediction in output layer (classification layer) with help of probabilities computed by softmax layer.

IV. Experimental Results

Each classifier outlined in previous section has few learnable parameters, which are fine-tuned to obtain best results, described in this section. The 115811 samples in Turkish text dataset are split randomly into training set consists of 69487 samples (60% of total dataset) and testing set containing 46324 instances (40% of dataset). Same number of training and testing samples were fed to all three proposed methods and evaluated performance in terms of precision (P), recall (R), F-measure (F), and overall accuracy (OA).

All the experiments were carried out on Jupyter Notebook with Python 3.6 environment along with Keras and Tensor flow packages, running on Nvidia GeForce 410M (512MB RAM) and 2.3GHz Intel Core™ i5 processor of 8GB RAM.

A. Logistic Regression

The LR model is trained with 'lbfgs' optimizer. It achieved an overall accuracy of 95.0% on testing dataset, with recall: 95.0%, precision: 95.0%, and F-measure: 95.0%. Table 2 shows the results obtained for LR classifier.

TABLE II. Classification Accuracy For LR

True Labels	Predicted Labels		Evaluation Results (%)		
	Text	*Non-text*	*P*	*R*	*F*
Non-text	1307	20565	95.3	94.0	94.7
Text	23443	1009	94.7	95.9	95.3
Overall			95.0	95.0	95.0

Fig. 5 shows the Receiver Operating Characteristics (ROC) curve as performance evaluation of LR on Turkish dataset.

B. Support Vector Machine

Classification parameters plays vital role in SVM, therefore we preferred RBF kernel due to its popularity and effectiveness in classification. We empirically calculated and fine-tuned the parameter C=12 and gamma=1.0E4. SVM

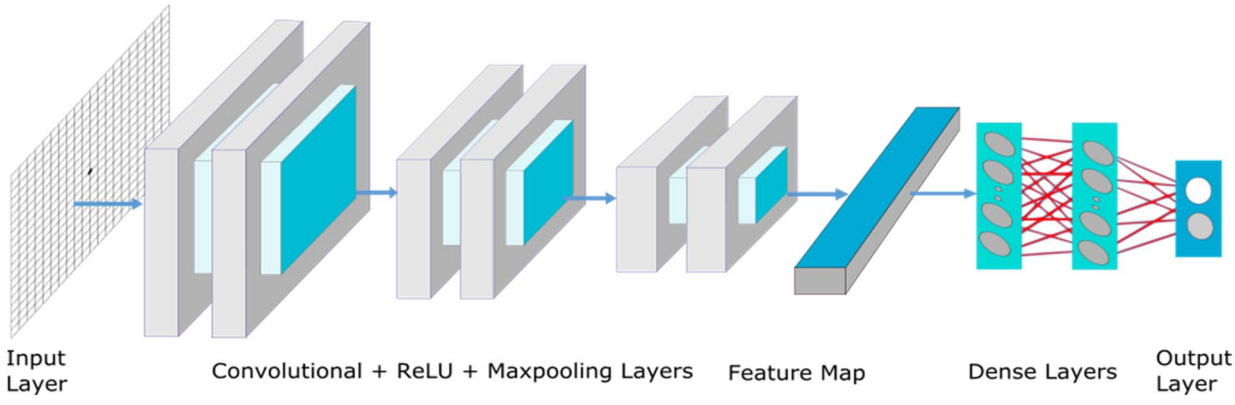


Fig. 3. Network architect of proposed CNN model.

produced promising results with an overall accuracy of 98.7%, recall: 98.7%, precision: 98.8%, and f-measure 98.7%, summarized in Table 3.

TABLE III. CLASSIFICATION ACCURACY FOR SVM

True Labels	Predicted Labels		Evaluation Results (%)		
	Text	*Non-text*	*P*	*R*	*F*
Non-text	399	21473	99.1	98.2	98.6
Text	24249	203	98.4	99.2	98.8
Overall			98.8	98.7	98.7

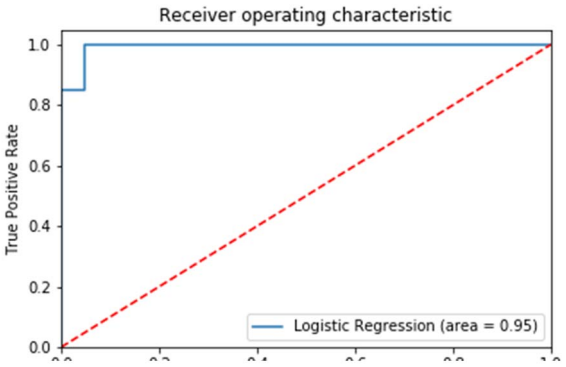


Fig. 6. ROC for LR model.

C. Convolutional Neural Network

For deep-learning approach, we proposed a CNN architect in Fig. 4. Image feature map is fed as input layer (size 32x128) to two consecutive convolutional layers, each with an output shape of 32x128x16, a kernel of size 3x3 and same size padding is turned-on. Followed by 2x2 max-pool layer, and dropout is fixed to 0.10. Next comes two more convolutional layers having a depth of 32 with 3x3 kernel and padding. Again 2x2 max-pool layer is placed for dimensionality reduction and dropout is set to 0.25. Two more convolutional layers with depth of 64 is placed followed by 2x2 max-pool layer and performed a dropout of 0.4. All convolutional layers have ReLU as activation function. The flattened features are fed to fully-connected layer with softmax activation function to predict the output either as text or non-text.

TABLE IV. CLASSIFICATION ACCURACY FOR CNN

True Labels	Predicted Labels		Evaluation Results (%)		
	Text	*Non-text*	*P*	*R*	*F*
Non-text	71	21801	99.8	99.7	99.8
Text	24424	28	99.7	99.9	99.8
Overall			99.7	99.8	99.8

The model is trained with Adam optimizer while learning-rate of 0.001. The proposed CNN model accomplished an overall accuracy of 99.8%, recall: 99.8%, precision: 99.7%, and F-measure: 99.8%. The results are tabulated in Table 4, and Fig. 6 shows the loss and accuracy curves for this trained model.

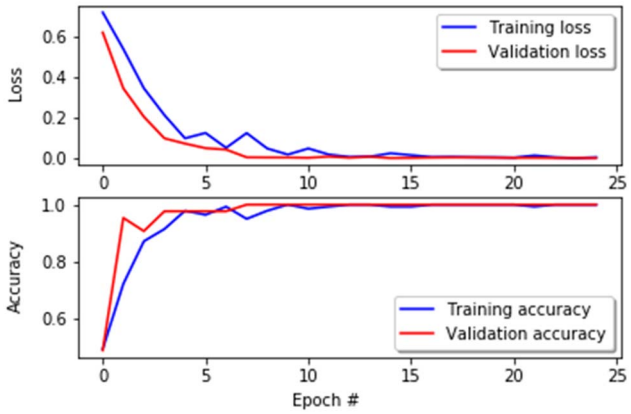


Fig. 5. Loss/Accuracy curves for CNN

V. CONCLUSION

In this study, we performed comparative analysis of deep learning and machine learning algorithms for detecting whether an image has text or not. Even though text detection is language independent, but we focused on Turkish language as it's still an under-stressed area. For this, a dataset compiled from Turkish videos were used to train LR and SVM models as machine learning methods while CNN model was trained as deep learning approach. Experimental results showed that CNN outperformed other two conventional methods on this dataset. CNN achieved an accuracy of 99.8%, while SVM secured 98.7% and LR produced 95.0% prediction accuracy on our dataset. Although methods obtained promising results, but cross-validation is worth exploring in our future work.

REFERENCES

[1] J. Rasheed, A. Jamil, A. Yahyaoui, and A. S. A. Madey, "Automatic Video Indexing and Retrieval System for Turkish Videos Experimental results showed that our proposed method," in *press of The 28th IEEE Conference on Signal Processing and Communications Applications*, 2020.

[2] J. Rasheed, A. Jamil, H. B. Dogru, S. Tilki, and M. Yesiltepe, "A Deep Learning-based Method for Turkish Text Detection from Videos," in *2019 11th IEEE International Conference on Electrical and Electronics Engineering (ELECO)*, 2019, pp. 935–939.

[3] A. Jamil, J. Rasheed, and B. Bayram, "Local statistical features for multilingual artificial text detection from video images," in *International Conference on Advance Technologies, Computer Engineering and Science (ICATCES)*, 2019, no. 2nd, pp. 256–260.

[4] T. Q. Phan, P. Shivakumara, and C. L. Tan, "A laplacian method for video text detection," *Proc. Int. Conf. Doc. Anal. Recognition, ICDAR*, no. January, pp. 66–70, 2009.

[5] X. Bai, B. Shi, C. Zhang, X. Cai, and L. Qi, "Text/non-text image classification in the wild with convolutional neural networks," *Pattern Recognit.*, vol. 66, no. December 2016, pp. 437–446, Jun. 2017.

[6] X. Zhou *et al.*, "EAST: An Efficient and Accurate Scene Text Detector," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, vol. 2017-Janua, pp. 2642–2651.

[7] W. He, X.-Y. Zhang, F. Yin, and C.-L. Liu, "Deep Direct Regression for Multi-oriented Scene Text Detection," in *2017 IEEE International Conference on Computer Vision (ICCV)*, 2017, vol. 2017-Octob, pp. 745–753.

[8] S. Tian, S. Lu, and C. Li, "WeText: Scene Text Detection under Weak Supervision," in *2017 IEEE International Conference on Computer Vision (ICCV)*, 2017, vol. 2017-Octob, pp. 1501–1509.

[9] B. Shi, X. Bai, and C. Yao, "An End-to-End Trainable Neural Network for Image-Based Sequence Recognition and Its Application to Scene Text Recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 11, pp. 2298–2304, Nov. 2017.

[10] S. Bin Ahmed, S. Naz, M. I. Razzak, and R. Yousaf, "Deep learning based isolated Arabic scene character recognition," in *2017 1st*

International Workshop on Arabic Script Analysis and Recognition (ASAR), 2017, pp. 46–51.

[11] O. Zayene, M. Seuret, S. M. Touj, J. Hennebert, R. Ingold, and N. E. B. Amara, "Text Detection in Arabic News Video Based on SWT Operator and Convolutional Auto-Encoders," *Proc. - 12th IAPR Int. Work. Doc. Anal. Syst. DAS 2016*, pp. 13–18, 2016.

[12] E. Dikici and M. Saraçlar, "Sliding Text Recognition in Broadcast News," *IEEE 16th Signal Process. Commun. Appl. Conf.*, pp. 8–11, 2008.

[13] T. Som, D. Can, and M. Saraçlar, "HMM-based sliding video text recognition for Turkish broadcast news," *24th Int. Symp. Comput. Inf. Sci. Isc.*, pp. 475–479, 2009.

[14] D. Küçük and A. Yazici, "Exploiting information extraction techniques for automatic semantic video indexing with an application to Turkish news videos," *Knowledge-Based Syst. Elsevier*, vol. 24, no. 6, pp. 844–857, 2011.

[15] N. Cristianini and J. Shawe-Taylor, *An Introduction to Support Vector Machines and Other Kernel-based Learning Methods*. Cambridge University Press, 2000.

IEEE Third International Conference on Data Stream Mining & Processing
August 21-25, 2020, Lviv, Ukraine

Online Reduced-Order Kernel Regression for Data Processing in Sensor Network

Leonid Lyubchyk
Computer Mathematics and Data Analysis Dept.
National Technical University
"Kharkiv Polytechnic Institute"
Kharkiv, Ukraine
Leonid.Liubchyk@khpi.edu.ua

Olga Kostiuk
Computer Mathematics and Data Analysis Dept.
National Technical University
"Kharkiv Polytechnic Institute"
Kharkiv, Ukraine
Olha.Kostiuk@khpi.edu.ua

Abstract—**Online data processing in sensor networks problem is of particular relevance in connection with the widespread use of such systems for restoring physical fields from local measurements, for example, in environmental monitoring systems. In order to reduce model dimensionality, clustering of sensors net is carried out previously. In this paper, we consider the problem of restore fields of complex structure using data from sensor network. Nonlinear distributed regression kernel-based method is applied using sequential regularization based on optimal estimates concordance. A generalization of this approach to non-stationary case is also considered, for which recurrent regularized kernel-based learning algorithms are proposed, using both moving average and sliding window approaches.**

***Keywords*—data agregation, distributed estimation, kernel method, online learning, recurrent estimation, regularization, sensor network.**

I. INTRODUCTION

Decentralized estimation of multidimensional functional dependencies problem according to the data of distributed measurements has recently gained particular urgency in connection with the widespread use of wireless sensor networks that allow parallel measurements in real time [1]. Such systems are widely used in environmental monitoring systems, control and study of atmospheric phenomena, meteorological observation and many other applications. Typical is the task of estimating two-dimensional temperature or optical fields from distributed spatial measurements [2]. In sensor networks, mutual exchange of information is possible, which can significantly improve the estimation accuracy [3].

The restoration of functions from data obtained by the sensor network is a typical task of multivariate regression, for the solution of which it is advisable to use machine-learning methods. At the same time, it has its own peculiarities related both to the complex, essentially non-linear, nature of the functions being restored. Moreover, only small amount of *a priori* information is available, as well as the nature of the processing of data obtained using a network of a large number of sensors that are heterogeneously distributed in space and connected by network data transmission channels.

Particular difficulties arise when it is necessary to restore in real time non-stationary fields that change in time according to an unknown law.

For estimating of fields parameters from a variety of local observations, kernel machine learning methods have recently been widely used [4, 5]; the advantage of which is the ability to effective restore nonlinear multidimensional functions of a complex structure. Meanwhile, their successful application in information processing tasks in sensor networks is associated with the need to solve a number of additional tasks.

Using the classical kernel approach, the complexity of the regression model is determined by the number of measurement points, which, when using data obtained by a network of sensors, can be extremely large. In this case, preliminary data aggregation with the construction of some sensor clusters is necessary, which will limit the complexity of the estimated model.

The second problem is related to the choice of regularization method and parameters, since the problem under consideration is incorrect and requires special measures to stabilize the solution.

Of particular interest is the development of recurrent methods for estimating the parameters of kernel regressions, which allows real-time processing of the data stream received from the sensor network.

In this paper, we propose a methodology and algorithms for assessing kernel regression from distributed aggregate measurements obtained by preliminary clustering of data [6]. A two-stage regularization procedure is proposed, in which, at the first stage, a simplified linear model of the desired dependence is constructed, and at the second stage, the obtained estimates of linear model parameters are used as *a priori* information for estimating kernel model based on the principle of estimates optimal concordations [7, 8].

A recurrent implementation estimation method based on the use of "sliding window" principle [10,11] is also

978-1-7281-3214-3/20/$31.00 ©2020 IEEE

proposed, which provides real-time estimates of non-linear kernel model from aggregated data stream.

II. Problem Statement

The problem of unknown two-dimensional function $f(\mathbf{x})$ reconstruction using measurements of function values at Cartesian position of sensors $\mathbf{x}_i$ is considered

$$\begin{aligned} &y = f(\mathbf{x}), \quad y_i = f(\mathbf{x}_i) + \xi_i, \ i = \overline{1,n}, \\ &\mathbf{x}_i = (x_i^1, x_i^2)^{\mathrm{T}}, \quad \mathbf{Y}_n = (\mathrm{y}_1, \mathrm{y}_1, ..., \mathrm{y}_n)^{\mathrm{T}}, \end{aligned} \tag{1}$$

Clusterization of multiple sensors into L clusters is preliminarily carried out $\Omega_1, \Omega_2, ..., \Omega_L$, where $|\Omega_k|$ – number of elements in cluster,

$$\overline{\mathbf{x}}_k = \frac{1}{|\Omega_k|} \sum_{i \in \Omega_k} \mathbf{x}_i, \quad \overline{y}_k = \frac{1}{|\Omega_k|} \sum_{i \in \Omega_k} y_i, \tag{2}$$

aggregated measurements averaged over clusters.

Then the model of aggregated measurements takes the following form

$$\overline{\mathbf{X}}_L = (\overline{\mathbf{x}}_1, \overline{\mathbf{x}}_2, ..., \underset{\sim}{\mathbf{x}}_L)^{\mathrm{T}}, \quad \overline{\mathbf{Y}}_L = \overline{\mathbf{X}}_L \cdot \mathbf{w} + \overline{\xi}_L \tag{3}$$

It is necessary, according to the results of aggregated measurements $\{\overline{\mathbf{X}}_L, \overline{\mathbf{Y}}_L\}$, to restore unknown function $f(\mathbf{x})$. Consider both linear and nonlinear model of unknown function

- $f(\mathbf{x}) = \mathbf{x}^{\mathrm{T}} \cdot \mathbf{w}, \quad \mathbf{w} = (w^1, w^2)^{\mathrm{T}},$
- $f(\mathbf{x}) = \boldsymbol{\varphi}^{\mathrm{T}}(\mathbf{x}) \cdot \mathbf{c},$

where $\boldsymbol{\varphi}^{\mathrm{T}}(\mathbf{x}) = (\varphi_1(\mathbf{x}), ..., \varphi_M(\mathbf{x}))$ – coordinate function vector, $(w^1, w^2)^{\mathrm{T}}$, $\mathbf{c}^{\mathrm{T}} = (c_1, ..., c_M)$, $M >> 1$ – vectors of linear and nonlinear models unknown parameters respectively.

III. Linear model estimation

To estimate the parameters of a linear model from aggregated data on cluster centers we will use the regularized functional

$$J(\mathbf{w}) = \left\| \overline{\mathbf{Y}}_L - \overline{\mathbf{X}}_L \cdot \mathbf{w} \right\|^2 + \alpha \cdot \mathbf{w}^{\mathrm{T}} \mathbf{w}, \tag{4}$$

then ridge regularized estimate has the form:

$$\hat{\mathbf{w}} = (\alpha \mathbf{I}_2 + \overline{\mathbf{X}}_L^{\mathrm{T}} \overline{\mathbf{X}}_L)^{-1} \overline{\mathbf{X}}_L^{\mathrm{T}} \overline{\mathbf{Y}}_L. \tag{5}$$

For linear model with bias $f(\mathbf{x}) = \mathbf{x}^{\mathrm{T}} \cdot \mathbf{w} + w_{\mathrm{o}}, \quad \mathbf{w} = (w^1, w^2)^{\mathrm{T}}$ estimates of unknown parameters, taking into account the conditions for their normalization, have the following form:

- Least Squares Estimation (Bias Ridge Regression)

$$\begin{aligned} &\hat{\mathbf{w}} = (\alpha \mathbf{I}_2 + \overline{\mathbf{X}}_L^{\mathrm{T}} \mathbf{P}_n \overline{\mathbf{X}}_L)^{-1} \mathbf{P}_n \overline{\mathbf{X}}_L^{\mathrm{T}} \overline{\mathbf{Y}}_L, \\ &\hat{w}_0 = L^{-1} (\mathbf{I}_2 - \overline{\mathbf{X}}_L^{\mathrm{T}} (\alpha \mathbf{I}_2 + \overline{\mathbf{X}}_L^{\mathrm{T}} \mathbf{P}_n \overline{\mathbf{X}}_L)^{-1} \mathbf{P}_n \overline{\mathbf{X}}_L^{\mathrm{T}}) \overline{\mathbf{Y}}_L, \\ &\mathbf{P}_L = \frac{1}{L} \begin{pmatrix} L-1 & -1 & ... & -1 \\ -1 & L-1 & ... & -1 \\ ... & ... & ... & ... \\ -1 & -1 & ... & L-1 \end{pmatrix}; \end{aligned} \tag{6}$$

- Support vector method estimation

$$\begin{aligned} &\hat{\mathbf{w}} = \theta \cdot \overline{\mathbf{X}}_L^{\mathrm{T}} 1_L^{\mathrm{T}} \mathbf{Q}_L \overline{\mathbf{Y}}_L, \qquad \hat{w}_0 = \gamma \cdot \theta \cdot 1_L^{\mathrm{T}} \mathbf{Q}_L \overline{\mathbf{Y}}_L, \\ &\mathbf{Q}_L = (\gamma \mathbf{I}_L + \overline{\mathbf{X}}_L \overline{\mathbf{X}}_L^{\mathrm{T}})^{-1}, \quad \theta = (1_L^{\mathrm{T}} \mathbf{Q}_L 1_L)^{-1}. \end{aligned} \tag{7}$$

The obtained estimates are used as *a priori* information at the second stage of model estimation.

IV. Nonlinear kernel model estimation

In accordance with the general methodology of kernel estimation, we choose the coordinate functions from the following condition

$$\boldsymbol{\varphi}^{\mathrm{T}}(\mathbf{x}) \cdot \boldsymbol{\varphi}(\overline{\mathbf{x}}_k) = \kappa(\mathbf{x}, \overline{\mathbf{x}}_k) = \exp(-\mu \left\| \mathbf{x} - \overline{\mathbf{x}}_k \right\|^2), \quad \mu > 0, \tag{8}$$

Then aggregated measurements takes the form

$$\begin{aligned} &\overline{\mathbf{Y}}_L = \boldsymbol{\Phi}_L \mathbf{c} + \overline{\xi}_L, \ \boldsymbol{\Phi}_L^{\mathrm{T}} = (\boldsymbol{\varphi}(\mathbf{x}_1), .., \boldsymbol{\varphi}(\mathbf{x}_L)), \\ &\overline{\xi}_L = (\overline{\xi}_n, .., \overline{\xi}_L), \quad \overline{\xi}_k = \frac{1}{|\Omega_k|} \sum_{i \in \Omega_k} \xi_i. \end{aligned} \tag{9}$$

We obtain estimates of the model parameters from the minimization of regularized functional

$$I(\mathbf{c}) = \left\| \overline{\xi}_L \right\|^2 + \gamma \left\| (\mathbf{c} - \mathbf{c}_0) \right\|^2, \quad \overline{\xi}_L = \overline{\mathbf{Y}}_L - \boldsymbol{\Phi}_L \mathbf{c}. \tag{10}$$

Solution of corresponding optimization problem is:

$$\begin{aligned} &\hat{\mathbf{c}} = \mathbf{c}_0 + \boldsymbol{\Phi}_L^{\mathrm{T}} \mathbf{A}_\gamma^{-1} (\overline{\mathbf{Y}}_L - \boldsymbol{\Phi}_L \mathbf{c}_0), \\ &\mathbf{A}_\gamma = \gamma \mathbf{I}_L + \mathbf{K}_L, \ \mathbf{K}_L = \boldsymbol{\Phi}_L \boldsymbol{\Phi}_L^{\mathrm{T}} = \left(\kappa(\overline{\mathbf{x}}_i, \overline{\mathbf{x}}_j) \right)_{i,j=1}^{L,L}. \end{aligned} \tag{11}$$

To obtain *a priori* values of nonlinear kernel model parameters, we use obtained estimates of linear approximation model parameters.

In accordance with the principle of matching estimates, we introduce functional [6 - 8]:

$$I_0(\mathbf{c}_0) = \left\| \overline{\zeta}_L \right\|^2 + \omega \left\| \mathbf{c}_0 \right\|^2, \ \overline{\zeta}_L = \overline{\mathbf{X}}_L \hat{\mathbf{w}} - \boldsymbol{\Phi}_L \mathbf{c}_0 \tag{12}$$

Using Lagrange function

$$L(\mathbf{c}_0, \overline{\zeta}_L, \nu) = I_0(\mathbf{c}_0) + \nu^{\mathrm{T}} (\overline{\mathbf{X}}_n \mathbf{w} - \boldsymbol{\Phi}_n \mathbf{c}_0 - \overline{\zeta}_L) \tag{13}$$

we obtain *a priori* values of nonlinear model parameters

$$\mathbf{c}_0^* = \mathbf{\Phi}_n^{\mathrm{T}} \mathbf{A}_\omega^{-1} \mathbf{X}_L \hat{\mathbf{w}}, \ \mathbf{A}_\omega = \omega \mathbf{I}_L + \mathbf{K}_L. \qquad (14)$$

Finally, non-linear kernel regression model is:

$$\begin{aligned} &\hat{f}(\mathbf{x}) = \sum_{i=1}^{L} g_i \cdot \kappa(\mathbf{x}, \bar{\mathbf{x}}_i), \\ &g_i = \mathbf{A}_\gamma^{-1} \bar{\mathbf{Y}}_L + (\mathbf{I}_L - \mathbf{A}_\gamma^{-1} \mathbf{K}_L) \mathbf{A}_\omega^{-1} \bar{\mathbf{X}}_n \hat{\mathbf{w}}. \end{aligned} \qquad (15)$$

Thus, regularized kernel estimate of multidimensional regression depends only on the measured variables.

V. Recurrent on-line model estimation

Consider the measurements sequence in the form of data stream $\{y_k\}_{k=1}^n$, $y_k = f(x_k) + \varepsilon_k$, $k = 0, \ldots$ where $f(\cdot)$ is an unknown nonlinear function, and ε_k is a noise discrete random process, $\mathbf{E}\{\varepsilon_k\} = 0$, $\mathbf{E}\{\varepsilon_k^2\} = \sigma^2$.

In the kernel-model framework, identification model is $y_k = \hat{f}(x_k) = \phi^{\mathrm{T}}(x_k) c_k + \varepsilon_k$, $k = \overline{0,n}$,

In accordance with kernel-model technique, introduce kernel matrix $\mathbf{K}_n = \mathbf{\Phi}_n^{\mathrm{T}} \mathbf{\Phi}_n$, , which may be computed directly without reference to the feature vectors, because $\mathbf{K}_n = \left\| k_{i,j} \right\|$, $k_{i,j} = \kappa(x_i, x_j)$, $i, j = \overline{1,n}$.

Takes into consideration the auxiliary (dual) variables vector $\boldsymbol{\lambda}_n$, such as $c_n = \mathbf{\Phi}_{n-1} \boldsymbol{\lambda}_n$ [9].

At that measurements prediction is

$$\hat{y}_n = \hat{f}(x_n) = \varphi^{\mathrm{T}}(x_n) \mathbf{\Phi}_{n-1} c_n = \mathbf{k}_{n-1}^{\mathrm{T}}(x_n) \boldsymbol{\lambda}_n, \qquad (16)$$

where $\mathbf{k}_{n-1}^{\mathrm{T}}(x_n) = (\kappa(x_n, x_0) \ (\kappa(x_n, x_1) \ldots \kappa(x_n, x_{n-1}))$ - kernel vector, $\hat{y}_k$ - vector of measurements prediction, λ_n is an auxiliary variables vector estimate at instant n, which should be obtained by training sample $\{y_k, x_k\}_{k=0}^{n-1}$, and in accordance with "kernel trick", express in terms of only kernel matrix $\mathbf{K}_{n-1}$.

Recurrent kernel-model identification algorithms should ensures on-line $\boldsymbol{\lambda}_n$ estimates. Sliding window kernel model approach [10, 11] consider for estimation at instant n only last $n - s$ observations, so observation vector is $\mathbf{y}_{n,s} = (y_{n-s+1} \ldots y_n)^{\mathrm{T}}$. Consequently, kernel matrix $\bar{\mathbf{K}}_{n,s} = \mathbf{\Phi}_{n,s}^{\mathrm{T}} \mathbf{\Phi}_{n,s}$ has a fixed dimension $(s \times s)$ and observation equation takes the form $\mathbf{y}_{n,s} = \mathbf{\Phi}_{n-1,s}^{\mathrm{T}} \mathbf{w} + \boldsymbol{\varepsilon}_{n,s} = \mathbf{K}_{n-1,s} \bar{\boldsymbol{\lambda}}_n + \boldsymbol{\varepsilon}_{n,s}$.

Consider the "sliding" estimation cost function includes at any instant $n+1$ *a priori* information term determined by previously estimate. Using the representation $\mathbf{w} = \mathbf{\Phi}_{n,s} \bar{\boldsymbol{\lambda}}$, the optimization problem is:

$$\left\| \mathbf{y}_{n+1,s} - \mathbf{K}_{n,s} \bar{\boldsymbol{\lambda}} \right\|^2 + \gamma^{-1} (\bar{\boldsymbol{\lambda}} - \bar{\boldsymbol{\lambda}}_n)^{\mathrm{T}} \mathbf{K}_{n,s} (\bar{\boldsymbol{\lambda}} - \bar{\boldsymbol{\lambda}}_n) \to \min_{\bar{\boldsymbol{\lambda}}}. \qquad (17)$$

Condition of optimality leads to normal equations:

$$\left(\bar{\mathbf{K}}_{n,s}^{\mathrm{T}} \bar{\mathbf{K}}_{n,s} + \gamma^{-1} \bar{\mathbf{K}}_{n,s} \right) \bar{\boldsymbol{\lambda}}_n = \bar{\mathbf{K}}_{n,s}^{\mathrm{T}} \mathbf{y}_{n+1,s} + \gamma^{-1} \bar{\mathbf{K}}_{n,s} \bar{\boldsymbol{\lambda}}_n. \qquad (18)$$

Recurrent dual vector estimate takes the form:

$$\bar{\boldsymbol{\lambda}}_{n+1} = \left(\gamma^{-1} \mathbf{I}_S + \overline{\mathbf{K}}_{n,s} \right)^{-1} \left(\gamma^{-1} \bar{\boldsymbol{\lambda}}_n + \mathbf{y}_{n+1} \right) \qquad (19)$$

At last, it is necessary to put forward the updating algorithm for inverse regularized sliding kernel matrix $\bar{\mathbf{K}}_{n,s}^{-1}(\gamma)$. Using the approach, proposed in [11, 12], two step inverse regularized kernel matrix updating algorithm $\bar{\mathbf{K}}_{n-1,s}^{-1}(\gamma) \to \bar{\mathbf{K}}_{n-1,s-1}^{-1}(\gamma) \to \bar{\mathbf{K}}_{n,s}^{-1}(\gamma)$ was proposed, which use "downsizing" matrix $\overline{\mathbf{K}}_{n-1,s-1}(\gamma)$, determined from the sliding kernel matrix representation.

Using formula (16) and the obtained estimates of dual variables (19), we obtain an estimate of nonlinear model in the form of linear combination of kernel functions, localized in the centers of clusters of sensor network.

VI. Setting the aggregation centers

A significant increase of distributed reconstruction accuracy can be achieved through rational clustering of sensors in the network, taking into account the quality of the estimation by optimizing data aggregation centers.

Recurrent center optimization procedure is as follows:

$$\bar{x}(n+1) = \bar{x}(n) + \gamma(n) \cdot \nabla_x \mathbf{\Psi}(\bar{x}(n)), \qquad (20)$$

where $\gamma(n)$ is the experimentally selected step factor.

As a functional for setting up cluster centers, we take

$$\begin{aligned} \Psi(\bar{x}(n)) = &\sum_{m=1}^{L} g_m \cdot \exp\{-\mu \left\| \bar{x}_k(n) - \bar{x}_m(n) \right\|^2\} + \\ &+ \sum_{m=1}^{L} \left\| \bar{x}_k(n) - \bar{x}_m(n) \right\|^2. \end{aligned} \qquad (21)$$

At that, to adjust the *k*-th center, it is necessary to take the sum over all $m \neq k$, but when differentiating, the term with the index $m = k$ will disappear.

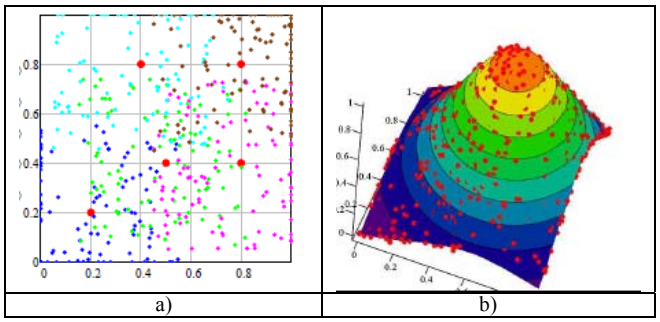
As a result, the formation of aggregated groups of sensors is carried out according to the following rule:

$$\Omega_k(n) = \{x_i\}_k, \ \left\| x_i - \bar{x}_k(n) \right\|^2 < \left\| x_i - \bar{x}_m(n) \right\|^2, \ m \neq k. \qquad (22)$$

VII. Numerical experimet

The obtained results was illustrated by a series of computational experiments on the restoration of a two-dimensional function from measurements, taken by a network of sensors on a plane.

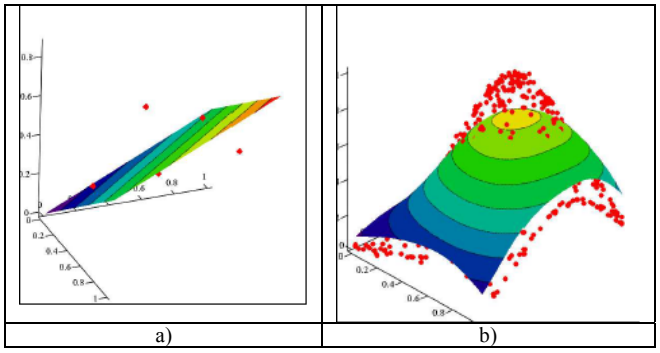
In Fig. 1 shows the distribution of sensors with the formation of clusters a) and a model function with the measurements results b).



a) b)

Figure 1. Initial data for regression recovery

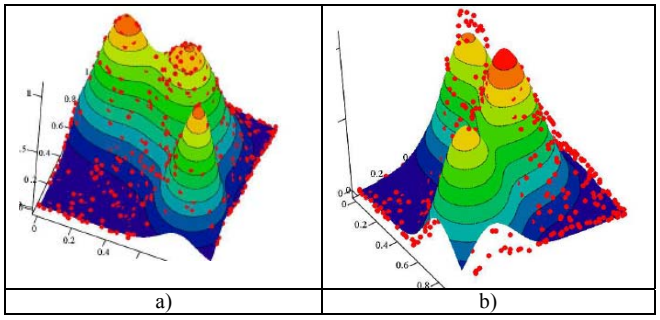
Fig. 2 presents the results of estimation of linear a) and nonlinear kernel model b).



a) b)

Figure 2. Regression recovery results

In Fig. 3 presents the results of evaluating a complex nonlinear kernel model, a) measurement data of model function, b) recovery results.



a) b)

Figure 3. Nonlinear regression recovery results

The experimental results indicate the efficiency of the proposed approach.

VIII. CONCLUSION

A method of multidimensional nonlinear regression for complex functions restoration using measurements, obtained by sensor network is proposed. To reduce the dimension of kernel model, preliminary clustering of the sensor network is carried out. A feature of the method is the implementation of two-level regularization, when estimates of linear model parameters are used as *a priori* information for constructing estimates of non-linear model by optimal estimates concordance.

Further research may be related to construction of "second level" algorithms for tuning regularization parameters in real time. For this purpose, it seems to be efficient to use the dynamic cross-validation method [13, 14], which makes it possible to assess the quality of non-stationary models, obtained using data stream data from sensor network.

REFERENCES

[1] Handbook of Wireless Sensor Networks: Issues and Challenges in Current Scenario's. Editors: Singh P.K., Bhargava B. K., Paprzycki M., Kaushal N.C. Hong, W.-C. (Eds.). Springer, 2020.

[2] W. Gao, J. Chen, C. Richard, and J. Huang, "Diffusion adaptation over networks with kernel least-mean-square," in IEEE CAMSAP, 2015.

[3] Fahmy H. M. A. Wireless Sensor Networks. Concepts, Applications, Experimentation and Analysis, Springer, 2016.

[4] Ban-Sok Shin, Henning Paul, and Armin Dekorsy. "Distributed Kernel Least Squares for Nonlinear Regression Applied to Sensor Networks". 2016 24th European Signal Processing Conference (EUSIPCO), 2016.

[5] Paul Honeine, Mehdi Essoloh, C´edric Richard and Hichem Snoussi. "Distributed regression in sensor networks with a reduced order kernel model", IEEE Global Telecommunications Conference (GLOBECOM), 2008.

[6] V. Strijov, V. Shakin, "Index construction: the expert-statistical method", Environmental research, engineering and management, No. 4 (26), pp.51-55, 2003.

[7] L. Lyubchyk and G. Grinberg, "Preference function reconstruction for multiple criteria decision making dased on machine learning Approach", Recent developments and new directions in Soft Computing, L.A. Zadeh et al. (Eds), Springer, pp. 53-63, 2014.

[8] L. Lyubchyk and G. Grinberg, "Online Ranking Learning on Clusters", In Proceeding of 2-nd IEEE International Conference on Data Stream Mining & Processing, Lviv, Ukraine, 2018.

[9] C. Richard, J. C. M. Bermudez and P. Honeine, "Online prediction of time series data with kernels," IEEE Trans. on Signal Processing, vol. 57, no. 3, pp. 1058–1067, 2009.

[10] Y. Engel, S. Mannor and R. Meir, "The kernel recursive least-squares algorithm", IEEE Transactions on Signal Processing, 2004, vol. 52, pp. 2275 - 2285.

[11] Van Vaerenbergh, V. Javier, I. Santamar. "Nonlinear System Identification using a New Sliding Window Kernel RLS Algorithm ", Journal of Communications, 2007, Vol. 2, No. 3.

[12] L. Lyubchyk, V. Kolbasin and R. Shafeev, "Nonlinear signal reconstruction based on recursive moving window kernel method", In 8-th International Conference on Intelligent Data Acquisition and Advanced Computing Systems (IDAACS'2015), Warsaw, Poland, 2015.

[13] M.Karn, P. Nedoma, and V.Smidl, "Cross Validation on Controlled Dynamic Models: Bayesian ApproachH", 16th Triennial World Congress, Prague, Czech Republic, 2005.

[14] C. Bergmeir and J. Benítez, "On the use of cross-validation for time series predictor evaluation", Information Sciences, v.191, pp. 192–213, 2012.

IEEE Third International Conference on Data Stream Mining & Processing
August 21-25, 2020, Lviv, Ukraine

Improvement of the Algorithm of Determination Parameter of Photography in the Conditions of Light Sensitivity

Maksym Korobchynskyi
Military-Diplomatic Academy named after Eugene Bereznyak
Kyiv, Ukraine
maks_kor@ukr.net

Mykhailo Slonov
Military-Diplomatic Academy named after Eugene Bereznyak
Kyiv, Ukraine
slonovmu@gmail.com

Myhailo Rudenko
Military-Diplomatic Academy named after Eugene Bereznyak
Kyiv, Ukraine
ruminik33@ukr.net

Oleksandr Maryliv
Military-Diplomatic Academy named after Eugene Bereznyak
Kyiv, Ukraine
delavar65000@gmail.com

Valentyn Pylypchuk
Military-Diplomatic Academy named after Eugene Bereznyak
Kyiv, Ukraine
olserg1982@gmail.com

Volodymyr Moldovan
Military-Diplomatic Academy named after Eugene Bereznyak
Kyiv, Ukraine
vl.mold@ukr.net

***Abstract*—The concept of critical modes of observation of an object with the help of discrete photographic equipment and their influence on the quality of a digital image is defined. A feature of such modes is the inability to obtain a guaranteed quality of the digital image of the object of observation. During photography, this is characterized by a significant change in the level of external light sensitivity of the subject, the possible distance of photography, the operational shift of the image and other factors that affect the image quality. But observation in these conditions increases the activity, completeness and timeliness of documenting the image of the object. The article offers a formalized description of the successful observation of the object. It contains the concept of a given probability of recognition of the object of observation of its image. It determines the rational level of image quality. The parameters of the photographic equipment which can provide the predicted qualitative image are offered. Such parameters of the photographic equipment are its light sensitivity and resolution. The algorithmic relationship between the limit values of light sensitivity and resolution is determined. It allows assessing the degree of achievement of the highest perfection of a discrete iconic tool, taking into account the specified depth of recognition of the object by its image. An algorithmic assessment of the influence of photographic noise on image quality is proposed. The algorithms developed in the article are verified by a computational experiment. It allowed us to estimate the rational ranges of light sensitivity that won't affect the resolution of the digital image of the observed object.**

Keywords—observation, photography, light sensitivity, photographic equipment, resolution, discrete visual means, object of observation.

I. INTRODUCTION

Documentation of the image of the object of observation is needful to solution problems in agriculture, ecological monitoring, geodesy, mapping, etc. Successful solution of such problems allows carrying out observations with the minimum influence of external conditions on the received results. There are conditions in which automatic systems for adjusting the parameters of discrete species surveillance work without taking into account the quality of the digital image. Under these conditions, it isn't possible to guarantee the receipt of information about objects with a given depth of recognition of its image. As a result, it is impossible to predict the results of observation [1–2]. When observing by photography, such limitations are characteristic in the case of changes in the level of external lighting of the subject, the operational shift of the image, the possible distance of photography and other factors that affect the image quality.

The image quality decreases sharply due to its pixelation when observations need to be made at dusk. This is a consequence of a significant increase in photographic noise [3, 4]. Photography from the maximum possible distance requires the use of maximum optical zoom. In this case, the minimum resolution and angle of view of the lens limit the ability to control the quality of the digital image on its image on the display. The ability to take high-quality photo image at dusk from the maximum possible distance increases the activity and timeliness of documenting the image of the object. Photography with predicted image shift will allow observing from the vehicles without losing the specified image quality. Thus, the task of predicting image quality when photographing in its atypical conditions is relevant.

In [5–7] the general questions of increase of possibilities of recognition of the image of object during photography are considered. In [2, 8] are presented an algorithm for predicting the recognizable properties of photographic equipment (PE) with taking into account the operational shift of the image and the results of checking its adequacy for photographing objects in real conditions. Features of photography of a distant object are considered in [6].

For photography in low light conditions, the hardware recommendations are described in [9]: to increase the light sensitivity of the receiver, which is software implemented in the PE for the shooting mode "Auto". Adherence to the above recommendations makes some sense but only in a limited range of image pixelation intensity. Such an effect hasn't been evaluated, although it is known that visual pixelization of the image significantly reduces its quality.

Thus, determining the features of the relationship between light sensitivity and resolution of the PE receiver in

978-1-7281-3214-3/20/$31.00 ©2020 IEEE

critical photography modes is an unsolved part of the problem of obtaining the predicted image quality.

The purpose of the article is determination relationship between the limits of light sensitivity and resolution in photography to perform tasks and identify ways to expand the capabilities of photography of the object in external conditions.

II. METHOD

Today PE is a complex and perfect software and hardware complex. It documents the objects observed in different environmental conditions. Taking into account the conditions of observation in the PE is due to automatic or manual adjustment of its parameters.

A formalized description for observation with the help of PE of the j-th object ($j = 1, \ldots, J$) from the distance S_j and the given reliability of its results (the probability P of the image recognition isn't less than its setpoint P_s) will look like [2]:

$$\begin{cases} P_S \geq P, \\ P = P\left[S_j, d_{PE}, f, l_j = \min, j = 1, J\right] \end{cases}, \quad (1)$$

where d_{PE} is the linear size of the PE distinguishing element (matrix pixel); f is the focal length of the PE lens; l_j is the linear size of the recognition feature of the object.

In (1) combines the requirements to the rational image quality of the image (P), the conditions of photography (S), the features of the object of photography (l) and the desired values of the influential parameters of PE (d_{PE}, f). Compliance with (1) will allow predicting the observation of an object with a given level of image quality.

The modes of automatic operation of the PE in atypical conditions are critical, in which the photographic image in the PE is remembered, but its expected quality (according to the set pixel size of the image) isn't guaranteed. For such cases, the household PE provides manual control of photography parameters. In this case the photographer must have sufficient personal experience of such setting. It is necessary to be guided in the circumstances that translate the normal mode of operation of the PE to a critical. In addition, manual adjustment is provided only in professional photo cameras. This feature isn't available in PE, in which the function of photography is secondary. In professional species equipment critical operating modes aren't usually distinguished.

Thus, it is important to determine the limits of PE parameters and the relationship between them, which will provide a predictable quality image. Necessary in this case are the following parameters of the PE namely its light sensitivity S_{ISO} and resolution d. The resolution of PE d is considered to be the linear size $d_{PE} = d$ of the element of the light-sensitive matrix (matrix pixel) – the current value of the resolution, which remembers the image according to the line of values in a particular sample. The intensity of the receiver's response to the input signal is characterizes by light sensitivity S_{ISO}. When taking a photo, the input signal is the level of illumination of the receiver with light reflected or emitted from the subject. The output signal is the amount of response of the photo receiver in the form of an image on the display.

At the limit of detection of the image segment E, the output signal i and from m pixels (elementary receivers of the light-sensitive matrix – photo detector) will be equal to:

$$i = \frac{m_1}{m} i_1, \quad (2)$$

where m_1 is the number of elementary receivers with the output signal from them i_1.

If $m_i < m_1$ the segment isn't recognized, if $m_i > m_1$ the segment is recognized better. When $m_i = m_1$ the value of i will be the criterion and $i = i_{cr}$. The value of m_i is determined by the difference in photography or its dependent value, namely the resolution. The maximum resolution R of the light-sensitive matrix with an elementary receiver with side d will be:

$$R = \frac{1}{2d}. \quad (3)$$

The number m of elementary receivers in the segment E in its square configuration with side $l_j = l$ will be equal to the following:

$$m = \left(\frac{l}{d}\right)^2 = 4l^2R^2. \quad (4)$$

The minimum signal of the each elementary receiver i_1 depends on its area Q, the material of manufacture β and technological efficiency of manufacture and operation η:

$$i_1 = \beta\eta Q = \beta\eta d^2. \quad (5)$$

Given the relations (3)–(5) of (2) can be reduced like:

$$m_1 = \frac{i}{i_1} m = \frac{i}{\beta\eta d^2} = \frac{l^2}{d^2} = \frac{il^2}{\beta\eta d^4}. \quad (6)$$

If the (6) signal from m_1 receivers exceeds the criterion value H_{cr} which determines the light sensitivity of the S_{ISO} photo detector the exposure H of the segment E will be sufficient. If the exposure of the segment has acquired a criterion value (the criterion value has acquired a signal $i = i_{cr}$ from segment E) we will have:

$$H_{cr} \equiv m_1 = \frac{i_{cr} l^2}{\beta\eta d^4}. \quad (7)$$

The light sensitivity of S_{ISO} is determined as follows:

$$S_{ISO} = \frac{k}{H_{cr}}. \quad (8)$$

The coefficient k is a constant and is given in the regulatory documentation. It is necessary to combine (7), (8) and bring to the following form:

$$\frac{S_{ISO}}{d^4} = \frac{k\beta\eta}{i_{cr} l^2} = \text{const}, \quad (9)$$

or

$$S_{ISO} R^4 = \frac{k\beta\eta}{i_{cr} l^2} = \text{const}. \quad (10)$$

Equation (9) and (10) indicates that the image of the recognition feature l of the object will be detected at the current level of perfection of PE (i_{cr}, β, η, k) and photo shooting conditions (η, k) only when the ratio between the

light sensitivity of the photo detector and its resolution of a certain value. This confirms the previous conclusion about the inexpediency of an infinite increase in light sensitivity when photographing at dusk. A rational way is to increase the size d of the elementary receiver. For PE this will correspond to the transition to a smaller pixel size of the light-sensitive matrix.

Given (9) the condition of photography can be considered inequality:

$$\frac{S_{ISO}}{d^4} \geq \frac{k\beta\eta}{i_{cr}l^2}. \quad (11)$$

Indicators of perfection (perfection of the selected light-sensitive material β, perfection of manufacturing technology η, detection of caviar, standard coefficient k) are in the right part of (11). They are constant values for each individual case of photography. Thus, the whole right part is a constant of the object and the level of perfection of the PE.

The PE parameters that change for each PE type and sample are in the left part. Therefore, according to (11), in order to achieve a smaller resolution d in the image, it is possible only by proportionally reducing the light sensitivity of the S_{ISO}.

Given the above, (11) characterizes the practical possibility of achieving the highest perfection of the iconic means, in this case – the PE sample. If for a particular sample of PE the equality between the left and right parts of (11) is preserved – the limit values of distinction and light sensitivity are reached. Further studies on the possibility of achieving greater S_{ISO} with the achieved resolution d may not be successful. It is necessary to move to fundamentally different light-sensitive materials, technologies, methods of obtaining and processing images.

III. Experiment And Discussion

According to (11), there will be the following limit between the limit values of light sensitivity and PE resolution:

$$S_{ISO} \geq d^4 \frac{k\beta\eta}{i_{cr}l^2}. \quad (12)$$

Determine the possible values of photography parameters in the right part of (12). The coefficient of proportionality k for household photo cameras is equal to 1 [10, 11]. The material of elementary receivers determines the quantum efficiency of the exposure process, that is, the ratio between the number of photons that hit the light-sensitive surface and the number of photoelectrons β created by them. Will assume that each photon initiates the emergence of one photoelectron (β = 1). The coefficient η is its efficiency and characterizes the technological efficiency of manufacture and operation of PE. The maximum possible value in this case will be one.

The value i_{cr} can be accessed through the contrast sensitivity of the human eye and will be 0.02 [12, 13].

The image length of the recognition feature l will be nd, where n is the number of pixels in the image. It will be $3d$ for detection, $6d$ for recognized and $12d$ for identified the object. Given the above, we move from (11) to the following:

$$S_{ISO} \geq \frac{d^4}{0.02n^2d^2} = \frac{d^2}{0.02n^2}. \quad (13)$$

Let's perform mathematical calculations for the Canon Power Shot SX420 IS Black [14]. Its light sensitivity of Canon Power Shot SX420 IS Black will fluctuate in the range of S_{ISO} = 100 … 1600 units ISO, the pixel size of the light-sensitive matrix will be 20 Mpx, the pixel size of the image will be 5152 × 3864 px [15]. The smallest value of the resolution will be equal to 2 μm = 2 × 10^{-6} m. Mathematical calculations of the limit light sensitivity when photographing with different depth of recognition presented in Fig. 1.

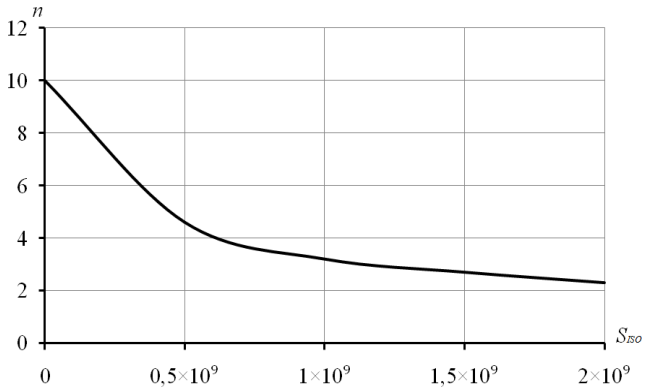


Fig. 1. The relationship between light sensitivity and the specified depth of recognition image.

The results of the calculations using (11) and (13) shown in the graph show that together with the change of the task to the depth of recognition of the object by its image, the requirements for the limit value of light sensitivity of the PE change. When changing the detection task for identification the limit light sensitivity will change more than an order of magnitude.

Calculations of the limit light sensitivity don't match the passport data for this type of photo camera. This is due to the far from optimal technology of manufacture and operation of the PE, opportunities for further improvement of the light-sensitive receiver, the influence of operational factors, especially in critical modes of operation of the PE.

Photographic equipment manufacturers are offered a procedure for changing the requirements for light sensitivity, which can be considered rational in terms of its relationship with distinction and capabilities. Depending on the task of the depth of recognition of the object by its photographic image, the requirements for the limit light sensitivity vary slightly more than an order of magnitude. Such possibilities for change of light sensitivity are provided also in the Canon Power Shot SX420 IS photo camera. The same applies to the PE installed in the Galaxy A50 Smartphone [16].

In Coolpix AW100 [17], Panasonic TZ18 Lumix [18] photo cameras, the situation is different, where the specified value changes by two or more orders of magnitude. Thus, for predicted high-quality photographs, it is important to observe the range of calculated changes in the limit values of light sensitivity when photographing.

The generally accepted approach to exposure management in PE allows changing its value within 3 … 4 orders of magnitude, namely to ensure the efficiency of PE to the level of illumination in 20 … 30 lux. In summer it corresponds to 19 … 20 o'clock in the evening and 6 … 7 o'clock in the morning. This illumination of the object is provided by a table lamp, which is installed at a distance of

1.0 … 1.5 m from it. However, it is desirable to take photos in the darker hours of the day, with a given or predicted image quality. Small contrast in the photo image will be more appropriate to the real conditions of photography. For example, the statistically determined average contrast value of a photo image in the corner of the lens field of view in PE "Landscape" photography is 0.18.

We can compare the required range of changes in light sensitivity in the PE with the real possible value. The illumination of the Earth's surface on a sunny summer day varies from $E_{max} = 10^5$ lux (12 … 13 hours) to $E_{min} = 0.22 \ldots 0.32$ lux on a lunar cloudless night. The dynamic luminance range of the E_{max} / E_{min} object will be $2.8 \times 10^5 \ldots 5.1 \times 10^5$ it's about five orders of magnitude. Accordingly, the parameters of the PE that determine the exposure of the image must be changed the same number of times. Approximately 1 order in exposure control can be provided by changing the shutter speed from 1/500 sec to 1/50 sec. Higher shutter speeds are possible in the PE but for shooting conditions their use will result in a significant image shift. By changing the aperture number in the range of 2.4 ... 32.0 and taking into account the quadratic effect on the exposure value it can be controlled within at least two orders of magnitude.

Also, the exposure control parameter is the light sensitivity of the receiver. According to Fig. 1, the multiplicity of light sensitivity change even for high-contrast images shouldn't exceed one order of magnitude.

IV. Conclusion

Thus, the article defines the relationship between light sensitivity and the difference when observing an object in critical photography conditions. This makes it possible to obtain a guaranteed and planned level of information about the object of observation by photography. Algorithm of determination parameter of photography in the conditions of light sensitivity is determined. This allows you to assess the degree of achievement of the highest perfection of the discrete iconic tool, taking into account the specified depth of recognition of the object by its image. Developed in the article algorithm are verified by a computational experiment. This allowed estimating the rational ranges of light sensitivity, which won't affect the resolution of the image of the observed object. Is estimated the impact on the depth of recognition of the object by its image of the photosensitivity of the camera in order to maintain the minimum possible value. The range of change of sensitivity of the camera by the order provided in photo cameras of the leading firms-manufacturers can be considered recommended. Is estimated the multiplicity of light sensitivity change for images of different contrast. To solve the problem of observation by photography under critical conditions can offer the following:

- Search for more light-sensitive substances for matrix receivers.
- Development of algorithms for finding compromise modes of operation of photographic equipment in critical conditions with minimal impact on the quality of the photographic image.

Further research in this area should focus on the development of hardware and software algorithms that are adapted to the causes of each case of critical conditions increase the accuracy of determining the relationship between the limit values of light sensitivity and resolution of the photo camera.

References

[1] M. Korobchynskyi, M. Slonov, M. Rudenko, O. Maryliv, "Assesment of the Effect of Image Shift on the Results of Photo-Video Recording", Conference proceedings, 2020 IEEE 40th International Conference on Electronics and Nanotechnology (ELNANO), Igor Sikorsky Kyiv Politechnic Institute, pp. 641–645, April 2020. Doi: 10.1109/ELNANO50318.2020.9088766.

[2] M. Korobchynskyi, A. Mariliv, A. Bohuslavets, S. Tsybulskyi, E. Sablina, V. Nechepurenko, "Algorithmic Model of the Cyclic Changes in the Temperature of the Solid Under the Effect of Convective Heat Exchanges with the Environment", Conference proceedings, 2019 15th International Conference on the Experience of Designing and Application of CAD Systems (CADSM), Lviv Polytechnic National University, pp. 4/6–4/12, February – March 2019. Doi: 10.1109/CADSM.2019.8779276.

[3] J. Jarvis, C. Wathes, "Mechanistic modeling of vertebrate spatial contrast sensitivity and acuity at low luminance", Vis Neurosci, vol. 3, pp. 169–181, 2012.

[4] M. Born, E. Volf, "Electromagnetic theory of propagation, interference and diffraction of light", Principles of Optics, vol. 6, pp. 327–329, 1980.

[5] S. Babichev, B. Durnyak, V. Zhydetskyy, I. Pikh, V. Senkivskyy, "Application of Optics Density-Based Clustering Algorithm Using", 2019 IEEE 14th International Conference on Computer Sciences and Information Technologies (CSIT), Lviv, Ukraine, September 2019. Doi: 10.1109/STC-CSIT.2019.8929869.

[6] V. Vasiliev, O. Mashkov, V. Frolov, "Method and technical meansures of the ecological monitoring", Environmental science, vol. 1 (5), pp. 57–67, 2014.

[7] M. Dmitriev, O. Panchenko, O. Derkachov, I. Rutcovska, "Definition of regional condition at the airfield surface", VNAU, vol. 1, pp. 161–164, 2008.

[8] V. Lytvyn, I. Peleshchak, V. Vysotska, R. Peleshchak, "Satelite Spectral Information Recognition Based on the Synthesis of Modified Dynamic Neural Natworks and Holographic Data Processing Techniques", 2018 IEEE 13th International Scientific and Technical Conference on Computer Sciences and Information Technologies (CSIT), 2018. Doi: 10.1109/STC-CSIT.2018.8526595.

[9] E. Mitchell, Photographic Science. The University of North Carolina at Chapel Hill: John Wiley and Sons, 1984.

[10] Standart: Photography – Color reversal camera film, Determination of ISO speed. Published Standart ISO 2240:2003, 2003.

[11] Standart: Optical Transfer Function – Definitions and Mathematical Relationships. Published Standart ISO 9334, 1995.

[12] Jin Sun, Yuzhong Ma, Han Yang, Ning Wang, Yu Ding, Aiping Song, Yongwei Zhu, Juntong Xi, "Character Recognition Method for Low-Contrast Image of Numerical Instruments", Conference proceedings, 2018 IEEE 3rd International Conference on Image, Vision and Computing (ICIVC), Institute of Electrical and Electronics Engineers, pp. 157–163, June 2018. Doi: 10.1109/ICIVC.2018.8492727.

[13] S. Puzović, R. Petrović, M. Pavlović, S. Stanković, "Enhancement Algorithms for Low-Light and Low-Contrast Images", Proceedings, 2020 19th International Symposium INFOTEH-JAHORINA (INFOTEH), University of East Sarajevo, Faculty of Electrical Engineering, pp. 1–6, March 2020. Doi: 10.1109/INFOTEH48170.2020.9066316.

[14] Canon: Digital photo Canon Powershot SX420 IS. Manuel, pp. 20–22, 2018.

[15] M. Park, J. S. Jin, Suhuai Luo, "Locating the Optic Disc in Retinal Images", Proceeding, International Conference on Computer Graphics, Imaging and Visualisation (CGIV'06), Computer Society, pp. 141–145, July 2006. Doi: 10.1109/CGIV.2006.63.

[16] Samsung: Smartphone Galaxy A50. Manuel, pp. 57–60, 2019.

[17] Nicon: Digital photo Coolpix AW100. Manuel, pp. 20–21, 2011.

[18] Panasonic: Digital photo Panasonic TZ18 Lumix. Manuel, pp. 33–35, 2015.

Topic #2

Machine Vision and Pattern Recognition

IEEE Third International Conference on Data Stream Mining & Processing
August 21-25, 2020, Lviv, Ukraine

Hybrid Multidimensional Deep Convolutional Neural Network for Multimodal Fusion

Olena Vynokurova
GeoGuard,
Kharkiv National University of Radio Electronics
Kharkiv, Ukraine
vynokurova@gmail.com

Dmytro Peleshko
GeoGuard
Lviv/Kharkiv, Ukraine
dpeleshko@gmail.com

***Abstract* — The Hybrid Multidimensional Deep Convolutional Neural Network (HMDCNN) topology for the multimodal recognition of the speech, the face, the lips, and human gestures behavior is proposed. In this case a hybridization is understood to be compatible use of 2D and 3D convolutional neural networks in one multimodal architecture. Conducted researches relate to improving the understanding of complex dynamic scenes. The basic unit of the proposed hybrid system is deep neural network topology, which combines 2D and 3D convolutional neural network (CNN) for each modality with proposed intermediate-level feature fusion subsystem. Such a feature map fusion method is based on scaling procedure with a specific combination of pooling operation with non-square kernels and allows merging different type of modalities.**

Keywords— Multimodal fusion, hybrid system, deep convolutional neural network, 2D and 3D convolution and pooling

I. INTRODUCTION

By improving the quality of video semantic segmentation based on multimodal system we try to simplify affordance detection in cases of complex visual scenes. The essence of this simplification is the correct definition of the scene subjects and the actions classification in cases where the visual representation does not allow to correctly identify the functional interaction of objects and people. For example, in the case when

- it is not clear who of several people gave the instruction command.

- one and the same object of attention in different scenes can be identified by different modalities - in one scene, voice modality, and in another - only by visual one.

- there are several attention objects of one type. And only the accuracy of the pointer's trajectory or additional information in the voice stream (for example, the dish is large, and not small) allows identifying the attention object uniquely. For successful semantic segmentation in described cases, one or two modalities may not be sufficient.

Moreover, sometimes the physical parameters of the visual scene (changing illumination, point of view, etc.) require big datasets for acceptable recognition. In such cases, the addition of another modality makes it possible to avoid the use of a big dataset.

In order to improve a semantic segmentation, HMDCNN topology is proposed that not only increases the accuracy of the segmentation of the visual scenes in the indicated cases, but also gives the opportunity to expand the set of necessary modalities specialized for cases of complex scenes.

In the developed HMDCNN topology a set of most used standard modalities is used. But if necessary, the set can be completed by modalities for a particular case of affordance detection. In this case, the precision of semantic segmentation will not increase.

The fusion procedure of modalities is one of the main parts in the HMDCNN topology that was developed. The modern artificial intelligence systems are the systems with the integrated video stream data processing where each of the streams is a selected modality. The fusion of the several modalities in single processing is lead to additional informativeness. This informativeness should not only improve the quality of processing but also move the robotics systems closer to natural communication processes.

The main purpose of this moving is to essentially simplify the interfaces of interaction with robots, which are endowed with elements of artificial intelligence. In such case, the user cannot change his habits at the level of household interaction with smart devices, and, accordingly, the processes of adapting the "coexistence" of a person and a smart device will be much easier.

From the point of view of modern methods of data streams, the processing several modalities is a parallelized initial processing of individual modalities and their fusion at the final stage. The fusion processes are already well developed, and their classification can be represented as [1]:

• Input-level features fusion is the fusion of data at the level of input functions, like the merge of unweighted features vectors.

• Intermediate features fusion (feature-level fusion) where each modality can be studied using the unimodal neural network (not necessarily Deep Neural Network, DNN). The weights, which obtained during the training of each unimodal DNN are combined together and the training process continues to the decision level.

• Decision-level fusion is so called an ensemble of fusion or late fusion when the modalities merge at the decision level.

In the case when the multimodal system contains audio modality among others it is recommended to use Feature-level fusion or Decision-level fusion [2].

II. RELATED WORKS

Processing video stream in multimodal systems in the vast majority focuses on gestures, lips, faces, bodies / poses. Most modern multimodal systems are based on existing or modifications of existing models for individual modalities

978-1-7281-3214-3/20/$31.00 ©2020 IEEE

recognition. Among the results obtained in the field of recognition of gestures, it is should be noted researches based on the using shallow networks [3] and deep or 3D networks [4-8].

In [3], an approach based on a skins model is proposed. The basic idea of identification (estimation) of motion is a background model (in YUV space) and a Kalman filter on two next frames. In fact, due to the movement assessment, resistance to the color of the skin is achieved. Symptoms of speech recognition are Mel-scale Frequency Cepstral Coefficients (MFCC). The main advantage of the proposed approach is that a large number of states can be used to classify gestures. This means that the gesture can be described in great detail. And the accuracy of recognition is very high.

However, this number of states is also a problem - for the fusion it is necessary to reconcile the number of states of various modalities. The correspondence between the states obtains based on a threshold. Defining such threshold may also be a problem.

One of the most successful models for recognizing gestures is described in [9]. The feature of the method is the use of two classifiers. The first of these is the classifier of motion, which is determined by normalized vectors in a given direction. And the second - a trajectory classifier that contains a skin detector, a normalized skeleton (skeleton) representation of one or two hands. The localization of the scene is carried out by the method of the threshold in the space YCrCb.

One of the most successful models for recognizing gestures is described in [9]. The feature of proposed method is the use of two classifiers. The first of these is the classifier of motion, which is determined by normalized vectors in a given direction. And the second one is a classifier for trajectory that contains a skin detector, a normalized skeleton representation of one or two hands. The localization of the scene is performed by the threshold method in the space YCrCb.

The main advantages of the method are:

• gesture representation by a set of subgestures.

• defining the trajectory of the gesture movement by a special point of gesture ("centroid" of the ROI area). The direction and acceleration of the gesture motion are determined using such “centroid”.

• simplicity of practical implementation and high performance and precision of method at acceptable conditions in which recognition takes place.

The main problems of the developed methodology are:

• continuous gesture duration for an interval of 2 seconds. Accordingly, the continuation of the gesture after a pause is already a new gesture.

• significant noise of the gesture pattern, repetitive subgestures. Correspondingly, there will be problems for identifying natural gestures.

• the gesture identification method is dependent on the background model. Accordingly, the recognition efficiency will be different on different backgrounds in different lighting conditions. And this problem will deepen when zooming or moving the camera.

• gesture recognition occurs in the plane, so it is impossible to spread the technique on the 3D model.

Typical approaches using Hidden Markov Model (HMM) and CNN are now being developed in various uses of LSTM or 3D networks. For example, such models are proposed in [5-7, 10-13].

III. Hybrid Multidimensional Deep Convolutional Neural Network for Multimodal Fusion

HMDCNN topology is a hybridization of 2D and 3D deep convolutional neural networks which are coupled by proposed fusion subsystem for processing multimodal data streams.

The proposed HMDCNN consists of five sequential processes of multimodal data stream processing. The first process is the generation of data tensors from audio and video modalities streams. The second process is a parallelized subsystem for generation of feature map for each modality stream. Such subsystem consists of 2D or 3D convolution and pooling operation. The proposed network topology is presented in Fig.1. In the HMDCNN, the size of the input, output, kernel, and stride of operations in Sequential models 1-4 are shown in Tab. I for video modalities and Tab. II for audio modality.

The feature map of each modality (the necessary condition is the same dimension of the feature map of each modality) is fed to the input of the process of feature matching and fusion, which will be described in detail in subsection 4.3. Tables IIIa-c show the size of the input, output, kernel, and stride of operations in the fusion subsystem for different number of modalities. Feature Map Unions is a block for modalities feature concatenation. Feature Fusion Union is a block for modalities feature concatenation after fusion operation. Fusion Polling is a fusion operation which is based on 2D pooling operation with non-square kernels.

In Tables I-III, the spatial size of the 2D or 3D convolution kernel for audio and video modalities has the following dimension $Cd \times Hd \times Wd \times Kd$, where Cd is the kernel size in the time dimension (channels), Hd and Wd are the kernel dimensions of the modality frame in height and width respectively, and Kd is the number of kernels (filters).

An important feature of the network for streaming video modality is its 3D pooling operation with the parameter pooling stride equal two in three dimensions in order to increase robustness to the moving ROI effect and to maintain ROI movement features in the neighborhood of the pooling kernel. The 3D convolutional operations are performed to find the correlation between high-level temporal and spatial information by fusion among them. No zero-padding is used in the network topology.

By geometric shape the extracted features were close both to the square (a head, a gesture, the eyes) as to the rectangle (lips, gesture, palm). Therefore, in the developed network topology in the process of formation feature map of each modality for the 2D and 3D convolution operation both square and non-square filters were used.

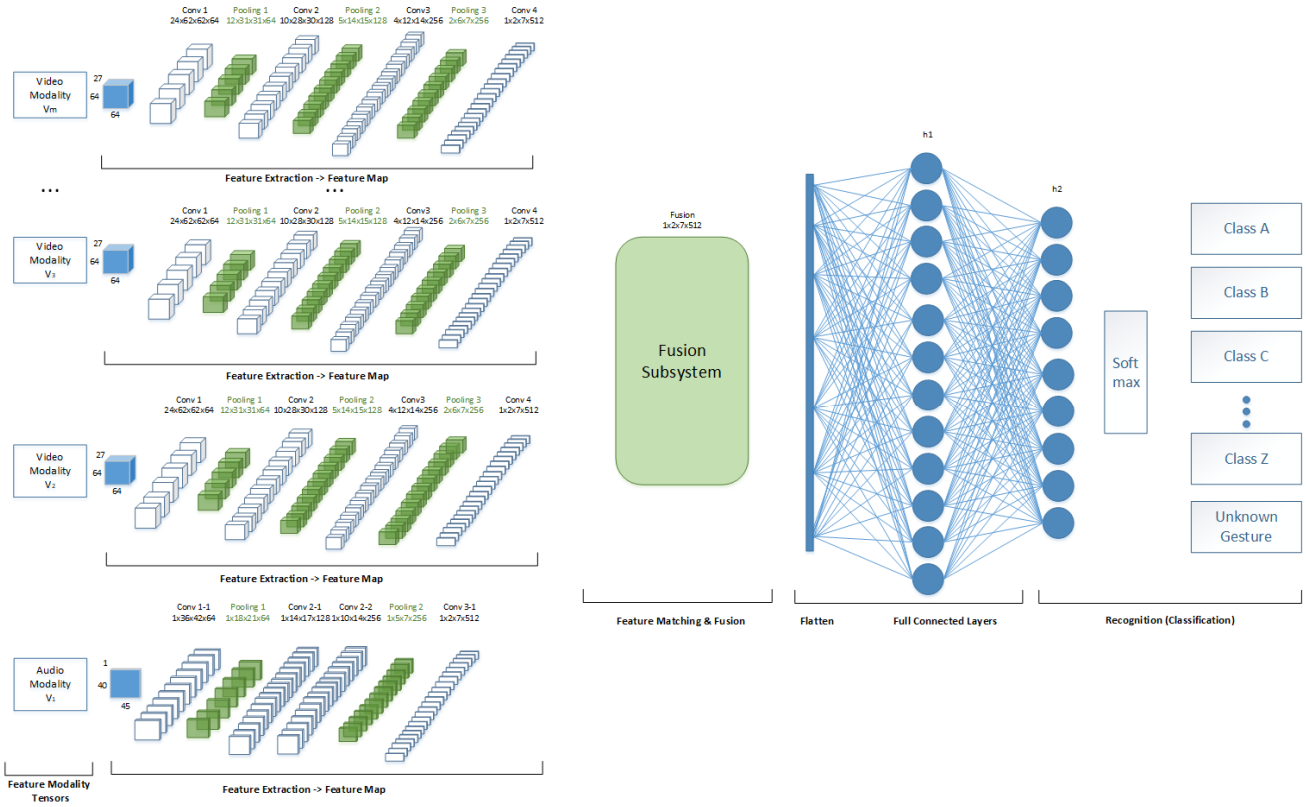


Fig. 1. Hybrid Multidimensional Deep Convolutional Neural Network Topology

TABLE I. SIZE OF OPERATIONS IN NETWORK TOPOLOGY FOR VIDEO MODALITY PROCESSING (GENERATING VIDEO MODALITIES FEATURE MAP)

Layer	Input-size	Output-size	Kernel	Stride
Conv1	27×64×64×1	24×62×62×64	4×3×3	1
Pool1	24×62×62×64	12×31×31×64	2×2×2	2×2×2
Conv2	12×31×31×64	10×28×30×128	3×4×2	1
Pool2	10×28×30×128	5×14×15×128	2×2×2	2×2×2
Conv3	5×14×15×128	4×12×14×256	2×3×2	1
Pool3	4×12×14×256	2×6×7×256	2×2×2	2×2×2
Conv4	2×6×7×256	1×2×7×512	2×5×1	1

TABLE II. SIZE OF OPERATIONS IN NETWORK TOPOLOGY FOR AUDIO MODALITIES (GENERATING AUDIO MODALITY FEATURE MAP)

Layer	Input-size	Output-size	Kernel	Stride
Conv1	1×40×45×1	1×36×42×64	1×5×4	1
Pool1	1×36×42×64	1×18×21×64	1×2×2	1×2×2
Conv2-1	1×18×21×64	1×14×17×128	1×5×5	1
Conv2-2	1×14×17×128	1×10×14×256	1×5×4	1
Pool2	1×10×14×256	1×5×7×256	1×2×2	1×2×2
Conv3	1×5×7×256	1×2×7×512	1×4×1	1

TABLE IIIA. SIZE OF OPERATIONS IN FUSION SUBSYSTEM OF NETWORK TOPOLOGY FOR PROCESSING FEATURE MAP OF MODALITIES (FOR 2 MODALITIES)

Layer	Input-size	Output-size	Kernel	Stride
G1	1×2×7×512 1×2×7×512	1×4×7×512	-	-
Θ1	1×4×7×512	1×3×7×512	1×2×1	1

TABLE IIIB. SIZE OF OPERATIONS IN FUSION SUBSYSTEM OF NETWORK TOPOLOGY FOR PROCESSING FEATURE MAP OF MODALITIES (FOR 4 MODALITIES)

Layer	Input-size	Output-size	Kernel	Stride
G1	1×2×7×512 1×2×7×512	1×4×7×512	-	-
G2	1×2×7×512 1×2×7×512	1x4x7x512	-	-
Θ1	1×4×7×512	1×2×7×512	1×3×1	1
Θ2	1×4×7×512	1×2×7×512	1×3×1	1
Ω	1×2×7×512 1×2×7×512	1×4×7×512	-	-
Θ3	1×4×7×512	1×3×7×512	1×2×1	1

Θ_j - Fusion Polling; Ω_j - Feature Fusion Union; G_j - Feature Map Union

TABLE IIIC. SIZE OF OPERATIONS IN FUSION SUBSYSTEM OF NETWORK TOPOLOGY FOR PROCESSING FEATURE MAP OF MODALITIES (FOR 4 MODALITIES)

Layer	Input-size	Output-size	Kernel	Stride
G1	1×2×7×512 1×2×7×512	1×4×7×512	-	-
G2	1×2×7×512 1×2×7×512	1x4x7x512	-	-
G3	1×2×7×512 1×2×7×512	1x4x7x512		
Θ1	1×4×7×512	1×2×7×512	1×3×1	1
Θ2	1×4×7×512	1×2×7×512	1×3×1	1
Θ3	1×4×7×512	1×2×7×512	1×3×1	1
Ω	1×2×7×512 1×2×7×512 1×2×7×512	1×6×7×512	-	-
Θ4	1×6×7×512	1×3×7×512	1×2×1	1×2×1

Θ_j - Fusion Polling; Ω_j - Feature Fusion Union; G_j - Feature Map Union

At the stage of the formation of the modalities feature map, square filters were used for the pooling operation, which completely corresponds to the methodology of images sub-sampling.

It should be noted that the fusion subsystem combines the features of the selected modalities to each other. And the sub-sampling operation of the dimension of the features in the middle of the modality does not occur. Therefore, the pooling operation is used in the fusion subsystem with non-square filters.

Since a convolutional neural network with non-square kernels provides a transition from low-level attributes to higher-level attributes, this fact is lead to extraction and processing temporal features at the lower level that are connected with speech functions.

Thus, at the output of the process of generating feature map for each modality stream, we have a unique set of features of the same dimension (in our case, the shape = 1×2×7×512).

Batch normalization and Dropout operation were used for all layers. Except for the last layer, all layers used the activation function Parametric ReLU (PReLU) [14]

$$\psi(o^{[i]}) = \max(0, o^{[i]}) + a_i \min(0, o^{[i]})$$

where $o^{[i]}$ is the network layer output, a_i is the parameter of the steepness of the negative part of the function.

The activation function of PReLU is a synthesis of the ReLU function and in comparison with the activation function of ReLU, PReLU has demonstrated better data stream processing.

After that, the results are fed to the last process, which corresponds to the recognition (classification) of gestures.

Optimizing tuned parameters of the 3DDCNN is based on cross entropy criterion optimization:

$$E(k) = -\sum_{j=1}^{Z} y_j(k) \log(\hat{y}_j(k))$$

where $y_j(k)$ is label of class j for k observation, $\hat{y}_j(k)$ is predicted label of class j for k observation.

To tune the parameters, we will use the root mean square propagation (rmsProp) [15] learning algorithm, which can be written in the form

$$W(k+1) = W(k) - \frac{\eta}{\sqrt{S[g^2](k) + \varepsilon}} g(k),$$

where $W(k)$ is the tuned parameters of network, k is instant of discrete time or number of observation, η - learning rate ($0 < \eta < 1$), $S[g^2](k)$ is the moving average at k discrete time

$$S[g^2](k) = \gamma S[g^2](k-1) + (1-\gamma) g^2(k).$$

Here $g(k) = \nabla_W E(k)$ is the gradient of the optimization criterion by the tuned parameters W, $0 < \gamma < 1$.

IV. Experiments and Results

To confirm effectiveness and compare the proposed approach with known approaches, it was chosen the Oxford-BBC Lip Reading in the Wild (LRW) Dataset described in [16].

For comparison of the proposed approach with known ones, the HMDCNN topology was adapted for two modalities – the audio (extracted MFCC) and video (features of lips motion) modalities, where the proposed fusion subsystem is used for the coupling modalities.

The dataset consists of up to 1000 occurrences per target word for training, 50 occurrences per word for testing and 50 occurrences per word for validating of 500 different words, spoken by hundreds of different speakers. All videos are 29 frames in length, and the word occurs in the middle of the video. The word duration is given in the metadata, from which you can determine the start and end frames.

Thus, the features tensors of the modalities were generated based on the dataset. For video modality, the tensor dimension was 27 × 64 × 64 (27 channel visual feature tensor with size 64 × 64 pixels) and, accordingly, an audio modality has 1 × 40 × 45 dimension, which corresponds to 40 MFCCs, which were calculated on sections 20 msec without overlap on a full audio signal with a length of 900 msec. Then, a matrix of eigenvectors with a dimension of 1 × 40 × 45 was generated based on matrix with MFEC coefficients.

Fig.4 and Tables I, II, IIIa shows layer's dimensions of Hybrid Multidimensional Deep Convolutional Neural Network for solving recognition task based on Oxford-BBC Lip Reading in the Wild (LRW) Dataset.

The rmsProp training algorithm had the following parameters: learning step 0.0001 with decay parameter 10^{-6}. The number of epochs was about 15. Dropout parameters in the sequential models were taken 0.1.

In the paper [17] authors have presented main results for processing Oxford-BBC Lip Reading in the Wild (LRW) dataset based on different approaches. These approaches are 7 type of deep networks:

- Baseline - the multi-tower VGG-M,
- N1 - the 2D convolution ResNet, while the back-end is based on temporal convolution.
- N2 - the same model as N1, but with 3D convolution.
- N3 - Deep Neural Network (DNN) of approximately the same number of parameters (~20M).
- N4 - the same model as N1 based on a single-layer Bi-LSTM instead of temporal convolution.
- N5 - the same model as N1 based on a double-layer Bi-LSTM instead of temporal convolution.

- N6 - the same model as N5, but trained end-to-end, using the weights of N5 as starting point.
- N7 - is also trained end-to-end and the sole difference with N6 is that the outputs of the two directional LSTMs are concatenated together instead of added together.

All models trained 15 to 20 epochs. All networks have approximately the same number of parameters (~20M). The experimental results for proposed HMDCNN and approaches above are shown in Table IV.

TABLE IV. THE RESULTS OF PROCESSING OXFORD-BBC LIP READING IN THE WILD (LRW) DATASET BASED ON DIFFERENT APPROACHES

Network	Accuracy
Baseline (VGG-M)	61.1%
N1 CNN with 2D conv (Res Net)	69.6%
N2 CNN with 3D conv (Res Net)	74.6%
N3 DNN 3D conv	69.7%
N4 Network based on LSTM	78.4%
N5 Network based on Bi-LSTM	79.6%
N6 The same as N5, but trained end-to-end, using the weights of N5 as starting point	81.5%
N7 It is also trained end-to-end and the sole difference with N6 is that the outputs of the two directional LSTMs are concatenated together instead of added together	83.0%
Proposed HMDCNN	**86%**

The accuracy of the proposed model is 86.0%, which corresponds to the improvement of the result relative to the baseline VGG-M network by 25.1% and an increase in accuracy relative to the state-of-the art by 3%. It should be noted that the provided network contains almost 2 times less configurable parameters.

V. CONCLUSIONS

As was already mentioned the proposed HMDCNN is intended to improve the understanding of the robot systems of various complex dynamic scenes. That is, in cases where the identification of action or subject is not unambiguous.

The developed HMDCNN topology and fusion method give an opportunity to expand the set of new modalities without restrictions (the limitations are only computing resources of the operating environment). In this case, the processing of these modalities can be both in 3D and in 2D dimension. The level of data preprocessing can be different: from convolution operation sequences to deep network processing.

It should be noticed that HMDCNN topology allows using the different patterns of fusion. In the process of developing HMDCNN, several variants of the fusion operation have been developed. The presented fusion procedure in the paper is chosen according to the criterion of the highest accuracy of the classification problem based on the selected datasets. The question of the choice of fusion procedure demands the individual research and in this paper has not been considered.

ACKNOWLEDGMENT

This research was supported by Samsung R&D Institute Ukraine, which kindly provided computing power for practical experiments.

We are very grateful to Rob Cooper at BBC Research for help in obtaining the LRW.

REFERENCES

[1] N.-B Chang, and K. Bai. (2018). "Multisensor data fusion and machine learning for environmental remote sensing". CRC PressTaylor & Francis Group. doi: 10.1201/b20703. Available at https://www.researchgate.net/publication/325347741_Multisensor_data_fusion_and_machine_learning_for_environmental_remote_sensing

[2] P. Wu, H. Liu, X. Li, T. Fan, and X. Zhang. "A Novel Lip Descriptor for Audio-Visual Keyword Spotting Based on Adaptive Decision Fusion." IEEE Transactions on Multimedia, vol. 18., no. 3, pp.326-338, March 2016.

[3] S.-T. Cheng, C.-W. Hsu, and J.-P. Li. "Combined Hand gesture-speech model for human action recognition." Sensors, 13: 17098-17129, 2013.

[4] P. Tzirakis, G. Trigeorgis, M. A. Nicolaou, B. Schuller, and S. Zafeiriou. "End-to-end multimodal emotion recognition using deep neural networks". IEEE Journal of Selected Topics in Signal Processing, 11(8):1301-1309, 2017.

[5] J. Yue-Hei Ng, M. Hausknecht, S. Vijayanarasimhan, et al. "Beyond short snippets: Deep networks for video classification." Proc. of the IEEE Conference on Computer Vision and Pattern Recognition, 4694-4702, 2015.

[6] K. Simonyan, A. Zisserman. "Two-stream convolutional networks for action recognition in videos." Advances in Neural Information Processing Systems, 1:568-576, 2014.

[7] C. Feichtenhofer, A. Pinz, A. Zisserman. "Convolutional two-stream network fusion for video action recognition." Proc. of the IEEE Conference on Computer Vision and Pattern Recognition, 1933-1941, 2016.

[8] Z. Hu, H. Youmin, J. Liu, B. Wu, D. Han, T. Kurfess. "3D Separable Convolutional Neural Network for Dynamic Hand Gesture Recognition." Neurocomputing, 318: 151-161, 2018.

[9] M. Favorskaya, A. Nosov, A. Popov. "Localization and recognition of dynamic hand gesture based on hierarchy of manifold classifiers." The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, XL-5/W6, 2015. (open access - https://www.int-arch-photogramm-remote-sens-spatial-inf-sci.net/XL-5-W6/1/2015/isprsarchives-XL-5-W6-1-2015.pdf)

[10] E. Tsironi, P. Barros, C. Weber, S. Wermter. "An analysis of convolutional long short-term memory recurrent neural networks for gesture recognition." Neurocomputing, 268: 76–86, 2017.

[11] P. Wu, H. Liu, X. Li, T. Fan, and X. Zhang. "A Novel Lip Descriptor for Audio-Visual Keyword Spotting Based on Adaptive Decision Fusion." IEEE Transactions on Multimedia, vol. 18., no. 3, pp.326-338, March 2016.

[12] R. Ding, C. Pang and H. Liu. "Audio-visual keyword spotting based on multidimensional convolution neural network." In Proc. 2018 IEEE International Conference on Image Processing, Athens, Greece, October 2018, pp.4138-4142

[13] Argones Rúa, E., Bredin, H., García Mateo, C. et al. "Audio-visual speech asynchrony detection using co-inertia analysis and coupled hidden markov models." Pattern Anal Applic 12, 271–284 (2009). https://doi.org/10.1007/s10044-008-0121-2

[14] Kaiming He et al. "Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification". arXiv:1502.01852

[15] rmsProp. http://ruder.io/optimizing-gradient-descent/index.html#rmsprop

[16] J. S. Chung, A. Zisserman, "Lip Reading in the Wild. Asian Conference on Computer Vision", 2016 [http://www.robots.ox.ac.uk/~vgg/data/lip_reading/lrw1.html]

[17] T. Stafylakis, G. Tzimiropoulos, "Combining Residual Networks with LSTMs for Lipreading". INTERSPEECH 2017: 3652-3656

IEEE Third International Conference on Data Stream Mining & Processing
August 21-25, 2020, Lviv, Ukraine

Interpretability of Neural Network Binary Classification with Part Analysis

Oleksii Gorokhovatskyi
Informatics and Computer Technologies
Simon Kuznets Kharkiv National University of Economics
Kharkiv, Ukraine
ORCID: 0000-0003-3477-2132

Olena Peredrii
Informatics and Computer Technologies
Simon Kuznets Kharkiv National University of Economics
Kharkiv, Ukraine
ORCID: 0000-0003-0390-1931

Volodymyr Gorokhovatskyi
Informatics
Kharkiv National University of Radio Electronics
Kharkiv, Ukraine
ORCID: 0000-0002-7839-6223

***Abstract*—In this paper, we describe searching for an explanation of some particular decision of shallow convolutional neural network classification result. The method is the division of the initial input image, followed by the hiding and classification of separate parts. As a result of local interpretation search, we get a pair of images, each of which represents hidden parts of the initial image but classified differently. We tested the proposed approach for a shallow convolutional neural network model of different training quality using the Food-5K dataset. The average quantity of images that are used for explanations is estimated. We compared the analysis of the division of the images with known hiding of the superpixels method and showed that the success explanation rate is better for the proposed method.**

Keywords—superpixel; classification explanation; black box model; interpretation; shallow convolutional neural network; division of image, part analysis.

I. INTRODUCTION

Most state-of-the-art approaches in computer vision problems related to pattern recognition use artificial neural networks (ANN). They allow to get very good results but not able to provide any explanations to the interested person about reasons, which influence this decision. This drawback sometimes raises the question of whether we can trust such automated artificial neural network decisions [1, 2]. This problem becomes more acute in case we are talking about sensitive environments making a decision for like person identification, medical diagnostic and health care, safety, and security, etc.

Last years the usage of AI systems became the subject of regulations, e.g. General Data Protection Regulation (GDPR) [3, 4], introduced in European Union in 2016 and implemented in 2018, guarantees the right to have a local explanation about the logic involved to automatic decision-making systems. Authors of Montreal Declaration of responsible AI development proposed such interpretability statement [5]: "The decisions made by AIS affecting a person's life, quality of life, or reputation should always be justifiable in a language that is understood by the people who use them or who are subjected to the consequences of their use. Justification consists in making transparent the most important factors and parameters shaping the decision, and should take the same form as the justification we would demand of a human making the same kind of decision". Finally, policy guidelines on Artificial Intelligence, which contain the "Transparency and explainability" chapter, were adopted by OECD and partner countries in May 2019 [6].

Different methods to get explanations [7] are known. Backpropagation (measures the influence of each pixel on the common result exploring backpropagation of a signal) and perturbation (creates some modification of initial signal and measures how classification result changes) are the most common ones.

LIME (Local Interpretable Model-Agnostic Explanations) [8, 9] and SHAP (SHapley Additive exPlanations) [10 – 12] are amongst the most famous methods, used to search for interpretations. Authors base LIME on the construction of a separate explainable model that uses the perturbation input features vector as a training set. The issues of this approach relate to the need to select some limited quantity of initial features and the overall instability [10].

The core of SHAP is the feature analysis based on game theory, which allows calculating the contribution (Shapley value) of every feature in the global solution. The major drawback of this approach is a significant computational complexity since it is necessary to review all possible combinations of all features.

Causal interpretations of the black-box model are proposed in [13]. By causality, authors mean the investigation of change of output value after the changing of single input parameter while other parameters remain the same.

Another way to understand how the model works is the search for counterfactuals [14 – 16]. According to [4], counterfactual is a statement that describes what initial values should be to achieve desirable output. This statement should be as close as possible to the decision in a particular case, so minimal changes should be applied to get the desired result from the current.

Different modifications of neural network model to implement interpretability functionality are proposed. Search for critical data routing paths that inspect the intermediate layer outputs and build numerical "gates" those each input sample should pass to be classified correctly was introduced in [17]. Such analysis is stable to adversarial attacks. Construction of the special interpretable convolutional neural network (CNN) with automatic detection of different filters in high convolutional layer is proposed in [18]. An Attention Branch network that can improve performance, explanation, and accuracy compared to traditional CNNs is proposed in [19]. Paper [20] describes the building of a decision tree, which is trained on the convolutional features of high layers.

978-1-7281-3214-3/20/$31.00 ©2020 IEEE

The contribution of the paper is the analysis of parts of images to build local perturbations of the initial image to find explanations of some particular classification result and investigations of its properties.

II. Superpixels

One of the first ideas for superpixels was introduced in [21] to replace the processing of a huge quantity of pixels with the processing of a reduced quantity of superpixels – regions, which preserve properties of images. Superpixel analysis was applied in LIME [8, 9] to get explanations about the influence of separate parts of the image on the classification. However, the effective selection of separate superpixels and combinations of them are unclear, except for the brute force method. Papers [22 – 24] refer to iterative replacing of some image fragments with the mean value preserving, but this method is computationally expensive. Since that time a lot of different superpixel algorithms are developed [25, 26]: graph based segmentation, quickshift, compact watershed, normalized cuts, SLIC, and others [27].

One of the most popular methods to build superpixels is SLICO [28, 29], that requires only the quantity of superpixels to be set up to segment the image. SLICO uses K-means clustering to merge similar pixels into subpixels. Limitations of SLIC/SLICO relate to superpixel's borders, which are somewhat ragged, don't adhere to content boundaries, and quantity of superpixels formed, which may not follow exactly specified required value [30, 31]. Additionally, SLIC may provide too uniform or too inhomogeneous results in case of an inappropriate quantity of superpixels chosen. For example, as you can see in Fig. 1, the first image contains 22 superpixels, but was generated for 20 ones, the second image contains 8 superpixels, but quantity parameter was set up to be 10. Even such minor differences may reduce the performance for some methods of superpixels combining (e.g. exhaustive search).

We tested SLICO implementations from [29] and [32] in this paper.

Fig. 1. Examples of SLICO [29] segmentation results

III. Part Analysis of Image Classification Process

We will consider another way to get and hide various parts of images without involving content/semantic analysis. We assume that we already have some neural network classifier, trained beforehand, and we use it to recognize whether this image contains food.

Let's look at an example of the work of the proposed method, the initial image is shown in Fig. 2. It has been classified as "food" quite confidently, corresponding classifier outputs are 0.00001103 for "non-food" and 0.99998581 for "food" classes, respectively.

Fig. 2. The initial image

Our first step is to split the image into 4 parts and generate 4 new images (Fig. 3) with each part hidden one by one in turn. It is possible to hide them with some specific or average color, or blur. It worth noting that this color should not be sensitive for content, e.g. the usage of color typical for the object being classified is not the appropriate choice.

Fig. 3. First division step and the selection of sample image

We classify each of the new four images independently. All samples in this case are classified as "food" with following output values: 0.99991733, 0.99952668, 0.99988151, 0.9998247. We have to select one of them to work with further, so choose the least confident decision, sample 2.

Choosing the image at the previous stage as a basic one, we again split it into smaller parts (Fig. 4), generate corresponding samples with hidden parts, and classify them. Again, all images were classified as "food" and we selected the sample with the weakest decision.

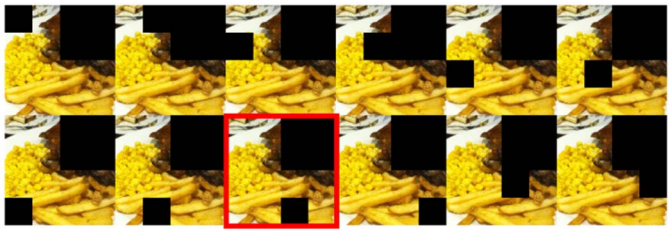


Fig. 4. Second division step and the selection of sample image

After the split at the next stage, we finally get one sample classified as "non-food" (Fig. 5, highlighted with a red border). Now we find a complement of this image with swapping back regions to original ones and vice versa and verify the result of its classification. Both images, the first one is of "non-food" class and the second one represents "food" one is shown in Fig. 6. This is the successful finish of the division algorithm.

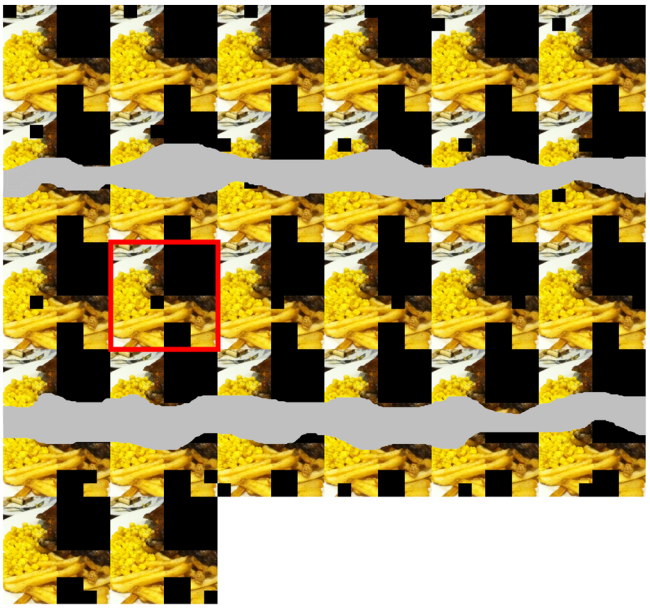


Fig. 5. Third division step and the selection of sample image

If both final images are of the same class, the algorithm fails. Here, we may step back and repeat the last division stage for another sample. If there are few sample images at any stage, which are classified as a required class, we may test all of them and receive few results.

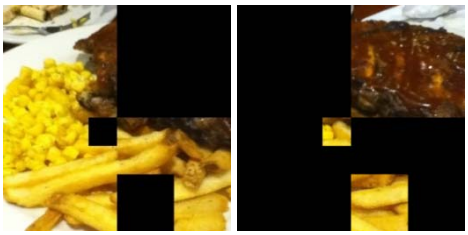


Fig. 6. Pair of complement images as the result of division algorithm

IV. Modeling

Investigations of shallow ANN architecture usage in solving different tasks [33 – 39] and other methods of ANN optimization like pruning [40, 41] show that the accuracy of smaller nets may be close enough to deeper and state-of-the-art ANNs. Additionally, it is more convenient and effective to use tiny architectures to build cascades and ensembles of neural networks to make a cooperative decision. We assume that the architecture of ANN/CNN is shallow if it consists of only up to ten layers of any type.

We created a shallow network that consists of two convolutional layers, two maxpooling and dropout ones, one hidden dense layer with 512 neurons, and the last output layer with 2 neurons for two corresponding food/non-food classes (Fig. 7). The size of the initial image was 64x64 pixels, optimizing is done by Adam, batch size during training was set up to 128 images. Keras [42] software was used to work with neural network models.

Fig. 7. Shallow CNN architecture used for food/non-food classification

We used Food-5K dataset [43] for training and testing. It contains 2500 food and 2500 non-food images. 3000 of them were used as training ones, 1000 for testing, and 1000 for validation stages. It is mentioned in [44], that fine-tuned GoogLeNet (22 layered CNN) allows achieving 99.2% of accuracy.

We wanted to investigate additionally how the quality of a trained neural network influences explanation procedure, so we fix the state of the network after 5, 10, 15, 20, 25, and 30 training epochs. Corresponding classification accuracies, achieved on a test set, are: 91.6%, 95.25%, 97.9%, 98.3%, 98.95%, 99.6%.

We will use “explanation success rate” (ESR) metrics to compare the quality of explanations search for manual hiding of superpixels and division: if the method allows finding such an image, that helps to explain classification result, we will consider this experiment as successful (even if initial classification result was incorrect). We will test all images from the test set and accumulate an explanation success rate (ESR).

We limit SLICO method with 10 superpixels and combinations of only 1, 2, and 3 hidden superpixels because of poor performance. We apply early stopping if the explanation is found for a smaller quantity of hidden superpixels. The proposed division approach was limited with a minimal size of a part, that makes sense (both height and width should be greater than 32), and early stopping is applied if the explanation is found for parts of a bigger size.

We confirm explanation by both methods to be successful only in case if hiding of meaningful parts of the image (superpixels or parts) leads to the change of classification results and, at the same time, classification of only these parts (hiding all another) preserves initial classification results.

Result for all inspected networks and methods are combined in Fig. 8. The size of a point (as well as the label inside) shows the number of superpixels, that was used for this network. The color of a point relates to the accuracy of the specific network achieved on the testing set. OX axis shows the success rate for explanations found for “non-food” class, OY – corresponding value for “food” class. This does not apply to the division approach, so these all points are of a unit size with corresponding network accuracy label.

The network that classifies images correctly for 91.6% cases is not effective for searching for explanations both with superpixel and our approaches (Fig. 8). There is a huge success rate imbalance between two classes: explanations for “food” class were found in over 90% cases, whilst the same values for “non-food” class are below 40%. Approximately the same is true for the network after the next training stage (95.25% of test set accuracy), but the division of an image works better in this case. All other trained networks are similar in terms of success rate, but results achieved for 8 and 10 superpixels seem to be more balanced between two separate food/non-food classes.

Finally, the division algorithm shows much better results for searching for explanations for “non-food” cases compared to all previous superpixel-based approaches and just a bit worse ESR of “food” cases.

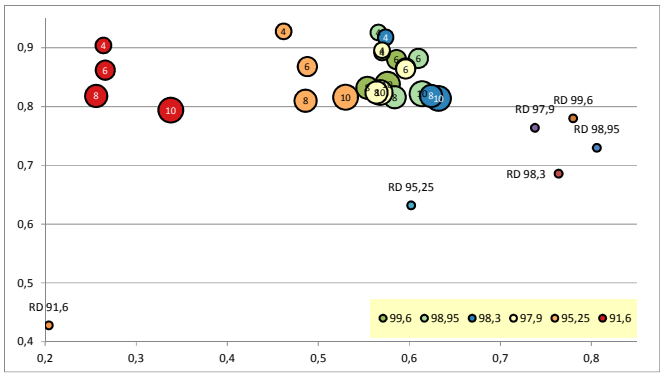


Fig. 8. Explanation success rate for different training stages of shallow CNN and division approach

Let's compare the quantity of images required to classify for superpixels and division approaches. Assuming we have n superpixels and want to create and process all combinations of k of them (e.g. hiding all pairs of superpixels means $k=2$, hiding all triples means $k=3$ and so on), the total images with hidden parts to be investigated is $\frac{n!}{k!(n-k)!}$. So, the total number of images, each of which has 4 hidden superpixels out of 10 is 210.

In case we look at division, the quantity of images on the first step is $q_1 = 2^2$, on the second – $q_2 = 2^4 - 2^2$, for the third – $q_3 = 2^6 - 2^4 - 2^2$, forming in such a way common expression: $q_n = 2^{2n} - \sum_{k=1}^{n-1} 2^{2k}$. This function grows much faster (Fig. 9), than corresponding one for superpixels, e.g. for $n=4$ we will have 172 images, but their size will be 256 times less, than size of initial image. For most of images this threshold is enough to stop because of small size of parts. So, division is suitable when we need to inspect influence of pretty small parts of image on the overall classification result. That is not convenient to implement with superpixels because we would need a lot of them.

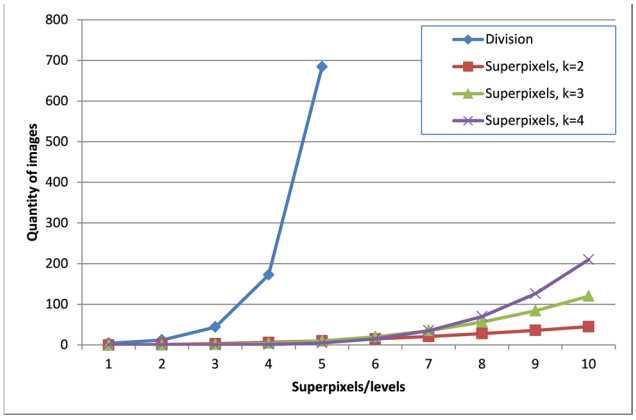


Fig. 9. Quantity of images required for testing of superpixel combinations and partial analysis

The statistics of the most accurate network (99.6% of correct classification rate) is shown in Table I. 707 explanation attempts were successful using hiding of separate superpixel approach compared to 780 ones for split approach. 247 cases out of 707 were found hiding just one superpixel out of 10, 294 – hiding two superpixels, 166 – hiding three superpixels. 162 successful cases out of 780 for the division were achieved splitting image into 2 parts, 293 – into 4 parts, 282 into 8 parts and 43 – splitting into 16 parts.

TABLE I. SUPERPIXELS VS DIVISION COMPARISON

Switching superpixels off				
Quantity of superpixels	Quantity of successfully explained images			Average images per explanation
	1 hidden	2 hidden	3 hidden	
10	247	294	166	50.38
8	224	280	189	27.87
6	274	300	159	10.98
4	292	277	162	4.76
Division and switching parts off				
Level1 q_1	Level2 q_2	Level3 q_3	Level4 q_4	Average images per explanation
162	293	282	43	30.72

The average quantity of images that requires classification to get an explanation is shown in the last column of Table I. The performance of the division is nearly comparable to searching for explanations for 8 superpixels. But division in this case found 780 explanations, while hiding superpixels one by one only 693.

Sometimes geometrical division of the image protects from the situation, when the object of interest in image falls into the entire superpixel. One of such example is gender recognition by face, examples of classification explanations are shown in Fig. 10. Parts of an image responsible for the correct class ("male" in the first case, "female" in the second) are highlighted. Second image is processed with division into three parts at each stage. The usage of superpixel-based method is not convenient here because it splits the part with a face in the nonregular fashion.

Fig. 10. Example of gender classifications explanations

V. CONCLUSIONS

The investigation describes the idea of the usage of the division of an image into non-intersection pieces, hiding them, in turn, one by one with black color to generate perturbation sample images. These images are used to search for an explanation of a particular decision, made by the black box model in a form of artificial convolutional shallow neural network.

One advantage of the part analysis is the success explanation rate, which is comparable to the method that is based on searching for superpixels and hiding them in turn. The proposed approach may be useful in cases when analysis of the context of the image is slow or makes no sense (e.g. with an analysis of regular images like faces). Future research of the division approach may relate to split not only

in quantities which are degrees of two, but three, or even different values at different stages.

Additionally, we showed that the explanation success rate depends on the quality of the network and may be quite low even for a network of good accuracy (e.g. 91.6% of correct classification rate in our case).

REFERENCES

[1] D. Gunning, "Explainable Artificial Intelligence (XAI)," [Online]. Available: https://www.cc.gatech.edu/~alanwags/DLAI2016/(Gunning)%20IJCAI-16%20DLAI%20WS.pdf. [Accessed: February 15, 2020].

[2] D. Gunning, "Explainable Artificial Intelligence (XAI)," [Online]. Available: https://www.darpa.mil/attachments/XAIProgram Update.pdf. [Accessed: February 15, 2020].

[3] General Data Protection Regulation GDPR. [Online]. Available: https://gdpr-info.eu. [Accessed: February 15, 2020].

[4] S. Wachter, B. Mittelstadt, and C. Russell, "Counterfactual Explanations Without Opening the Black Box: Automated Decisions and the GDPR," Harvard journal of law & technology, vol. 31, pp. 841-887, 2018.

[5] The Montreal Declaration for a responsible development of artificial intelligence. [Online]. Available: https://www.montrealdeclaration-responsibleai.com. [Accessed: February 15, 2020].

[6] Forty-two countries adopt new OECD Principles on Artificial Intelligence. [Online]. Available: https://www.oecd.org/science/forty-two-countries-adopt-new-oecd-principles-on-artificial-intelligence.htm. [Accessed: February 15, 2020].

[7] J. Wagner et al. "Interpretable and Fine-Grained Visual Explanations for Convolutional Neural Networks," in Proceedings of the IEEE International Conference on Computer Vision and Pattern recognition (CVPR), pp. 9097-9107, 2019.

[8] M. T. Ribeiro, S. Singh, and C. Guestrin, "Why Should I Trust You? Explaining the Predictions of Any Classifier," [Online]. Available: https://arxiv.org/pdf/1602.04938.pdf. [Accessed: September 28, 2019].

[9] M. T. Ribeiro, "LIME - Local Interpretable Model-Agnostic Explanations," [Online]. Available: https://homes.cs.washington.edu/~marcotcr/blog/lime/. [Accessed: September 28, 2019].

[10] C. Molnar, "Interpretable Machine Learning. A Guide for Making Black Box Models Explainable," [Online]. Available: https://christophm.github.io/interpretable-ml-book. [Accessed: September 28, 2019].

[11] SHAP (SHapley Additive exPlanations), [Online]. Available: https://github.com/slundberg/shap. [Accessed: January 12, 2020].

[12] S. M. Lundberg, and S. Lee, "A Unified Approach to Interpreting Model Predictions," In: I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, R. Garnett. (eds.) Advances in Neural Information Processing Systems, pp. 4765–4774, 2017.

[13] Q. Zhao, and T. Hastie, "Causal interpretations of black-box models," Journal of Business & Economic Statistics, pp. 1–10, 2019, doi: 10.1080/07350015.2019.1624293.

[14] C. Chang, E. Creager, A. Goldenberg, and D. Duvenaud, "Explaining image classifiers by counterfactual generation," [Online]. Available: https://openreview.net/pdf?id=B1MXz20cYQ. [Accessed: January 12, 2020].

[15] A. Kanehira, K. Takemoto, S. Inayoshi, and T.Harada, "Multimodal Explanations by Predicting Counterfactuality in Videos," in Proceedings of the IEEE International Conference on Computer Vision and Pattern recognition (CVPR), pp. 8594-8602, 2019.

[16] O. Gorokhovatskyi, O. Peredrii, V. Zatkhei, and O. Teslenko, "Investigation of Random Neighborhood Features for Interpretation of MLP Classification Results," In: Lytvynenko V., Babichev S., Wójcik W., Vynokurova O., Vyshemyrskaya S., Radetskaya S. (eds). Advances in Intelligent Systems and Computing, pp. 581–596, 2020, doi: 10.1007/978-3-030-26474-1_40

[17] Y. Wang, H. Su, B. Zhang, and X. Hu, "Interpret Neural Networks by Identifying Critical Data Routing Paths," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 8906-8915, 2018, doi: 10.1109/CVPR.2018.00928.

[18] Q. Zhang, Y. N. Wu, and S. Zhu, "Interpretable Convolutional Neural Networks," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 8827-8827, 2018, doi: 10.1109/CVPR.2018.00920.

[19] H. Fukui, T. Hirakawa, T. Yamashita, and H. Fujiyoshi, "Attention Branch Network:Learning of Attention Mechanism for Visual Explanation," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 10705-10714, 2019, doi: 10.1109/CVPR.2019.01096.

[20] Q. Zhang, Y. Yang, H. Ma, and Y. N. Wu, "Interpreting CNNs via Decision Trees," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 6254-6263, 2019, doi: 10.1109/CVPR.2019.00642.

[21] X. Ren, and J. Malik, "Learning a classification model for segmentation," in Proceedings of the Ninth IEEE International Conference on Computer Vision, vol. 1, pp. 10-17, 2003, doi: 10.1109/iccv.2003.1238308.

[22] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Object Detectors Emerge in Deep Scene CNNs," [Online]. Available: https://arxiv.org/pdf/1412.6856.pdf. [Accessed: February 02, 2020].

[23] P. Dabkowski, Y. Gal, "Real time image saliency for black box classifiers," in Proceeding of the 31st International Conference on Neural Information Processing Systems (NIPS), pp. 6970-6979, 2017.

[24] D. Seo, K. Oh, I. Oh, "Regional Multi-scale Approach for Visually Pleasing Explanations of Deep Neural Networks," IEEE Access, vol.8, pp. 8572–8582, 2019, doi: 10.1109/access.2019.2963055.

[25] D. Stutz, A. Hermans, and B. Leibe, "Superpixels: An Evaluation of the State-of-the-Art," Computer Vision and Image Understanding, vol. 166, pp. 1-27, 2018, doi: 10.1016/j.cviu.2017.03.007.

[26] Comparison of segmentation and superpixel algorithms. [Online]. Available: https://scikit-image.org/docs/dev/auto_examples/segmentation/plot_segmentations.html. [Accessed: September 28, 2019].

[27] D. Stutz, "An Up-to-Date List of Superpixel Algorithms," [Online]. Available: https://davidstutz.de/date-list-superpixel-algorithms. [Accessed: September 28, 2019].

[28] R. Achanta, A. Shaji, K. Smith, et al., "SLIC Superpixels Compared to State-of-the-art Superpixel Methods," in: Proceedings of IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 34 (11), pp. 2274 – 2282, 2012, doi: 10.1109/tpami.2012.120.

[29] Image and Visual Representation Lab IVRL, "SLIC Superpixels," [Online]. Available: https://ivrl.epfl.ch/research-2/research-current/research-superpixels. [Accessed: September 28, 2019].

[30] X. Xie, G. Xie, X. Xu, L. Cui, and J. Ren, "Automatic Image Segmentation With Superpixels and Image-Level Labels," IEEE Access, vol. 7, pp. 10999–11009, 2019, doi: 10.1109/access.2019.289194.

[31] J. Zhao, R. Bo, Q. Hou. M. Cheng, and P. Rosin, "FLIC: Fast Linear Iterative Clustering with Active Search," Computational Visual Media, vol.4(4), pp. 333-348, 2018, doi: 10.1007/s41095-018-0123-y.

[32] SLICO#: SLICO Superpixel Segmentation in C# (.NET) [Online]. Available: https://github.com/junjiez/SLICOSharp. [Accessed: February 15, 2020].

[33] O. Gorokhovatskyi, "Shallow Convolutional Neural Networks for Pattern Recognition Problems," in Proceedings of the Second IEEE International Conference on Data Stream Mining & Processing, August 2018, Lviv, Ukraine, pp. 459 – 463, 2018, doi: 10.1109/dsmp.2018.8478540.

[34] M. D. McDonnell, and T. Vladusich, "Enhanced Image Classification With a Fast-Learning Shallow Convolutional Neural Network," in 2015 International Joint Conference on Neural Networks (IJCNN), 2015, doi: 10.1109/ijcnn.2015.7280796.

[35] J. Pan, E. Sayrol, X. Giro-i-Nieto, K. McGuinness, and N. E. O'Connor, "Shallow and Deep Convolutional Networks for Saliency Prediction," in Proceedings of the IEEE International Conference on Computer Vision and Pattern recognition (CVPR), pp. 598-607, 2016, doi: 10.1109/CVPR.2016.71.

[36] J. Ba, and R. Caruana, "Do Deep Nets Really Need to be Deep?" Advances in Neural Information Processing Systems (NIPS), pp. 2654-2662, 2014.

[37] S. Yu et al. "A shallow convolutional neural network for blind image sharpness assessment," PLoS ONE, vol. 12(5), 2016, doi: 10.1371/journal.pone.0176632

[38] H. T. Le, C. Cerisara, and A. Denis, "Do Convolutional Networks need to be Deep for Text Classification?" [Online]. Available:

https://arxiv.org/pdf/1707.04108.pdf. [Accessed: September 28, 2019].

[39] C. Szegedy et al. “Going deeper with convolutions,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), p. 1-9, 2015, doi: 10.1109/cvpr.2015.7298594

[40] J. Frankle, G. K. Dziugaite, D. M. Roy, and M. Carbin, “Stabilizing the Lottery Ticket Hypothesis,” [Online]. Available: https://arxiv.org/pdf/1903.01611.pdf. [Accessed: September 28, 2019].

[41] J. Frankle, and M. Carbin, “The Lottery Ticket Hypothesis: Finding Sparse, Trainable Neural Networks,” [Online]. Available: https://openreview.net/pdf?id=rJl-b3RcF7. [Accessed: September 28, 2019].

[42] F. Chollet, “Keras,” [Online]. Available: https://github.com/fchollet/keras. [Accessed: February 15, 2020].

[43] Food Image Dataset [Online]. Available: https://mmspg.epfl.ch/downloads/food-image-datasets. [Accessed: February 15, 2020].

[44] A. Singla, L. Yuan, and T. Ebrahimi, “Food/Non-food Image Classification and Food Categorization using Pre-Trained GoogLeNet Model,” [Online]. Available: https://infoscience.epfl.ch/record/221610/files/madima2016_food_recognition.pdf. [Accessed: February 15, 2020].

IEEE Third International Conference on Data Stream Mining & Processing
August 21-25, 2020, Lviv, Ukraine

Fabric Defects Detection by Comparison of Clustered Samples

Roman Melnyk
Software Department
Lviv Polytechnic National University
Lviv, Ukraine
ramelnyk@polynet.lviv.ua

Yurii Havrylko
Software Department
Lviv Polytechnic National University
Lviv, Ukraine
yurii.havrylko@gmail.com

Ivan Mykulanynets
Software Department
Lviv Polytechnic National University
Lviv, Ukraine
bananachevalier@gmail.com

***Abstract*—In this paper the K-means clustering algorithm was used to blur segments of image and amplify defects regions of the fabric samples. The usage of the comparison formula to transformed images allows to detect coordinates of defects and to measure their intensity marked proportionally to the difference between the etalon and defective images.**

Keywords—fabric image, intensity, defects, clustering, pixel, comparison, silhouette.

I. INTRODUCTION

There are many works devoted to fabric defects detection algorithms. Many of them consider detection defects on images with repeated patterns.

Among them the papers [1,2] propose algorithms of the Fourier Transform to get the domain image. It is a base to extract features connected with the most significant frequency. Comparison with no defective structure was made for images got by the sliding window. The difference between the signal for the etalon image and a sample image is calculated. Variable are parameters of a size of sliding window, correlation coefficients etc. In the second work a spectral residual (SR) and "energy-map" (EM) methods are used for threshold segmentation. Test images are resized.

In the work [3] for texture-periodicity in wallpaper groups a family of Gabor wavelets are used. The proposed method is periodicity-based Images must have at least two periodic units. Each cropped image is split into several periodic blocks Ward's hierarchical clustering is used to identify defective and defect-free blocks.

In the paper [4] visual inspection systems for textile fabric defects detection is developed based on features extraction from segmented images.

The study [5] proposes the method to classify metal defects using computer vision and machine learning. Training models are used to a type of defect on the steel plate.

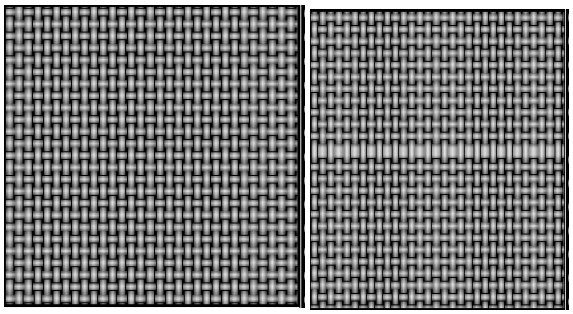
The work [6] presents the most powerful instrument for fabric defects detection as a special framework. It is based on gradient minimization and fuzzy c-means method. It allows to detect various types of defects. The algorithm has two steps: elimination of background and iterative clustering to get centers of defects.

By this review of the short publication list we conclude that there are many various mathematical instruments to detect fabric defects. They differ between themselves by complicated procedures, types of fabric defects and accuracy for given time costs. In common the majority of the above-mentioned approaches are quite time-consuming to be realized. In our paper we present very simple approach based on the clustering algorithm and comparison of images.

II. IMAGE ANALYSIS

To consider the approach for fabric defects detection we use some examples of the simulated images of a plain weave defects from the work [1].

Two samples are given in Fig. 1: the etalon image and with marked defects. Defects on the second sample are visible and contrast enough.



a) b)

Fig. 1. Etalon (a) and defective image (b)

The simplest approach to determine a location of defects consists in comparison by subtraction of two simulated images.

In our case we can not use the logical OR operation as it is usually used for binary images. For gray images we propose the specific logical OR operations which for different operands results red color:

$$
\begin{aligned}
&I_r(x,y) = I_d(x,y),\\
&\forall\,|\,I_e(x,y) - I_d(x,y)\,| < Tol,\\
&I_r(x,y) = RGB(|\,I_e(x,y) - I_d(x,y)\,|,00)\\
&\forall\,|\,I_e(x,y) - I_d(x,y)\,| \geq Tol
\end{aligned}
\qquad (1)
$$

where $I_r(x,y)$ is a pixel intensity of the resulting image, $I_e(x,y)$ is a pixel intensity of the etalon image, $I_d(x,y)$ is a pixel intensity of the fabric image with defects, Tol is a tolerance value to control a difference between an etalon and the controlled samples pixel intensity.

Without a tolerance specific logical OR operations, we present as follows:

$$I(I_s \vee I_s) = I_s, I(I_e \vee I_d), red(|I_e - I_d|)$$

Applying the formula (1) to images from Fig. 1 we get the resulting image with marked defects in Fig. 2.

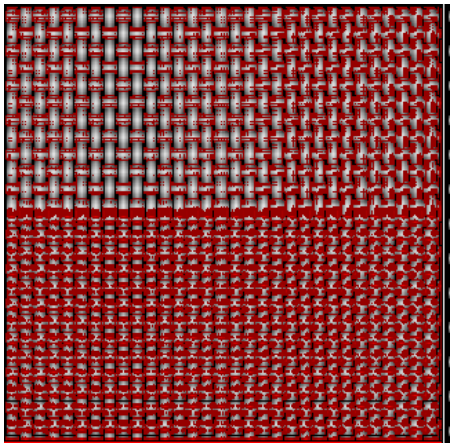


Fig. 2. Specific difference between the etalon and the image with defects

First of all, we see that the etalon image differs from the defective image essentially, mainly in the lower part. There are two reasons of this fact: the images were formed independently or one of them was shifted relatively to the other. That is why we see in the resulting image many red pixels.

To evaluate positions and levels of the different figures in the images we use a pixel intensity function in the W columns and H rows of the matrix:

$$\bar{I}(i) = 1/W(\sum_{j=1}^{W} I(i,j)), i = 1,2,...,H,$$

$$\bar{I}(j) = 1/H(\sum_{i=1}^{H} I(i,j), j = 1,2,...,W,$$

where $I(i,j)$ is pixel intensity in *i*-row and *j*-column ($1 \le i \le H, 1 \le j \le W$).

For the etalon and the defective image these functions in rows are presented in Fig. 3.

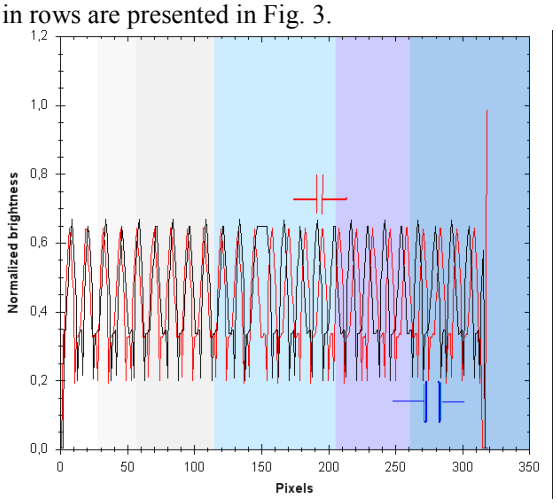


Fig. 3. Intensity functions for the etalon and defective image

Functions in Fig. 3 give us distances in pixels between the peaks of the same image:

$$
\begin{aligned}
&d_e = l_{ei+1} - l_{ei},\\
&\bar{I}_e(l_i) = max(l)\left|\bar{I}_e(l)\right|,\\
&\bar{I}_e(l_{i+1}) = max(l)\left|\bar{I}_e(l)\right|
\end{aligned}
$$

or distances between the peaks of the different images:

$$
\begin{aligned}
&d_{ed} = l_{ei+1} - l_{di},\\
&\bar{I}_d(l_i) = max(l)\left|\bar{I}_d(l)\right|,\\
&\bar{I}_e(l_{i+1}) = max(l)\left|\bar{I}_e(l)\right|
\end{aligned}
$$

where $\bar{I}_e(l_i)$, $\bar{I}_d(l_i)$ are a mean pixel intensity of the etalon and the image with defects, l_i is a point on the OY axis.

For the plots given in Fig.3. these distances are:

$$d_e = 11,\ d_{ed} = 3.$$

The distance between different function indicates how much the etalon object positions differ from defective one. The difference within the same function are being taken into account for a grid step covering the image for future clustering.

III. THE CLUSTERING K-MEANS ALGORITHM

To make the subtraction approach insensitive to small shifts and deviations of sizes we propose to transform two images participating in comparison. We use the K-means algorithm to get these two images. It partitions an image into *K* groups of image segments (clusters). By other words the K-means algorithm approximates the real image with the sample consisting of segments which pixels have the same intensity.

To obtain the elementary leaf as a basis to form clusters we choose a model as a rectangle. They are very simple for building and calculation. So, a virtual grid with horizontal and vertical steps and imaginary thin lines divides the image (Fig. 4) into rectangles.

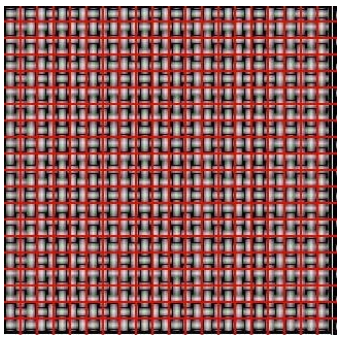


Fig. 4. Coverage of fabric image by a grid

K-means algorithm assigns segments to a cluster such that the sum of the distances calculated by chosen criteria between the cluster's centre and the cluster's members is at the minimum. The most acceptable feature of rectangle as the cluster's member is a mean intensity. The functions to be minimized by the algorithm are as following:

$$S = \sum_{i=1}^{m} \sum_{k=0}^{K} w_{ik} \mid I - \overline{I}_k \mid, S = \sum_{i=1}^{m} \sum_{k=0}^{K} w_{ik} [I - \overline{I}_k]^2 \quad (6)$$

where $w_{ik} = 1$ if the segment of mean intensity I_{ir} belongs to cluster k; otherwise, $w_{ik} = 0$. Also, $\overline{I}_k$ is the intensity of the i's cluster (centroid).

The hierarchical clustering algorithm for data includes the following steps:

S0. For all the points of the input set x_i, $x_j \in X$.

S1. Assignment of K cluster's centers x_i^0 $(i=1,...,K)$ having interval of the image intensity S:

$$x_i^0 = (S/(K+1)) * i, i = 1,..., K,$$

S2. Searching of leaf pairs by the similarity function:

$$\forall\ (x_i^0, x_j \ \text{count}\ F(x_i^0, x_j))$$

S3. Calculating and searching for pairs with the smallest distance value

$$F^*(x_i, x_j) = \min \quad F(x_i, x_j), \qquad i, j \in I,$$

and adding the members x_i, x_j, to of a new member (cluster) x_{nA}.

S3. (Optional) Recalculate the cluster's centers x_i^0 (i=1,…,K)

S3. Remove members x_j from the list for searching.

S4. End of the procedure (for all x_i, $x_j \in X$).

The algorithm builds a binary hierarchical rolling tree (dendrogram) of the points in the clusters by the proximity function.

An example of the clustering process as a dendrogram is shown in Fig. 5.

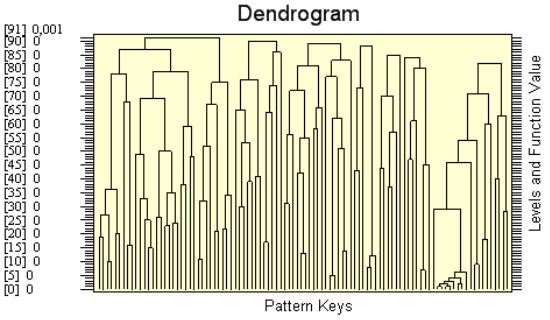
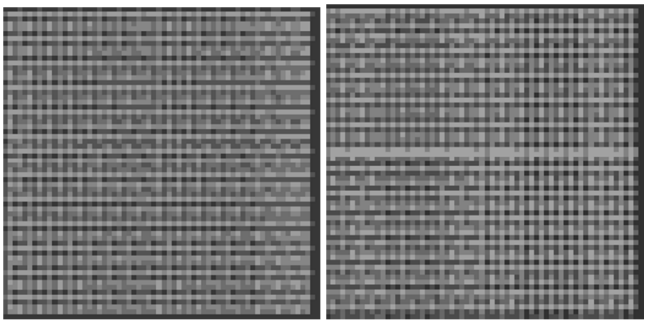


Fig. 5. Dendrogram of clustering process.

IV. DETERMINATION OF DEFECTS BY COMPARISON OF CLUSTERED IMAGES

To illustrate a work of the clustering algorithm we consider two images in Fig. 2 having dimension 319x323. If they are covered by the 32x32 grid then each rectangle has dimension 10x10.

After the rolling up process has been performed the clustered images for 8 clusters were formed and given in Fig. 6.



a) b)

Fig. 6. Fig.6 Clustered images: a - etalon, b - with defects

Applying the formula (6) to the clustered images (rectangles by 10x10) and the tolerance value 55 we obtain the resulting image in Fig. 7.

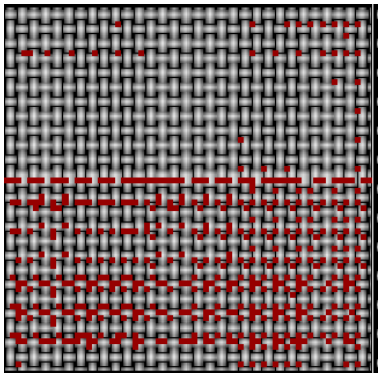


Fig. 7. Specific difference between clustered images

Comparatively to the original image from Fig. 2 we have much fewer red pixels. They are connected with different positions and intensity of sample details of the etalon and the defective image.

Positions and intensity of these details we can observe in Fig. 8, where the mean intensity functions in rows demonstrate them. Especially lower part of the images.

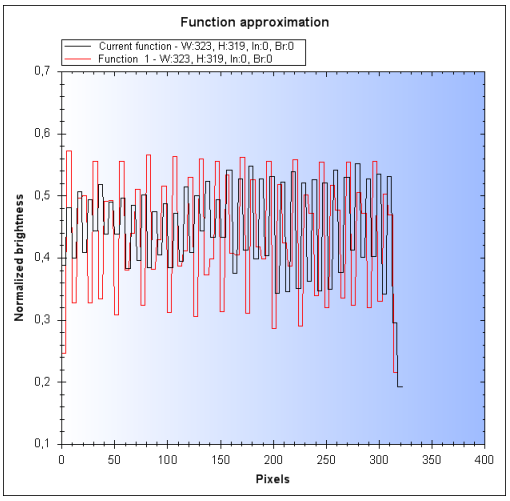


Fig. 8. Intensity functions in rows for clustered images: red is for the etalon, blue is for the defective image

So, we must change control parameters of the clustering algorithm. Having in mind that a distance between peaks of intensity functions in original images is near 11 pixels the minimal size of the rectangle must be of 11 pixels and less. If we cover the image by the 31x31 grid, then each rectangle has dimension 10x10. Corresponding clustered images are given in Fig. 9.

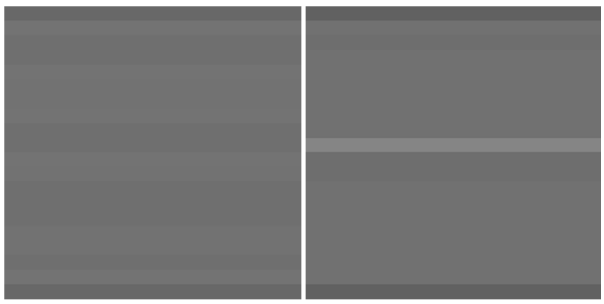


Fig. 9. Clustered images

Now we need to measure a tolerance value for the comparison formula (1). The first method being more universal is less accurate. We calculate the mean intensity function in rows or columns (2) and search the biggest distance betwee two plots

$$I_{tol} = \max\ (s) \,|\, \bar{I}_{ec}(s) - \bar{I}_{dc}(s) |, \; s = 0,\ldots,255 \qquad (7)$$

where $\bar{I}_{ec}(s)$, $\bar{I}_{dc}(s)$ are a mean pixel intensity of the clustered images: the etalon and of the defective fabric image.

In Fig. 10 we see two plots of the previous functions and can determine approximate value of a tolerance. The accurate tolerance value can not be known because the plots are for arithmetic mean intensity of pixels in rows.

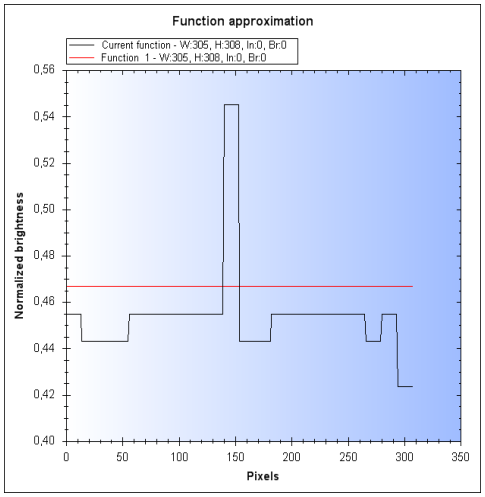


Fig. 10. Mean intensity functions in rows of clustered images

The second approach to measure the tolerance value is a metod based on so-called silhouettes. We find coordinates of pixels with greatest intensity in rows or columns :

$$I_w(x_i) = \{\max\ I(x_i, y_j)\}, \; 1 \le j \le H\}, \; 1 \le i \le W,$$
$$I_H(y_j) = \{\max\ I(x_i, y_j)\}, \; 1 \le i \le W\}, \; 1 \le j \le H. \qquad (8)$$

Then an accurate value of the tolerance is as a maximal difference between silhouettes of the clustered etalon and the defective image:

$$I_{tol} = \max\ (l) \,|\, I_{ewc}(l) - I_{dwc}(l) |. \qquad (9)$$

Two silhouettes for previous clustered images (with step 31) as a view from the OY plane are shown in Fig. 11. For them a value 17 of I_{tol} was measured. Then applying the formula (1) to the clustered images with Tol=17 we get the image with marked defects in Fig. 12.

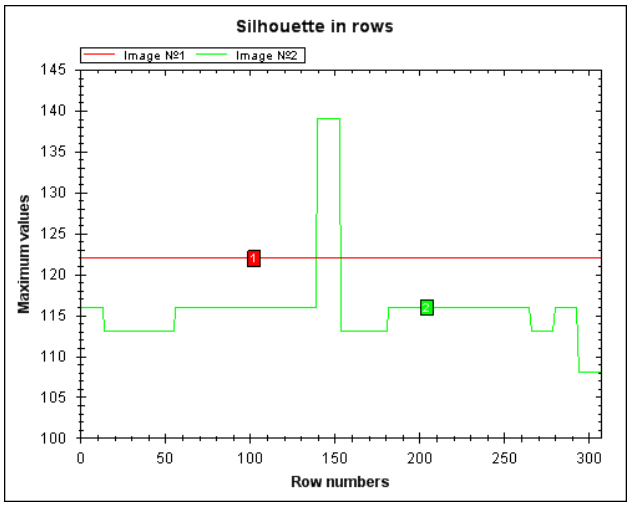


Fig. 11. Silhouette functions of clustered images

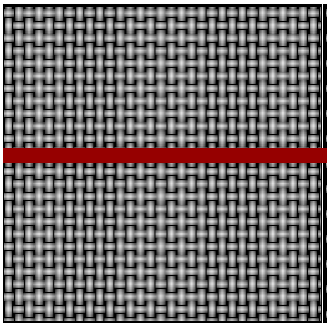


Fig. 12. Resulting image with marked defects

One more experiment we will do with the sample in Fig. 13 having smaller contrast of defects comparatively to mentioned above.

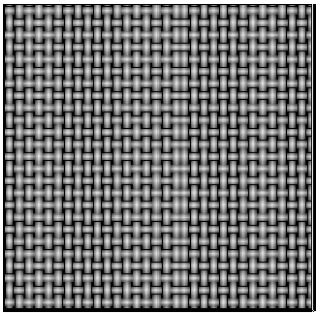
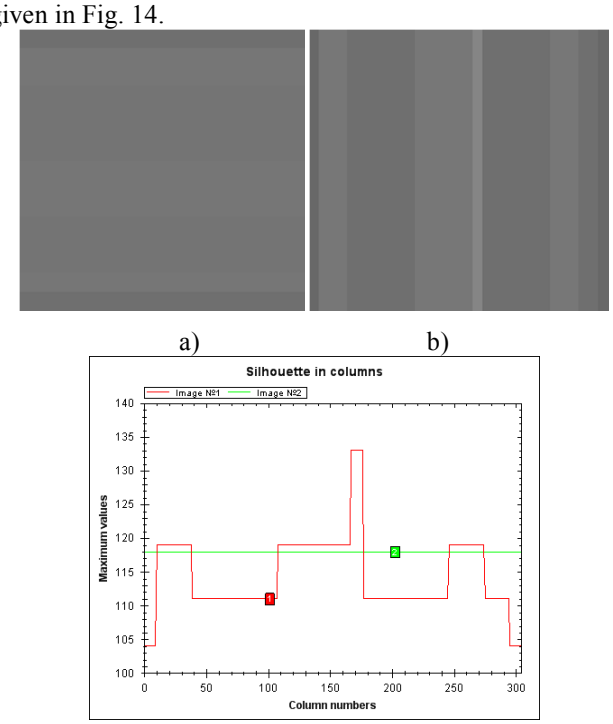


Fig. 13. Defects of small contrast.

Covering the image by the 31x1 grid then each rectangle has dimension 1x320. The corresponding clustered images (etalon and defective) and their silhouettes characteristics are given in Fig. 14.



a) b)

c)

Fig. 14. Clustered images (a,b) and their silhouettes (c)

Applying the formula (9) to the clustered images with Tol=13 we get the image with marked defects in Fig. 15.

From two experiments we conclude that clustering allows determining irregularities of a texture image even if they are not very visible and contrasting. The main question of the approach is detrmination of a step for the covering grid. The approach also allows to measure defects intensity by measurement of red pixels brightness.

$$I(x,y) = \{I(x,y), I_R(x,y) < Tolerance \;\; RGB(,0,0), I_R(x,y) \geq Tolerance$$

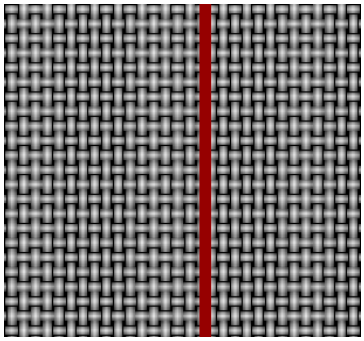


Fig. 15. Resulting image with marked defects

V. CONCLUSION

The algorithm based on comparison of a clustered fabric etalon image and a defective fabric image was developed. It allows detecting coordinates and intensity of defects in fabric images. The algorithm is based on the K-mean clustering which amplifies defects and blurs the lines of drawing. To determine a difference between an etalon and defects the silhouettes function or mean intensity in rows function were being used.

REFERENCES

[1] M. Tunák, A. Linka. "Simulation and recognition of common fabric defects." (2006) [https://www.semanticscholar.org/paper/Simulation-and-recognition-of-common-fabric-defects-Tunak-Linka/2200467a39f02406907ce51ae6087f4076f850d3]

[2] I. Ahn, C. Kim. "Finding Defects in Regular-Texture Images," in Proc. 16th Korea-Japan Joint Workshop on Frontiers of Computer Vision, Hiroshima, Japan, February 2010, pp. 478-480.

[3] V. Asha, N.U. Bhajantri and P. Nagabhushan. "Automatic Detection of Texture Defects using Texture-Periodicity and Gabor Wavelets" In: Venugopal K.R., Patnaik L.M. (eds) Computer Networks and Intelligent Computing. ICIP 2011. Communications in Computer and Information Science, vol 157. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-22786-8_69

[4] K.N.Sivabalan, D.Ghanadurai. "Detection of defects in digital texture images using segmentation." International Journal of Engineering Science and Technology, vol. 2, 2010, pp. 5187-5191

[5] G. Saurabh. "Use Machine Learning to Detect Defects on the Steel Surface", 2018 [https://software.intel.com/en-us/articles/use-machine-learning-to-detect-defects-on-the-steel-surface].

[6] H. Zhang, J. Ma, J. Jing and P. Li. "Fabric Defect Detection Using L0 Gradient Minimization and Fuzzy C-Means." Applied Sciences. (2019). https://doi.org/10.3390/app9173506

IEEE Third International Conference on Data Stream Mining & Processing
August 21-25, 2020, Lviv, Ukraine

The Fast Image Recognition System Based on Neuro-Fuzzy Units and its Online Learning for Data Stream Mining Tasks

Nonna Kulishova
Media Systems and Technologies Department
Kharkiv National University of Radioelectronics
Kharkiv, Ukraine
0000-0001-7921-3110

Yevgeniy Bodyanskiy
Control Systems Research Laboratory
Kharkiv National University of Radioelectronics
Kharkiv, Ukraine
0000-0001-5418-2143

Volodymyr Timofeyev
Department of Information Technologies
Kharkiv National University of Urban Economy
Kharkiv, Ukraine
timofeev2001@yahoo.com

***Abstract*— In recent years, the need for real-time pattern recognition applications has sharply increased. Along with deep and probabilistic neural networks, hybrid architectures such as neo-fuzzy networks and networks based on neo-fuzzy units turned out to be quite effective for solving this problem. In this paper, we consider the task of developing an architecture and a learning algorithm for a real-time pattern recognition system built on neo-fuzzy units. The objective function based on cross-entropy allows formulating a learning criterion that provides a high learning rate.**

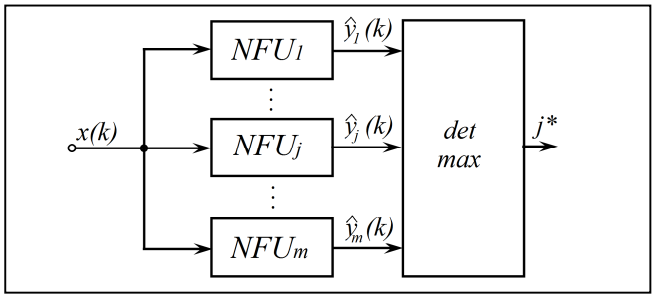
Keywords—neo-fuzzy units, real-time image recognition, cross-entropy learning criterion

I. INTRODUCTION

An important part of Data Mining and Data Stream Mining is general problem of pattern classification [1], for which many techniques have been worked out for now. The most powerful ones from accuracy point of view for today are deep neural networks (DNNs) [2, 3]. However, DNNs are characterized by a low duration of the learning process, which makes it difficult to use them in Data Stream Mining tasks, where the data from the training set enter to system sequentially in online mode. One of the fastest are probabilistic neural networks (PNN) [4], using so-called "lazy learning". Nevertheless, PNN needs large volumes of the training samples, the number of neurons in the patterns layer increases dramatically, which makes this system too cumbersome. In addition, in real tasks, the generated classes are, as a rule, mutually overlapped, which, of course, requires the use of the fuzzy classification apparatus. In this connection, the extended generalized neo-fuzzy network (EGNFN) and algorithm for its learning in real-time mode was proposed in [5]. The "building" blocks of EGNFN are neo-fuzzy neurons [6-8], and are a simplistic version of such universal approximators as Takagi-Sugeno-Kang neuro-fuzzy system. EGNFN showed good results in practical tasks solving, however, it has a five-layer architecture, which, of course, complicates its training. In tasks where information processing rate comes to the fore, this system loses its advantages. In this regard, we propose to use as nodes in the online pattern recognition system not standard elementary Rosenblatt perceptrons nor more "advanced" Yamakawa neo-fuzzy neurons, but so-called neuro-fuzzy units (NFU) [9-11], which are "hybrids" of these neurons and have considerable approximating properties and differ noncomplex learning, which is easy to organize in real time.

II. ARCHITECTURE OF IMAGE RECOGNITION SYSTEM, BASED ON NFU

The architecture of regarded system is presented in Fig. 1 and is essentially a set of parallel-connected neo-fuzzy units NFU_j ($j = 1, 2, ..., m$, m is amount of presumable classes in the training set). System output signals $\hat{y}_j$ are formed that are the levels of belonging of the presented images to certain classes. The output layer is an elementary maximum detector that extracts from first layer output signals the maximum value and



its corresponding class j^*, as it is implemented in standard PNNs.

Fig. 1. The image recognition system based on neuro-fuzzy units

Thus, the training sample is a vector sequence $x(k) = (x_1(k), ..., x_i(k), ..., x_n(k))^T \in R^n$, $X = \{x(1), ..., x(k), ..., x(N)\}$ where $k = 1, 2, ..., N$ in general case - the current discrete time, while the volume N is unlimited and can vary in time, $\hat{y}_j(k)$ describes reaction of neuro-fuzzy unit j to the signal $x(k)$, and the larger $\hat{y}_j(k)$ value, the higher certainty degree that image $x(k)$ appertain to class j. The assumption that the learning set X contains representatives of all m classes is also natural.

The image recognition system contains m NFU_js (one for each class) that process input information in parallel and independently, and if the training set data are fed to the neuro-

fuzzy units, the learning process comes down to nonlinear synapses NSU_{ji} synaptic weights adjusting [6-8].

Fig. 2 shows neuro-fuzzy unit (NFU_j) architecture, and Fig. 3 - nonlinear synapse NSU_{ji}.

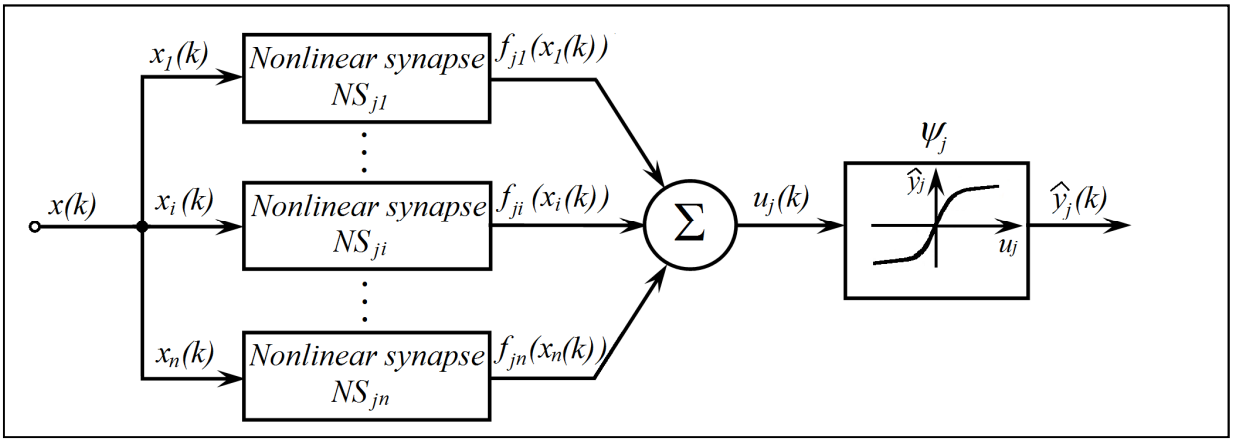


Fig. 2. Neuro-Fuzzy Unit (NFU_j)

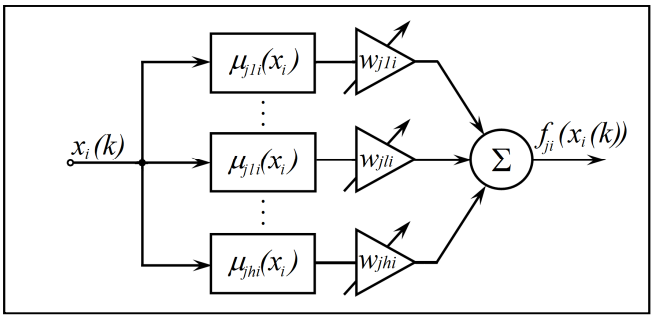


Fig. 3. Nonlinear Synapse NSU_{ji}

Each NSU_{ji} is built from h membership functions $\mu_{li}(x_i)$ and described by h synaptic weights w_{li}, which could be customized during learning sample data processing by optimizing the chosen goal function - learning criterion. As a membership function, it is advisable to use g-th degree B-splines ($g = 1, 2, 3, \ldots$) that respond the Ruspini conditions of unity decomposition:

$$\mu_{li}^{g}(x_i) = \begin{cases} \left\{ \begin{array}{l} 1,\ if\ x_i \in \left[c_{li}, c_{l+1,i}\right] \\ 0\ \text{otherwise,} \end{array} \right\} & \text{for } g = 1, \\ \dfrac{x_i - c_{li}}{c_{l+1,i} - c_{li}} \mu_{li}^{g-1}(x_i) + \\ + \dfrac{c_{l+g,i} - x_i}{c_{l+g,i} - c_{l+1,i}} \mu_{l+1,i}^{g-1}(x_i) & \text{for } g > 1 \end{cases}$$

where c_{li} are membership functions centers, rather randomly located in input signal determining interval.

If the rate comes to the fore in learning process, it is advisable to use second degree B-splines ($g = 2$), that are essentially standard triangular membership functions. Moreover, at each current time moment, only two neighbor membership functions are fired, i.e. in each non-linear synapse, only two synaptic weights are adjusted.

Thus, the nonlinear synapse output can be represented as

$$f_{ji}\left(x_i(k)\right) = \sum_{l=1}^{h} w_{jli}\mu_{jli}\left(x_i(k)\right)$$

and NFU_j output signal is

$$\hat{y}_j(k) = \psi_j\left(u_j(k)\right) = \psi_j\left(\sum_{i=1}^{n} f_{li}\left(x_i(k)\right)\right) =$$
$$= \psi_j\left(\sum_{i=1}^{n}\sum_{l=1}^{h} w_{jli}\mu_{jli}\left(x_i(k)\right)\right)$$

where $\psi_j(\bullet)$ is the nonlinear activation function adopted in Rosenblatt perceptrons and satisfying G. Cybenko theorem conditions [12], for example, sigmoid or

$$\psi_j\left(u_j(k)\right) = \tanh u_j(k).$$

In this case, the each neuro-fuzzy unit NFU_j output signal can be written finally as

$$\hat{y}_j(k) = \tanh\left(\sum_{i=1}^{n}\sum_{l=1}^{h} w_{jli}\mu_{jli}\left(x_i(k)\right)\right).$$

III. Online Learning of Image Recognition System Basing on NFUs

The learning of proposed system is based on the goal function associated with cross-entropy [1, 13]

$$E_j(k) = \frac{1}{2}\left(1 + y_j(k)\right)\ln\frac{1 + y_j(k)}{1 + \hat{y}_j(k)} +$$
$$+ \frac{1}{2}\left(1 - y_j(k)\right)\ln\frac{1 - y_j(k)}{1 - \hat{y}_j(k)}$$

where $y_j(k)$ is the exemplary signal, taking only two values: +1, if specific image $x(k)$ pertains to class j, and -1 otherwise. This criterion is always positive, except in the case $y_j(k) = \hat{y}_j(k)$ (perfect training). Applying the hyperbolic tangent activation function, we finally obtain the following learning rule

$$w_j(k) = w_j(k-1) + \eta(k)\left(y_j(k) - \hat{y}_j(k)\right)\mu_j\left(x(k)\right) = \\ = w_j(k-1) + \eta(k)\left(y_j(k) - \right. \\ \left. - w_j^T(k-1)\mu_j\left(x(k)\right)\right)\mu_j\left(x(k)\right) = \\ = w_j(k-1) + \eta(k)e_j(k)\mu_j\left(x(k)\right) \quad (1)$$

where $\eta(k)$ - is a learning rate parameter, $w_j(k) = \left(w_{j11}(k),...,w_{jl1}(k),...,w_{jh1}(k),...,\right.$ $\left. w_{j1i}(k),...,w_{jli}(k),...,w_{jhi}(k),...,\ w_{j1n}(k),...,w_{jhn}(k)\right)^T$, $\mu_j\left(x(k)\right) = \left(\mu_{j11}\left(x_1(k)\right),...,\mu_{jl1}\left(x_1(k)\right),...,\mu_{jh1}\left(x_1(k)\right),...,\right.$ $\mu_{j1i}\left(x_i(k)\right),...,\mu_{jli}\left(x_i(k)\right),...,\mu_{jhi}\left(x_i(k)\right),...,$ $\left.\mu_{j1n}\left(x_n(k)\right),...,\mu_{jhn}\left(x_n(k)\right)\right)^T$ - $(hn\times 1)$ -vectors.

Since the adjustable synaptic weights in the learning algorithm (1) linearly depend on the learning error $e_j(k)$, it can be optimized for rate, taking on form of one-step optimal Kaczmarz-Widrow-Hoff algorithm:

$$w_j(k) = w_j(k-1) + \frac{y_j(k) - \hat{y}_j(k)}{\left\|\mu_j\left(x(k)\right)\right\|^2}\mu_j\left(x(k)\right)$$

or in matrix form for the system as a whole

$$\mathbf{w}(k) = \mathbf{w}(k-1) + \frac{\left(\mathbf{y}(k) - \hat{\mathbf{y}}(k)\right)\boldsymbol{\mu}^T\left(x(k)\right)}{\left\|\boldsymbol{\mu}\left(x(k)\right)\right\|^2} = \\ = \mathbf{w}(k-1) + \mathbf{e}(k)\boldsymbol{\mu}^+\left(x(k)\right) \quad (2)$$

where $\mathbf{w}(k)$ is the $(m\times mhn)$ synaptic weights matrix, $\mathbf{y}(k) = \left(y_1(k),...,y_i(k),...,y_m(k)\right)^T$, $\hat{\mathbf{y}}(k) = \left(\hat{y}_1(k),...,\hat{y}_i(k),...,\hat{y}_m(k)\right)^T$ are $(m\times 1)$ vectors, $\boldsymbol{\mu}\left(x(k)\right) = \left(\mu_1^T\left(x(k)\right),...,\mu_j^T\left(x(k)\right),...,\mu_m^T\left(x(k)\right)\right)^T$ is the mhn vector of all membership functions, $(\bullet)^+$ is the pseudo-inversion symbol.

Algorithm (2) can be given additional filtering properties by introducing an outdated information exponential smoothing into it. As a result, we come to a matrix modification of learning procedure [14]

$$\begin{cases} \mathbf{w}(k) = \mathbf{w}(k-1) + \\ +\mathbf{r}^{-1}(k)\left(\mathbf{y}(k) - \mathbf{w}(k-1)\boldsymbol{\mu}\left(x(k)\right)\right)\boldsymbol{\mu}^T\left(x(k)\right), \\ \mathbf{r}(k) = \alpha\mathbf{r}(k-1) + \left\|\boldsymbol{\mu}\left(x(k)\right)\right\|^2, 0 \le \alpha \le 1 \end{cases} \quad (3)$$

where the component $0 \le \alpha\mathbf{r}(k-1) \le \sum_{\tau=1}^{k-1}\left\|\boldsymbol{\mu}\left(x(\tau)\right)\right\|^2$ plays the role of momentum term, providing the algorithm with additional regularization.

IV. Experiments

Many applications use information about user reactions now. For example, analysis of this information, allows changing interface versions, offering individual options or services, making decisions about actions in relation to the user (wake a driver who has fallen asleep). A person's reaction can be estimated by breathe frequency, heart rate, characteristics of speech, gestures or facial expressions.

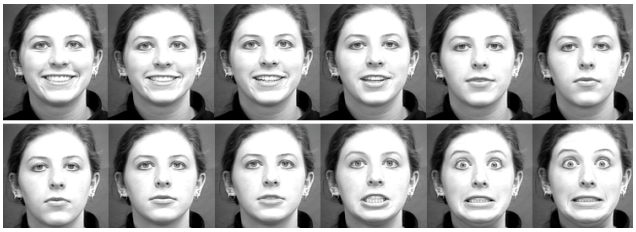
In [15], it was noted that only 7% of information in human communications carries speech, 38% is provided by voice modulations and 55% by facial expression. The person visage in a special way demonstrate not only a sentiment fleeting change, but also mood in general, state of health, and so on.

Such studies are increasingly being taken in multimedia applications that are used in education. Given the students individual reactions during interactive lessons, application can manage educational content. In the distance learning conditions, this approach becomes very effective, because it allows individualizing the educational process, make it flexible, but retaining its focus [16].

The proposed architecture was used to recognize emotions in a real-time video. To study developed architecture properties in detail, it was decided to use datasets that contain individual video frames that demonstrate the change in facial expression in dynamics. These frames can be played back as video also. Datasets CK+ and PICS [17, 18] were selected for the study. Their feature are facial images of people whose emotions change over time (fig. 4). In the database, photos representing 7 main emotions - neutral, distress, joy, aversion, rage, anxiety, amaze - were used.

Fig. 4. Images for system learning and testing

Characteristic points associated with facial features location are highlighted in the images (Fig. 5). They correspond to the well-known FACS system for describing



studied emotions by facial movements [19].

The task of emotions recognition peculiarity is that only a small data set can be provided for recognition system learning. A data set of 507 photos was generated in the study, 102 photos were taken from the PICS dataset, remaining 405 - from the CK+ dataset. For the experiment, images in which people placed in front, without head turning and tilting, the level of illumination is average, glare and dimming are missing were selected. People in the photo are present without glasses, tattoos, mustaches, beards or other artifacts on faces. The main criterion for selecting photos was the completeness of emotion alternation from least noticeable degree to greatest affective expression. In these datasets, not all samples contain such a variation range for each emotion. In addition, it was necessary to achieve equal representation of all seven above emotions in the experiment. However, it turned out that taking into account the main criterion for selecting photos; it was not possible to achieve equal number of samples.

From selected set of photos 60% are randomly selected for training the system; the remaining 40% are used for testing.

System learning errors are shown in fig. 6 – 8 and in the tables 1 – 3.

The membership functions number of in the experiments was different - from 5 to 11. The number of learning eras was varied from 1000 to 10000. The best combination of accuracy and speed is provided by 9 - 11 membership functions, with 7000 - 10000 eras.

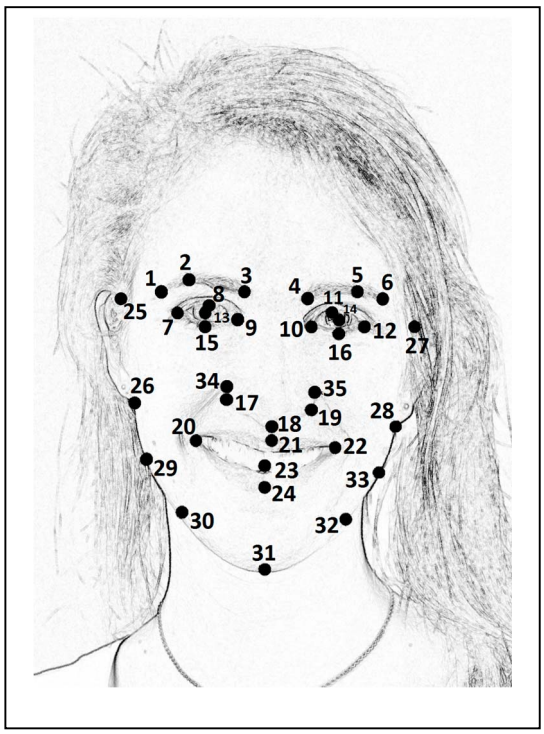


Fig. 5. Feature points for mood recognition

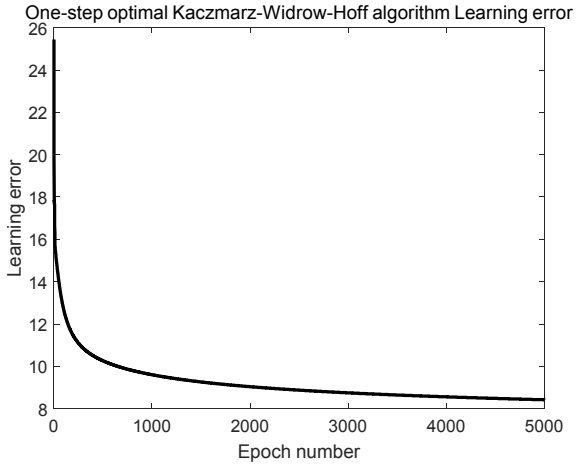


Fig. 6. System learning error for 9 membership functions, 5000 learning epochs

TABLE I. NUMBER OF UNRECOGNIZED IMAGES FOR 9 MEMBERSHIP FUNCTIONS, 5000 LEARNING EPOCHS

	Emotions						
	Rage	***Aversion***	***Anxiety***	***Joy***	***Distress***	***Amaze***	***Neutral***
Data set size	40	49	42	51	19	44	59
The percentage of unrecognized images,%	1.5	8.1	16.1	10.0	3.03	2.7	17.8
Total amount of unrecognized photos = 15.79%							

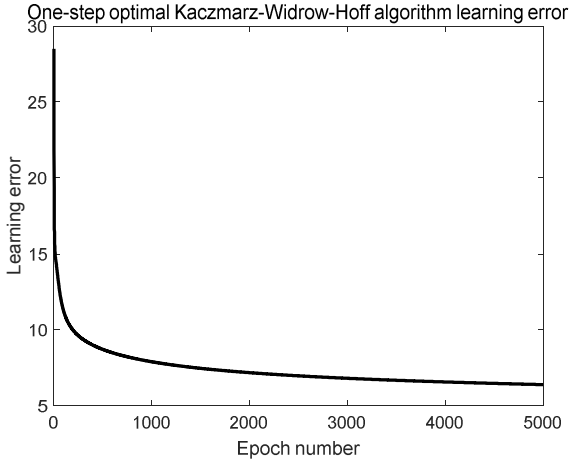


Fig. 7. System learning error for 11 membership functions, 5000 learning epochs

TABLE II. NUMBER OF UNRECOGNIZED IMAGES FOR 11 MEMBERSHIP FUNCTIONS, 5000 LEARNING EPOCHS

	Emotions						
	Rage	***Aversion***	***Anxiety***	***Joy***	***Distress***	***Amaze***	***Neutral***
Data set size	40	49	42	51	19	44	59
The percentage of unrecognized images,%	0	1.2	4.8	2.5	0	0	8.4
Total amount of unrecognized photos = 4.93%							

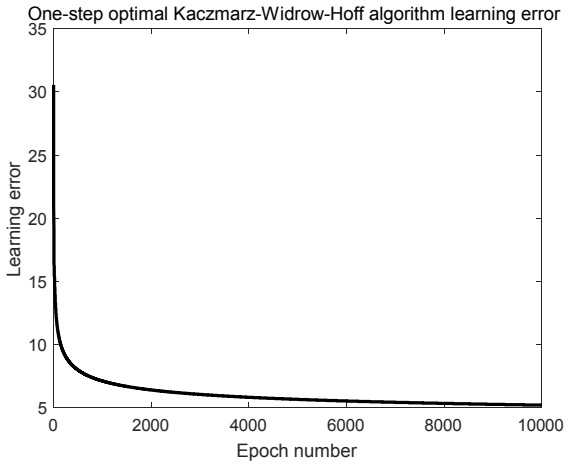


Fig. 8. System learning error for 11 membership functions, 10000 learning epochs

TABLE III. NUMBER OF UNRECOGNIZED IMAGES FOR 11 MEMBERSHIP FUNCTIONS, 10000 LEARNING EPOCHS

	Emotions						
	Rage	***Aversion***	***Anxiety***	***Joy***	***Distress***	***Amaze***	***Neutral***
Data set size	40	49	42	51	19	44	59

The percentage of unrecognized images,%	0	0	1.6	0	0	0	5.6
Total amount of unrecognized photos = 2.3 %							

CONCLUSIONS

The paper considers the problem of recognition in real time. When it comes to the classification of emotions by images or by video, objects classes overlap, do not have constant and clear boundaries, quality criteria are not defined. To solve this problem, authors propose pattern recognition system that built basing on neo-fuzzy units. The hybrid nature of these neurons guarantees high accuracy of approximation, which, in turn, allows for accurate recognition.

The real-time work requires the use of a learning algorithm optimal for speed. For the developed system, the authors propose algorithms that have filtering and adaptive properties.

The experiments showed a high rate of proposed system learning, and its accuracy in recognition problems solving.

REFERENCES

[1] C.M. Bishop, Pattern Recognition and Machine Learning. Singapore: Springer Science+Business Media, 2006, 738 p.

[2] D. Graupe, Deep Learning Neural Networks Design and Case Studies. Singapore: World Scientific, 2016.

[3] J. Goodfellow, Y. Bengio and A. Courville, *Deep Learning*. MIT Press, 2016.

[4] D.F. Specht, "Probabilistic neural network", Neural Networks, 3, № 1, p. 103 – 118, 1990.

[5] N. Kulishova, Ye. Bodyanskiy and I. Pliss, "The Extended Generalized Neo-Fuzzy Network and its Online Learning in Image Recognition Problem", Proc. of the 2019 10th IEEE International Conference on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications (IDAACS), vol. 1, p. 34 -39, Metz, France, September 18-21, 2019.

[6] T.Yamakawa, E. Uchino and T. Samatsu, "Wavelet neural networks emloying over-complete number of compactly supported non-orthogonal wavelets and their applications", Proc. of 1994 IEEE International Conference on Neural Networks, Orlando, FL, USA, p. 1391 – 1396, 1994.

[7] J. Miki, T. Yamakawa, "Analog implementation of neo-fuzzy neuron and its on-board learning," in Computational Intelligence and Applications, N.E. Mastorakis, Ed. Piraeus: WSES Press, 1999, p. 144 – 149.

[8] E. Uchino and T. Yamakawa, "Soft computing based signal prediction, restoration and filtering," in *Intelligent Hybrid Systems: Fuzzy Logic, Neural Networks and Genetic Algoritms*, Da Ruan, Ed. Boston: Kluwer Academic Publishers, p. 331 – 349, 1997.

[9] Ye. Bodyanskiy and S. Popov, "Neuro-fuzzy unit for real-time signal processing", Proc. IEEE East-West Design & Test Workshop (EWDTW' 06), Sochi, Russia, p. 403 – 406, 2006.

[10] Ye. Bodyanskiy, S. Popov and T. Rybalchenko, "Multilayer neuro-fuzzy network for short term electric load forecasting", Lecture Notes in Computer Science, 5010, Berlin Heidelberg: Springer-Verlag, 2008, p. 339 – 348.

[11] Ye. Bodyanskiy, S. Popov and M. Titov, "Robust learning algorithm for networks of neuro-fuzzy units", in Innovations and Advances in Computer Sciences and Engineering, T. Sobh, Ed. Dordrecht: Springer Science + Business Media B.V., 2010, p. 343 – 346.

[12] G. Cybenko, "Approximation by superposition of a sigmoidal function", Math. Contr. Sign. Syst., 2, p. 303 – 314, 1989.

[13] A. Cichocki and R. Unbehauen, Neural Networks for Optimization and Signal Processing, Stuttgart: Teubner, 1993.

[14] Ye. Bodyanskiy, V. Kolodyazhniy and A. Stephan, "An adaptive learning algorithm for a neuro-fuzzy networks", Lecture Notes in Computer Science, 2206, Berlin: Springer, 2001, p. 68 – 75.

[15] Mehrabian A., 1968. Communication without words // Psychology Today, vol. 2, No. 4. – Pp. 53 – 56.

[16] Using emotion recognition technology to assess the effects of different multimedia materials on learning emotion and performance /Chen Chih-Ming, and Hui-Ping Wang // Library & Information Science Research, 33, 244–255.

[17] P. Lucey, J.F. Cohn, T. Kanade, J. Saragih, Z. Ambadar, 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.

[18] 2D face sets. Available online: URL http://pics.psych.stir.ac.uk/2D_face_sets.htm (accessed on 03 of October 2018).

[19] P Ekman., W.V. Friesen, J.C. Hager, Facial Action Coding System, A Human Face, Salt Lake City, USA, 2002.

IEEE Third International Conference on Data Stream Mining & Processing
August 21-25, 2020, Lviv, Ukraine

The Principles of Organizing the Search for an Object in an Image and the Selection of Informative Features based on the Visual Perception of a Person

Vitaliy Boyun,
Department of Intelligent Real Time Video Systems
V.M.Glushkov Institute of Cybernetics NASU
Kyiv, Ukraine
vboyun@gmail.com

***Abstract*— A significant expansion of the scope of computer vision, in particular in real-time systems, places very high demands on them in terms of productivity and efficiency of information processing, and in feedback systems, it also requires information lag in it. Such requirements are not ensured by traditional approaches. The way out of the situation may be to use as a prototype the principles of organization of the human visual system, which has a very high selectivity of perception of video information. The paper presents a generalized dynamic model of the organization of these principles. It is proposed to use them to organize the search for an object in a coarse image of a scene, to track an object and, if necessary, to carry out its classification or recognition at a more detailed level.**

Keywords*— retinal neural network; receptive fields; image preprocessing; object search; informative signs; local (ring) organization of neurons; adaptation mechanisms; intelligent video systems

I. INTRODUCTION

Intelligent video cameras and real-time video systems play a large role in automation systems for production processes, visual quality control of products, robotics, security and military systems, automation systems for scientific and biomedical research, etc. Moreover, the range of their application, the requirements for them are constantly expanding. This is especially true for video systems with feedback, where the results of real-time information processing are used to control the process or for other actions. Such systems put forward increased demands not only on the performance of computing facilities, but also on the lag of information in the feedback loop, which are not provided within the framework of traditional approaches [1,2].

On the other hand, the human visual system has improved over millions of years and has reached an extremely high level of organization. Therefore, the phenomenon of vision provides an extremely many diverse solutions for computer vision systems. Despite the enormous amount of information in the image, and especially in the video sequence, the human visual-analyzing system very effectively and efficiently copes with these problems due to its extremely high selectivity [3-7].

There is a significant semantic gap in how a person perceives and describes an image and how the image is perceived by the video system. A person identifies the semantics of the image, and the video system represents the visual content of the image in the form of its low-level characteristics such as color, texture, orientation, shape, the presence of movement, etc. [3].

The second section briefly discusses the principles of organization of the human visual system, the third - the state of the problem, in the fourth - a generalized model of human perception in dynamics. The fifth section is devoted to the principles of organizing the search for an object in an image, tracking it and extracting informative features based on a person's visual perception.

II. THE PRINCIPLES OF ORGANIZATION OF THE HUMAN VISUAL SYSTEM

The following is a brief description of the structural features of the organization of the retina, the mechanisms and processes associated with the processing of video information in it [3,5,7,8].

The scene image projected on the retina is inverted along two axes (top-bottom and left-right) and is a set of individual points obtained from rods and cones. These points represent, respectively, different light intensities in shades of gray and different wavelengths, perceived by the brain as three basic colors (red, blue and green with their shades). Moreover, dots in shades of gray are mainly located on the periphery of the retina with a low spatial resolution, and colored dots, more densely packed, are located in the central fossa (fovea zone). This representation of the environment is extremely redundant and not informative for the brain; it simply cannot cope with such huge flows of information. In the process of evolution, selective mechanisms have been developed in the human visual system to extract the most informative features for the perception of the environment, reducing the redundancy of scene images by many orders of magnitude.

These mechanisms use the presence in the images of edges, borders of objects that differ in color, texture, shape, orientation, movement in space, which allows using different methods to distinguish between objects. The basic information in this case is the contrast, which, together with the retinal neurons (on- and off-centers) organized by the ring principle and lateral inhibition mechanisms, allows us to distinguish these primary signs of images of objects. Such information is used by the brain to evaluate fragments of the image and to control the further process of perception of the image by the retina depending on the goal (search for a

978-1-7281-3214-3/20/$31.00 ©2020 IEEE

specific object in the scene, tracking it or recognizing it). In the case of contemplation of scenes without a goal (for example, from a car window), perception is carried out reflexively without the intense involvement of the brain, we only see an enlarged picture of flickering objects.

Similarly, perception in tracking mode of a fast-moving object works. In this case, shorter eye movements, with a previously determined scale of the object, carry out tracking.

Visual information, before getting from the eye to the brain, passes through many layers of specialized retinal neurons on the retina. The retina is a complex network of photoreceptors (rods and cones) and nerve cells (neurons). The rods provide a perception of the brightness of achromatic light (in shades of gray), and cones - electromagnetic waves of different lengths, which are perceived by the human visual system as three basic colors (red, blue, green with their shades). The peripheral zone of the retina, on which the rods and partly cones are predominantly located, provides a wide field of view with a low spatial resolution. The dense placement of cones in the central fossa (fovea zone) provides clear color vision.

The peripheral retina and central fossa are organized according to the ring principle (the so-called on- and off-centers). In the on-center, the excitation zone occupies the central part, and around it there is an inhibition zone. In off-centers, the zones of excitation and inhibition are interchanged. Such centers of the peripheral retina are organized using horizontal and diffuse bipolar cells and extract larger spatial features. The centers of excitation-inhibition of the central fossa are organized using horizontal and small bipolar cells and extract more subtle features of the image. The sizes of receptive fields are controlled, respectively, by interplexiform and amacrine cells using horizontal cells. Such an organization allows you to extract spots or points that differ from the background in brightness, color, orientation, texture, etc.

The well-known phrase “the eye looks and the brain sees” quite accurately describes the process of perception, that is, thinking about stimulating our sensory receptors. In this case, ascending and descending processes are involved. Ascending processes are also called processes of transmitting information data from the receptors. And the descending processes, also called the processes of conceptualizing information data, are based on previously acquired knowledge, previous experience, understanding and expectations.

To increase sensitivity in conditions of insufficient illumination, the rods of the peripheral retina are combined into groups with the sum of signals from larger receptive fields, i.e. there is an exchange of spatial resolution for increasing sensitivity in brightness.

There is evidence of the existence of spatial frequency detectors in the visual system. Spatial frequency analysis is a simple and reliable way to describe and generalize the structural details of visual objects. Neurons are tuned to a limited frequency range.

At the level of the retina, a number of adaptation processes operate. In particular, it is the brightness adaptation to the level of illumination (10-12 orders of magnitude: from the threshold of sensitivity of night vision, which is ensured by the excitation of more sensitive rods of the retina, to the threshold of blinding brightness, which is limited by the level of perception of cones of color vision). The subjective brightness, which is perceived by the human visual analyzer, is a logarithmic function of the physical brightness of the light that has entered the eye. However, the range of brightness levels, which is simultaneously perceived by the eye, is about 3 orders of magnitude.

Excitation of retinal neurons in the presence of objects (spots) in the focused image on it, which differ from the background in brightness or color values, spatial frequencies, orientation, and the presence of movement, is used to quickly search for objects in the scene image. It consists in the fact that the place in the scene that caused the excitement, with the help of saccades, is refluxed (focused) into the central fossa for a more detailed analysis [6, 9,10].

In this case, the plasticity property of neurons is used — the formation of receptive fields of various sizes, the change in their shape, and the possibility of extracting informative features from an object that display various physical properties of the object [11]. Special (interlexiform and amacrine) cells control these processes by changing the sizes of receptive fields, the sizes of the excitation and inhibition zones of on- and off-centers, as well as neuron activation thresholds.

In accordance with the theory of integration of the features of the object, the perception of the object is carried out in two stages [11]:

- previous attention - the quick extraction of simple visible features of an object (perceptual primitives), is carried out in parallel throughout the retina automatically without conscious effort and concentration;
- focused attention - requires the observer's efforts and close examination of the object, is carried out sequentially with the help of the central fossa.

Information from diffuse bipolar cells of the peripheral retina and small bipolar cells of the central fossa is transmitted using Gm- and Gp-type ganglion cells, respectively, to the mago- and parvocellular regions of the lateral geniculate body (LGB). In the parvocellular region, color information from the central fossa is also added. These areas, like the retina, are organized according to the concentric ring principle, but operate with larger receptive fields, thus increasing the level of abstractness of the presentation of information. LGB, in addition, performs the function of switching the received and processed information into the corresponding layers of the primary visual cortex of the brain.

Thus, a significant part of the preliminary processing of information is carried out already at the level of the retina. The peripheral zone of the retina is oriented toward a coarse spatial perception of the scene, with a high sensitivity to perception of brightness. The central zone of the retina is focused on clear vision both in spatial resolution and in color.

Using the principles of organizing the retina of the eye allows us to more effectively organize the search for an object in the scene image and preliminary preparation of video data for recognition by extracting low-level informative features of the image.

III. State of the Problem

Quite often, the retina is considered as a structure with a logarithmic-polar organization [12], in the center of which is the central fossa (fovea zone), and the peripheral retina is located around. This provides a wide peripheral inspection and the possibility of a detailed perception of information in the central fossa. Attempts have been made to repeat this foveal organization of visual perception in technical and algorithmic models. One of the directions is the creation of foveal sensors with radial and hierarchical (pyramidal) organization of the receptive field [13]. The disadvantage of radial organization is the need to implement the management of the "look", i.e. direction of the optical axis of the sensor, which requires the use of a high-speed drive and its control system. In addition, the implementation of logarithmic-polar sensors and the subsequent processing of information on them also present significant difficulties.

To speed up the process of searching for an object in the scene, the most commonly used method is the pyramidal organization of the process of taking and processing information [14]. It consists in the fact that at first the image with the maximum resolution is read, and then, by smoothing and thinning, for example, using the Gauss or Laplace pyramids, the next layer of the pyramid is formed, whose dimensions are 4 times smaller than the original. This procedure is repeated several times until the desired level of coarsening is achieved. It is easier to find an object on a coarse image, and then, by applying the reverse procedure, it is possible to restore the original resolution of the image of the object for its recognition, measurement, etc. However, this approach, which was used to reduce the amount of information when transmitting images under conditions of limited channel capacity, does not fully meet the principles of organizing the human visual system and the requirements of real-time systems.

In [15], models of the organization of the retina and central fossa are presented, the principles of the organization of connections on layers of neurons of different specializations, adaptive mechanisms for the perception of lighting and contrast, and the path of signals from receptors to ganglion cells are considered. In the process of modeling, the reliability of these models was confirmed to a large extent; therefore, in some parts of our work, we focus on their use. Close approaches to the principles of processing information on the retina, in terms of extracting informative features, are used in convolution neural networks, which are a further development of the cognitron and neocognitron [16]. They serve as the first layers of recognizing artificial neural networks (ANN). However, convolution neural networks realize only a small part of the arsenal of techniques developed in the process of evolution and embedded in the human visual analyzer.

In particular, are not used:

- mechanisms of “attention” and techniques for quick search for an object on the coarse peripheral retina, followed by detailed analysis in the central fossa;
- adaptive mechanisms for controlling the dynamic range of perception of brightness, contrast and sensitivity in low light conditions.

This leads to a significant increase in the database of objects and scenes, as well as the training time of the ANN to improve the quality of classification and recognition.

In addition, convolution neural networks are currently not designed to connect to real cameras and to work in real time with video sequences.

Neurophysiologists, neuropsychologists, gestalt psychologists, cognitive psychologists, cybernetics consider the human visual analyzer from different angles, from their fields of knowledge, but there is no generalized representation of the process of perceiving the environment, and especially in dynamics. The situation is aggravated by the use in different publications of different terminology of the same processes and the conflicting interpretation of the results, which makes it difficult to understand the complete process of perception.

IV. Dynamic Model of the Processes of Searching for an Object in a Scene, Tracking it and Extracting Informative Features

The generalized model of the human visual system is multifunctional and consists of hundreds of local models that describe a number of structural, physical, biochemical, psychophysical mechanisms and processes. The process of perception of visual information by a person is dynamic, with many parameters that change in the process of perception, with many feedbacks. This is evidenced by a huge number of articles devoted to studies of specific properties, features, mechanisms, etc. [17-22]. However, there is practically no general idea and understanding of the process of perception by the visual system in dynamics for various modes.

In this work, a generalized model of the organization and functioning of the retina, as well as approaches to the implementation of its most important and fundamental principles, which could help increase the intelligence of computer vision systems, are considered from a general perspective. In accordance with this model, enlarged fields of receptors (rods) are created on the retina using horizontal cells, which create a center of excitation using diffuse bipolar cells. With the help of other horizontal cells from receptors adjacent to the receptors of the excitation zone, larger circular inhibition zones are formed, i.e. on- and off-centers are organized. This process is carried out along the entire periphery of the retina in parallel. The results of the work of these centers through Gm-type ganglion cells in parallel and topologically (i.e. with reference to the location on the retina) are transmitted to the LGB.

At the same time, the dimensions of the center and surround gradually decrease, i.e. such centers work, as it were, on different scales of perception. In each center, a search is made for the maximum response of the on-center with the specification of the size of the region. Amacrine cells control the outputs of on- and off-centers and control the change in the size of receptive fields, helping to search for the maximum response of these filters. It should also be noted that the search on the peripheral retina is carried out in parallel on many features: shades of gray, texture, orientation, movement, etc. The flows of this enlarged information through Gm-type ganglion cells enter the region of LGB magnocellular cells organized according to the same center-surround ring principle with local connections between neurons, but operating with higher receptive fields.

Magnocellular cells work with fuzzy images; they are more sensitive to the difference in surface illumination (gray on gray), which contributes to the perception of the depth of the scene. In addition, they quickly respond to stimuli, but the reaction quickly dies away, which is well suited for detecting movement on the scene.

The results of the generalization of information are transmitted from the LGB to the corresponding layers of the visual cortex. In the case of contemplation of the environment or tracking a fast-moving object, this process is repeated many times. If it is necessary to search for an object in the scene for its classification, more detailed examination or recognition, a finer analysis of image fragments that have fallen on the retina is performed. It is possible that the visual cortex of the brain, controlling the response maxima throughout the retina, sets priorities for their sequential "transmission" to the central fossa (fovea zone). The process of "transmission" is carried out by saccades (rapid eye movements) by controlling the focusing system. In the central fossa, organized by the same principle of excitation-inhibition, due to the denser packing of cones, a more detailed analysis, including color, of the obtained image fragment is carried out in order to extract informative features from it. The results of the analysis in the central fossa through ganglion cells of the Gp type enter the region of parvocellular cells of the LGB, which is also organized on the principle of center-environment. These results are also parallel and topologically transmitted to the corresponding cortical layers of the visual cortex. In particular, layer v1 is responsible for frequency and orientation analysis, layers v2 and v3 process information about the shape and position in space, layer v4 is responsible for the perception of color information, layer v5 is for movement.

The visual cortex is organized on a linear basis, reacts to lines, stripes, rectangular segments. It contains simple cells, complex and hypercomplex, i.e. it also, according to the hierarchical principle, increases the level of abstractness of the features of an object for recognition [23]. If the results of the visual cortex using a coarse image of a fragment with additional fine information from the central fossa do not satisfy the search target, then the next image fragment detected on the retina is transferred to the central fossa to check if the searches target matches. This process is repeated for all identified fragments until a final positive or negative result is obtained. If a more detailed analysis is needed, similar to the above, a fragment is analyzed with additional color information for classification, recognition of an object, determination of its geometric dimensions, etc.

The present work on the creation of a generalized model of human perception is not intended to accurately copy the functions of the retina and the processes occurring in it. The aim of the work is to understand the general structure and organization of the human visual system and use these principles to build specialized and problem-oriented video systems for various purposes, taking into account the state and capabilities of modern microelectronics. And, perhaps, at the same time to put before it those problems that will make it possible to more effectively solve the problems of finding an object in an image, tracking it, classifying, recognizing objects, scenes and situations in real time. The advantage of this approach is the use of a number of adaptive mechanisms to increase sensitivity in low light conditions, increase contrast, attention mechanisms and coarse-accurate presentation of the scene, which will significantly reduce the amount of databases and time for training ANN, link them to the processes of shooting video sequences and use in real time video systems.

V. Coarsening (Assumptions) Accepted in the Model

Instead of a logarithmic-polar retina representation system, a Cartesian image representation system with a uniform grid is adopted. Accordingly, the ring organization of on- and off-centers has turned into a square-ring.

Parallel processes of extracting all possible (or necessary for a given perception mode) features of an object at the level of the peripheral retina and central fossa, as well as the processes of their transmission to the higher layers of the brain, have been replaced by sequential or parallel-sequential ones.

The paper does not examine in detail the principles of organization of the primary visual cortex, since they are already actively used in the ANN. The results of the presented study can be used as the first layers of the ANN for preliminary preparation of information for recognition on them. This is an important point, since the analysis of large scenes is computationally significantly more expensive than recognizing the objects themselves, providing a reduction in the amount of information for recognition by many orders of magnitude, although the recognition process is more complicated in the logical sense.

VI. Implementation of the Process of Searching for an Object in an Image, Tracking it and Extracting Informative Features Using the Principles of Human Retinal Organization

The purpose of this section is to highlight the fundamental principles of the organization of the human retina and their implementation, taking into account the capabilities of the modern level of microelectronics. These fundamental principles, which, in our opinion, will help create more advanced computer vision systems, are [24,25]:

- the ability to change the resolution when working with the video sensor;
- the ability to expand the dynamic range of perception of brightness;
- ensuring the principle of ring organization of receptive fields;
- implementation of methods for coarse-accurate search for an object in an image;
- implementation of adaptive mechanisms to the level of lighting and contrast;
- organization of ascending and descending processes;
- specialization of neurons;
- a significant reduction in the amount of information in the video stream due to the selection of informative features.

Let us consider these principles in more detail.

The ability to change the resolution when working with a video sensor will allow, on the one hand, to provide a wide

field of view, and on the other hand, the possibility of a detailed analysis of the selected object. Image coarsening is an effective method of reducing the amount of information when searching for an object in an image, tracking it, and also in the mode of simple contemplation of the situation.

Unfortunately, the modern technology for the production of CMOS video sensors does not fully provide such capabilities. An exception is the possibility of thinning an image into rows and columns when it is being read, which allows 4 times to coarsen the representation of the image, as well as the ability to 4 or more times increase the resolution of a color image compared to the original physical resolution due to linear or quadratic interpolation between adjacent pixels of the image. It is proposed to provide coarsening of the image by summing square (rectangular) fragments of the image, as it is in the case of the peripheral retina of the human eye. Such summation can not only provide a reduction in image size, but also increase sensitivity in low light conditions. This process can be controlled (by analogy with amacrine cells) depending on the average or local brightness of the image by changing the sizes of the summed receptive fields and the gain of the on- and off-centers.

The expansion of the dynamic range of brightness perception can be provided by a nonlinear, for example, a logarithmic, scan of the reference voltage during parallel (throughout the matrix) analog-to-digital conversion (scanning method by parameter) [1,26]. If it is necessary to obtain a binarized image, it is also possible to use a nonlinear sweep of the reference voltage, but a threshold element and a trigger must be integrated in each element of the sensor matrix, which captures the reach of a threshold value by this matrix element [27]. A very important point is the provision of the principle of ring organization of receptive fields, since it equally effectively works both on the model of the peripheral retina and on the model of the central fossa. In addition, on the basis of the center-surround principle, one can distinguish many informative features of a higher level on the retina model (spots that differ from the background in brightness, color, orientation, dynamic characteristics - spatial frequencies, the presence of motion in the video sequence, etc.) that provide the search for objects in the scene. The center-surround principle is suitable not only for processing "raw" RGB colors according to Bayer model, read from the sensor matrix, but also when using the second opponent color theory RG and BY.

It is problematic to organize parallel processing of all signs and sequential on different scales, as is done on the retina, on modern video sensors. However, they can be implemented in parallel-serial or serial-parallel manner on a large number of GPU or on the FPGA. Although with large coarsening on the peripheral retina, it becomes necessary to repeatedly calculate the sums of square or rectangular fragments, this can be effectively solved using an integrated matrix [19]. The same principle can be applied to calculate the center-surround operator (with the same inhibition zone coefficients). In this case, the total amount of the entire center-surround fragment is calculated from 4 points, and then the center area is determined from 4 points, i.e. it takes only 7 operations of addition / subtraction and one shift operation for several digits (using masks with a central coefficient multiple of degree 2).

In the central fossa, according to the principle of *center-surround*, more detailed informative features are extracted (brighter / darker or different in color or orientation points that characterize the edges of areas or contours of objects, spatial frequencies of textures, etc. that stand out from the background). The *center-surround* principle is realized in computer vision using appropriate masks (Laplace, Roberts, Sobel, Prewitt, etc.), which are already widely used in real time image processing techniques.

The ability to change the resolution of the video sensor allows us to implement the method of coarse-accurate search for an object in the image, in which the search is performed on the coarse image for informative features of higher order, which significantly reduces the amount of processed information [6]. These features are transmitted to the central processor or to the recognition layers of the ANN, performing the functions of the primary visual cortex of the brain, where they are compared with the existing models of objects accumulated as a result of previous experience. So the ascending flows of information are organized. If the comparison of the features of the object and the model was unsuccessful, then, in accordance with the priorities, the coarse features of the next object selected in the scene are compared. In the case of a successful comparison of features, depending on the purpose of the search, an object can be accompanied on a coarse image or, if necessary, a given image fragment with a higher resolution can be read for a more detailed examination, classification or recognition of an object. The center of the selected object and its overall dimensions has already been obtained when searching for the object. The central processor in downstream channels controls these processes. With an insufficient level of illumination, by analogy with the grouping of rods on the peripheral retina, the summation of signals from rods from larger receptive fields is performed, which can significantly increase the sensitivity to light perception, i.e. exchange spatial resolution for increased sensitivity. An increase in contrast can be achieved by building up the brake rings around the exciting center, i.e. by using larger masks (e.g. 5–5, 7–7, etc.).

Thus, a significant (by several orders of magnitude) reduction in the amount of processed information is achieved due to the possibility of a coarse-accurate representation of the scene (coarse - when searching for an object or tracking it, and accurate representation of only fragments of the image of the object - when recognizing), using an integrated matrix when implementing the principle of center-surround, etc. [2]. At the same time, due to the selection of informative features, the amount of information transmitted to the highest levels of information processing is reduced by 3-4 orders of magnitude (for example, in the ANN). Using the feature vector allows us to organize other approaches to object recognition [17-19] directly in the most intelligent video camera. A significant increase in productivity can be achieved by constructing a multi layer matrix that combines parallel reading and processing of information (suitable for specialized applications) [25-29].

The advantage of this approach to the implementation of the search for an object and the extraction of features both in a coarse image and in a detailed one is the use of the same sensor. To speed up the process of searching and extracting features, it is advisable to introduce specialized processors into the sensor conveyor for technological image preparation to implement the above principles of constructing an integrating matrix. This will greatly facilitate the

implementation of not only on- and off-centers, but also the implementation of other types of filters (Haar, Hessians, etc.), which are used in other models.

The limited scope of theses does not allow us to display a more detailed picture of the perception process, the technical implementation of the components and the simulation results. They will be presented in a key report and article.

VII. The Discussion of the Results

The advantages of the proposed principles of organizing the search for an object in an image, tracking it and extracting informative features for recognition are more complete use of the arsenal of principles for processing information on the retina of a human visual analyzer. In particular, these are:

- simplification of the process of searching for an object in an image by using a hierarchical coarse-accurate representation of the scene;
- simplification and acceleration of the learning process by expanding the dynamic range of brightness perception and adapting the camera to low contrast conditions and insufficient lighting levels;
- simplification of hardware and software through the use of local connections between neurons in the organization of on- and off-centers and through the use the arsenal of methods developed in theory and practice for image processing to extract informative features;
- a significant increase in information processing productivity by implementing multi layer structures directly on the sensor matrix.

VIII. Conclusion

This work was carried out within the frameworks of fundamental competitive topics (VFC 200.15 and VFC 200.19), which were funded by the Presidium of the National Academy of Sciences (NAS) of Ukraine.

I express my sincere gratitude to the employees of the Department of Intelligent Real-Time Video Systems of the Institute of Cybernetics named after V.M. Glushkov of NAS of Ukraine, which took part in the creation of facilities, modeling and verification of theoretical provisions.

References

[1] V. Boyun *The dynamic theory of information. Fundamentals and applications.* Institute of Cybernetics of NASU, Kyiv, Ukraine, 2001, 326 p. (in Russian)

[2] V. Boyun *Intelligent selective perception of visual information. Informational aspects*. Artificial intellect. 2011. №3. pp.16-24. (in Ukrainian)

[3] H.R. Schiffmann, *Sensation and perception,* 5-th ed. SPb., Russia: Piter, 2003, 928 p. (in Russian)

[4] R. Gonsales, R. Woods, *Digital image processing*. Moscow, Russia: Technosphere, 2005, 1072 p. (in Russian)

[5] V. Boyun, *A human visual analyzer as a prototype for construction of the set of dedicated systems of machine vision*, Transactions of the International conf. "Artificial intelligence. Intelligent systems», II-2010, 2010, vol. 1, pp. 21-26. (in Ukrainian)

[6] V. Boyun, *Intelligent Selective Perception of Visual Information in Vision Systems,* Proceedings of the 6-th IEEE International Conference on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Application. *(IDAACS"2011)*, (Czech Republic, Prague. 15-17 September 2011), vol. 1, pp. 412-416, 2011.

[7] H.Kolb. *How the Retina Works: Much of the construction of an image takes place in the retina itself through the use of specialized neural circuits.*American Scientist. 2003.Vol. 91, No. 1, pp. 28-35

[8] T.Gollisch, M.Meister.*Eye smarter than scientists believed: Neural computations in circuits of the retina.* Neuron.2010 January 28; 65(2). Pp150-164: doi: 10.1016/j.neuron.2009.12.009

[9] Anderson D. *Cognitive psychology. 5-th ed. Piter*, SPb., Russia, 2002. 496 p. (in Russian)

[10] Tagare H.D., Toyama K., Wang J.G. *A Maximum-Likelihood Strategy for Directing Attention during Visual Search*, *IEEE Transactions on pattern analysis and machine intelligence*, vol.23, no.5, May 2001, pp. 490-500.

[11] Podvigin N.F., Makarov F.N., Shelepin Y.E. *Elements of structural and functional organization of visual oculomotor system. Nauka, Leningrad branch*. Leningrad, USSR. 1986, 227 p. (in Russian)

[12] A. Benoit, A. Caplier, B. Durette, J. Herault. *Using Human Visual System modeling for bio-inspired low level image processing. Computer Vision and Image Understanding*. New York, NY, USA, Elsevier Science Inc., 2010. Vol. 114, Issue 7. P. 758-773.

[13] ShelepinY., BondurkoV., Danilova M., *Fovea construction and pyramid structure model of vision system*, *J. Sensor systems*. No.1,1995, pp. 87-97, (in Russian).

[14] P. J. Burt, *Smart Sensing within a Pyramid Vision Machine. IEEE*, vol. 76, №8, pp. 175-185, 1988.

[15] S. Shah, M.D. Levine, *Visual information processing in primate cone pathway - Part I: A model*, *Part II: Experiments,* IEEE Transactions on Systems, Man, and Cybernetics. Part B: Cybernetics. 1996. Vol. 26, No. 2. P. 259-274, 275-289.

[16] A Krizhevsky, I Sutskever, GE Hinton. *Imagenet classification with deep convolutional neural networks* Part of: Advances in Neural Information Processing Systems 25 (NIPS 2012) http://papers.nips.cc/paper/4824-imagenet-classification-with-deep-convolutional-neural-network

[17] H. Yamasaki, T. Shibata. *A Real-Time Image-Feature-Extraction and Vector-Generation VLSI Employing Arrayed-Shift-Register Architecture.* IEEE Journal of Solid-State Circuits, vol.42, no.9, Sept.2007. P.2046-2053.

[18] C. Siagian, L. Itti. *Rapid Biologically-Inspired Scene Classification Using Features Shared with Visual Attention*. IEEE Transactions on Pattern Analysis and Machine Intelligence. Vol.29. NO.2 Febr. 2007. P.300-312.

[19] H.Bay et al. *Speeded-Up Robust Features (SURF). Elsevier, Computer Vision and Image Understanding 110 (2008), p.p.346-359.*

[20] Supin A.J. *Neuron mechanisms of visual analysis.* Nauka. Moscow, USSR. 1974. – 180 p. (in Russian)

[21] Shevelev I.A. Neurons *of visual cortex. Adaptability and dynamics of receptive fields*. *Nauka*, Moscow, USSR, 1984. – 220 p. (in Russian)

[22] Hubel D. H. Eye, *Brain and Vision*. NewYork: Scienceific American, 1988.

[23] Marr D. A *Computational Investigation into Human Representation and Processing of Visual Information*. W.H.Freeman and Company. New York. 1987.

[24] V. Boyun, *Directions of Development of Intelligent Real Time Video Systems*. Application and Theory of Computer Technology, [S.l.], vol. 2, n. 3, pp. 48-66, apr. 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.

[25] Boyun V.P., Voznenko L.O., Malkush I. F. *Principles of Organization of the Human Eye Retina and Their Use in Computer Vision Systems*. Cybernetics and Systems Analysis Springer Nature. Jan 1, 2019 DOI 10.1007/s10559-019-00181-0

[26] V.P.Boyun.*Device for determination and parameters of an object in an image.* Ukraine Patent for an invention, № 76597, BI №6, 10.01.13. (in Ukrainian)

[27] V.P.Boyun. *Sensor device for determination of location and center of gravity of an object.* Ukraine Patent for an invention, № 106292, BI №15, 11.08.14. (in Ukrainian)

[28] V.P.Boyun *Sensor device for determination of location and moments of inertia of an object in an image.* Ukraine Patent for an invention, № 106301, BI №15, 11.08.14. (in Ukrainian)

[29] V.P.Boyun. *Sensor matrix with image processing*. Ukraine Patent for an invention, № 109335, BI №6, 10.08.15. (in Ukrainian)

IEEE Third International Conference on Data Stream Mining & Processing
August 21-25, 2020, Lviv, Ukraine

Parallel multi-head dot product attention for video summarization

Bohdan Bilonoh
Department of Informatics
Kharkiv National University of Radioelectronics
Kharkiv, Ukraine
bohdan.bilonoh@nure.ua

Sergii Mashtalir
Department of Informatics
Kharkiv National University of Radioelectronics
Kharkiv, Ukraine
sergii.mashtalir@nure.ua

***Abstract*—The dominant video summarization deep learning models are based on recurrent or convolutional neural network with a complex architecures. The best performing models also use attention mechanism. We propose a novel method for supervised, keyframes based video summarization by applying a well known Transformer architecture. Current state of the art method based on self-attention netwok. These network is simple and robust but still lower than humman level. We propose model for video summarization based on a powerfull language architecture Transformer which performs a single feed forward and backward pass during training for the entire video sequence to frame score transformation. This approach was adapted for the sequential regression followed by binarized evaluation which allowed to compare predicted summary with human level. Experiments show that model is competitive in performance and have a good parallel training ability also. Our model achieves F_1 48.1 and 50.19 for SumMe in cannonical and augmented setting and 60.12 and 61.5 for TVSum.**

***Index Terms*—video summarization, sequence to sequence, self-attention, Transformer**

I. Introduction

Visual perception of information are becoming dominant part of our life nowadays. Video chatting, online courses and many other forms of information exchange are an integral part of modern reality. According to Cisco Visual Networking Index: Forecast and Methodology, 2016-2021 [1], by 2021 video will account for 80% of all global Internet traffic, excluding P2P channels. Consequently, improving video processing, such as video summarization, is a very important problem.

Video summarization is video represented as reduced sequence of still frames (images) called keyframes that describes information in the video the best. This paper shows video summarization technique based of the keyframes. Unsupervised methods that were widely used low level spatio-temporal features [2]–[4] and clustering techniques [5], [6]. Distance or cost function usage between keyframes does not always give high accuracy in summarization task because of strong bias in features. On the other hand, supervised learning approach can generalize original transofrmation function.

TvSum [7] and SumMe [8] were used as an data for experiments. Each dataset contains 15-20 users' annotations for each video. Mean F-score for human level is 0.34. So it may be extremely difficult to choose the right clastering metric similar to human annotation. Based on this we expect that the video annotation is a subjective problem which is appropriate concept for supervised summarization task.

Latest works for video summarisation are based on sequence to sequence architectures where convolutional or reccurant neural netowrks (CNN, RNN) with bi-directional long short-term memory (LSTM) [9] or gated recurrent unit (GRU) [10] cells. Also attention [11] widely used to increase generalizing ability of a neural network. But in machine translation task new era of models occurred. Transformer [12] architecture shows ability of atetntion mechanism without reccurent netowrks for language modeling. The greatest superiority of such kind of model is computational efficiency. It means that model with more weights can be trained in less time.

Latest model with the hieghts F_1 was developed by J. Fajtl et al. [13]. Attention based network became the main part of architecture but without recurrent mechanism. A. Vaswani et al. [12] shows that natural language processing (NLP) based on encoder-decoder architecture can be done without recurrent units, on dense layers and attentions only. Based on impressive success of Transformers [12], [14], [15] in NLP problems advisable to try this architecture for video summarization problem.

II. Related works

Latest progress in video summarization was inspired by naturale language processing (NLP) approaches such as encoder-decoder networks, sequence to sequence attention. This relates to the nature of the data because decision of current keyframe is related to the previous like in language model. Next we will discuss related methods to our work.

Huge performace improvement in sequence to sequence learning was archived by [12]. Authors proposed to avoid reccuernt units like LSTM or GRU and use only attention and dense layers to encode and decode language data. Based on work [11] multi-head attention was proposed which stacks parallel block of scaled dot-product attentions. This allows the model to jointly learning information from different representation subspaces at different positions. After attention each layer encoder and decoder contains batch of dense layers. Since model did not contain reccurence positional encodding was added. This new family of models is differs high ability of generalization for sequences and high computational efficiency

978-1-7281-3214-3/20/$31.00 ©2020 IEEE

on the other hand it is more memory usage for big amount of weight and time for trainig.

The pioneers of RNN with LSTM cell in video summarization were K. Zhang et al. [16]. They model temporal dependency with variable length to achive representative video summaries with compact characteristics. Ji et al. [17] introduced another encoder-decoder model for supervised video summarization. Based on deep bi-directional LSTM with attention network encodes the contextual information among input video frames. Mahasseni et al. [18] proposed adversarial network that minimizes the distance between video and summary. Zhou et al. [19] proposed encoder-decoder architecture trained by reinforcement learning to get state of the art results in unsupervised summarization. Reward function takes diversity and representativeness of generated summaries into account. Current state-of-art model was proposed by [13] where enceder-decoder architecture with only fully connected and soft self-attention layers were used.

III. Model

Architecture proposed in this work follows encode-decoder model but replaces recurrent units with stack of attentions. $X = (x_0, ..., x_N), x_i \in R^D$ is the input sequence for our model that produces an output sequence $Y = (y_0, ..., y_N), y_i \in [0, 1)$. The input represents a features after concolutional model which are grouped into sequences for each video. Fig. 1 shows the entire network in detail. An attention mechanism can be described as a dictionary based memory from which value can be extracted with specific key which in turn was mapped from query. Weighted sum of the values are prepresented the output. Compatibility function is used to calculate weights for value based on query and its key.

A. Encoder Stacks

The encoder is composed of a stacked layers each of which has two sub-layers. The first one is a multi-head self-attention mechanism [12], and the second is a simple, position-wise fully connected feedforward network Fig. 2a. Also a residual connection [20] and layer normalization [20] was used after each sub-layer.

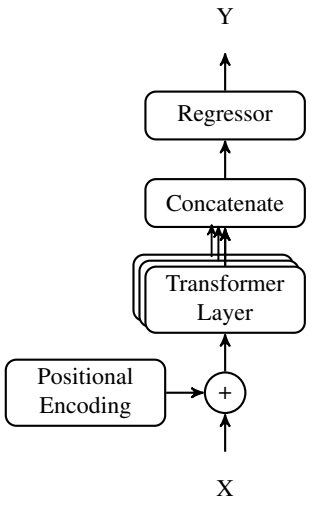


Fig. 1: Network architecture.

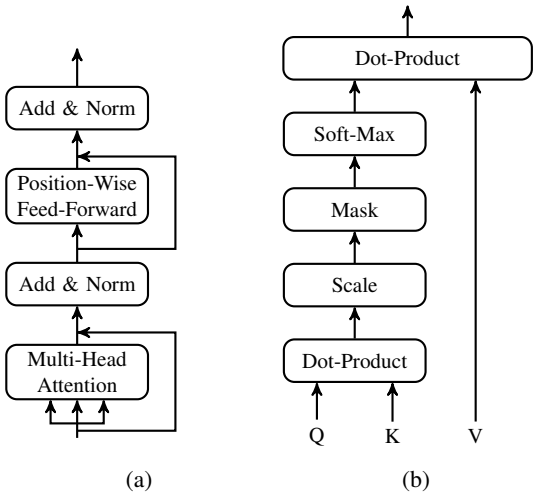


Fig. 2: (a) Encoder layers. (b) Scaled-Dot Product Attention.

B. Scaled Dot-Product Attention

Additive attention [11] and multiplicative attention [21] are mainly used attentions. This functions are compute almost the same output but in different ways. Multiplicative attention is better optimized for cimputational and memory usage because it can be executed on high speed matrix multiplication hardware.

The attention function computes on a matrix Q where queries packed together. Similar to the Q, K and V represents packed keys and values respectively. Output matrix compute as:

$$A(Q, K, V) = \sigma\left(\frac{QK_T}{\sqrt{d_k}}\right) \tag{1}$$

$$\sigma(z) = \frac{e^z}{\sum_{j=1}^{N} e^{z_j}} \tag{2}$$

where d_k is the dimension of the source hidden state. This is modified version of dot-product attention (Fig. 2b). It adds a scaling factor $1/\sqrt{d_k}$, motivated by the concern when the input is large, the softmax function may have an extremely small gradient, hard for efficient learning. Queries and keys form the imput with dimension d_k while values have d_v dimension.

C. Multi-Head Attention

Multi-head attention projects linearly d_{model}-dimensional keys, values and queries h times with different, learned linear projections to d_k, d_k and d_v dimensions. This allows us to perform attention function of these projections in parallel, yielding d_v-dimensional output values. Concatenate layer applied to build state vector, as depicted in Fig 3. Multi-head attention can be described as projection of multispaces for different positions that provides an opportunity to the model

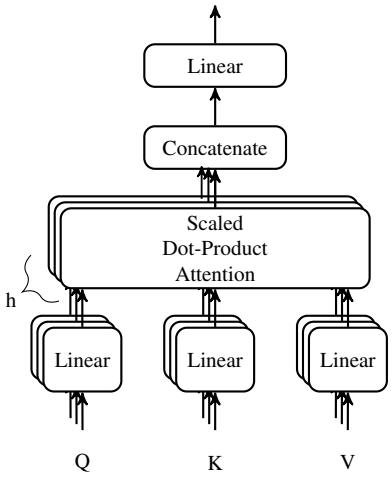


Fig. 3: Multi-Head Attention consists of several attention layers running in parallel.

attend hidden information. A single-head attention can be calculated as:

$$H(Q, K, V) = [h_0, ..., h_h]W^O \tag{3}$$

$$h_i = A(QW_i^Q, KW_i^K, VW_i^V) \tag{4}$$

where W_i^Q, W_i^K, W_i^V and W^O are parameter matrices to be learned.

D. Position-wise Feed-Forward Networks

Fully connected feed-forward network (FFN) was added to each layer in encoder to achive equivalent of applying two 1×1 convolution. The position-wise FFN consists of two dense layers that applies to the last dimension that has the same effect as 1×1 convolution layer. Position-wise was referred to the same two dense layers are used for each position item in the sequence.

$$F(x) = max(0, xW_1 + b_1)W_2 + b_2 \tag{5}$$

where W_1, W_2, b_1 and b_2 are parameter matrices to be learned. In addition, if two frames in the input sequence are identical, the according outputs will be identical as well.

E. Regressor Network

Each result vector for frame from encoder passes to feed-forward network with one neuron to compute regression score.

$$R(x) = S(Wx + b) \tag{6}$$

$$S(x) = \frac{e^x}{e^x + 1} \tag{7}$$

We use sigmoid activation function [22] to compute results in $(0, 1)$ range.

IV. Evaluation

A. Datasets Structure

TvSum [7], SumMe [8], OVP [23] and YouTube [23] were used to carry out all the experiments to correctly compare the results with previous works. Videos from TvSum and SumMe are currenly used for training and validation while OVP and YouTube are for augmentation. Table I provides datasets features.

TABLE I: Overview of the TvSum and SumMe features

Dataset	Videos	Video length (avg)	User annotations	Annotation type
SumMe	25	146	15-18	keyshots
TvSum	50	235	20	frame-level scores
OVP	50	98	5	keyframes
YouTube	39	196	5	keyframes

All keyframe annotations in OVP and YouTube have to be changed to the frame scores and binary keyshots due to the fact that TvSum is annotated by frame scores and SumMe with binary keyshots. This procedure is described in the following section IV-B.

B. Ground Truth Preparation

The model predicts frame scores which are need to be mapped to keyshots. [16] described two steps procedure: scene change points detection; subset with keyshots selection by total frame score maximization. [8] proposed for the total length of summary constraining only 15% of the original video length. Kernel Temporal Segmentation (KTS) proposed by [24] is used for change points detection. For each $i \in K$ detected shot s_i score calculates as:

$$s_i = \frac{1}{l_i}\sum_{n=1}^{l_i} y_{i,\alpha} \tag{8}$$

where i and l_i is the length of i-th shot whith $y_{i,\alpha}$ score for α-th frame. Then keyframes selection is done by Knapsack algorithm [7].

The SumMe contains frame scores calcualted from keyshot user summaries per frame (Fig. 4). So evaluation have to be performed on binary keyshot summaries while traning is done on frame scores. Temporarily segmenting technique was applied to map keyframe annotations to frame scores in

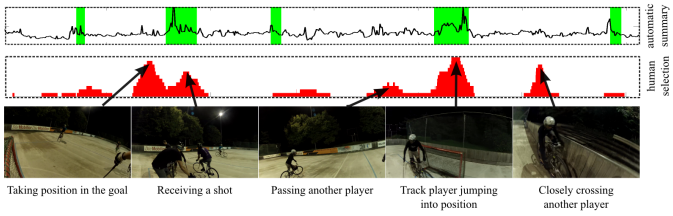


Fig. 4: Ground truth and corresponding keyframes for "Bike Polo" video from SumMe dataset.

OVP and YouTube. Total summary length is constracted by knapsack, however s_i (Eq. 8) is calculated as a ratio of number of keyframes within the keyshot and the keyshot length.

Ground truth data from [16], [18], [19] is adopted to make comperison more correct. Publicly available dataset [16], [19] which containes frame embeddings from pretrained CNN, keyshot labels, frame scores and scene change points are stored in one place. Embeddings were extracted from the GoogLeNet network [25] that was trained on ImageNet [26]. Each embeddings is the output of the pool5 layer and has 1024 dimensions.

Cross validation with 5 folds technique used for canonical and augmented data as proposed by [16]. For the canonical mode train/test data were randomly split into 5 folds for the TvSum and SumMe datasets. The train/test ration was selected as 80% to 20%. In the augmented also cross validation was applied with the same number of folds and splitting as above. One thing was done in with augmented setup is adding videos from the other datasets for training step. When was a training process for the SumMe in the augmented mode videos from TvSum, OVP and YouTube were combined and splitting was applied on all videos.

C. Evaluation

As state of the art work we followed evaluation protocol from [16], [19] and [18]. F_1 score was selected to measure performance between user and machine summaries.

$$F_1 = 2 \times \frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}} \times 100 \tag{9}$$

False negatives and true/false positives for F_1 can be presented as the ration of the machine predictions and ground truth.

The TvSum benchmarking approach proposed by [7] implies the F_1 calculation have to be done as an F_1 over videos. Each video score is an average between the model prediction and each of the user. The closest user summary to the prediction is selected for SumMe benchmarking as proposed by [27] and [28]. Evaluation done on multiple user summaries [16] due to the 15% of the original video length limitation for model summary [8].

D. Results

In Table II shows the comparison of human-level performance and pairwise F_1 score inside the user summaries. Table III presents results of the most recent state of art and ours models. Also human performance was added to show how well the models performed, which is calculated as pairwise F_1 of all user summaries and the ground truth. This fact leads to higher F_1 for human-level performance than user summaries because ground truth converted to the keyshots from an average of all user predictions which are longer than the discrete ones. $F_1 = 36$ reported by [7] is lower than the pairwise $F_1 = 53.8$ because KTS was used to convert each user summary to keyshots and each video was limited to 15% of the length and then calculated the pairwise.

In Table III shows F_1 score for previous models and ours in both canonical and augmented setups. We can see that our model performed better than SASUM$_{sup}$ but worse than VASNet. Attention mechanism in models allows extract more information form SumMe dataset compared to the TvSum, where most models are close to the human performance. Table I shows the reason of the small gain for TvSum dataset and it can be explained by the fact that video sequences are longer than in the SumMe.

TABLE II: Average pairwise F_1 calculated among user summaries and between human-level performance.

Dataset	**Pairwise F_1 score**	
	Among users annotations	**Training GT and users annotations (human performance)**
SumMe	31.1	64.2
TvSum	53.8	63.7

V. Conclusion

This paper propose a novel model for video summarization based on deep neural network with Transformer. This model performs a sequence sequential learning without reccurent units such a LSTM or GRU due to the attention mechanism. Based on previous and our results we shown than video summarization task is better solved in supervised mode. It was also demonstrated that attention based models performs better not only on language modeling and can be applied to almost all sequential data. We designed and tested the simplest Transformer architecture with positional encoding to establish a baseline method for such architectures.

We showed that video summarization with supervised learning based on attention outperforms canonical reccurent networks on the TvSum and SumMe benchmarks. Also adding positional encoding did not improve results. This can be explained by fact that complexity of different visual sequences with loss information on CNN extracting step and small amount of data. Despite this the simplicity of model without reccurent cycles is easier to implement and less resource expensive to run than RNN based methods with same number of weights, making more scalable and computationally efficient. Adding more stacks of Transformer and using deeper network for CNN feature extracting could improve the performance. We are considering these extensions for our future work.

TABLE III: Comparison of our model with the state of the art models for canonical and augmented settings and human performance

Model	**SumMe**		**TvSum**	
	Canonical	**Aug**	**Canonical**	**Aug**
dppLSTM [16]	38.6	42.9	54.7	59.6
M-AVS [17]	44.4	46.1	61.0	61.8
DR-DSN$_{sup}$ [19]	42.1	43.9	58.1	59.8
SUM-GAN$_{sup}$ [18]	41.7	43.6	56.3	61.2
SASUM$_{sup}$ [29]	45.3	-	58.2	-
VASNet [29]	49.7	51.1	61.4	62.4
Human	64.2	-	63.7	-
Proposed	**48.1**	**50.2**	**60.1**	**61.5**

REFERENCES

[1] C. V. N. Index, "Forecast and methodology, 2016–2021," *White Paper, June*, 2017.

[2] S. Mashtalir and V. Mashtalir, "Sequential temporal video segmentation via spatial image partitions," in *2016 IEEE First International Conference on Data Stream Mining & Processing (DSMP)*. IEEE, 2016, pp. 239–242.

[3] O. Kobylin, S. Mashtalir, and M. Stolbovyi, "Video clustering via multidimensional time-series analysis," in *Proceedings of the 9th International Conference on Information Management and Engineering*, 2017, pp. 60–63.

[4] S. Mashtalir and V. Mashtalir, "Spatio-temporal video segmentation," in *Advances in Spatio-Temporal Segmentation of Visual Data*. Springer, 2020, pp. 161–210.

[5] S. Mashtalir, M. Stolbovyi, and S. Yakovlev, "Clustering video sequences by the method of harmonic k-means," *Cybernetics and Systems Analysis*, vol. 55, no. 2, pp. 200–206, 2019.

[6] S. Mashtalir, O. Mikhnova, and M. Stolbovyi, "Sequence matching for content-based video retrieval," in *2018 IEEE Second International Conference on Data Stream Mining & Processing (DSMP)*. IEEE, 2018, pp. 549–553.

[7] Y. Song, J. Vallmitjana, A. Stent, and A. Jaimes, "Tvsum: Summarizing web videos using titles," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 5179–5187.

[8] M. Gygli, H. Grabner, H. Riemenschneider, and L. Van Gool, "Creating summaries from user videos," in *European conference on computer vision*. Springer, 2014, pp. 505–520.

[9] J. Schmidhuber and S. Hochreiter, "Long short-term memory," *Neural Comput*, vol. 9, no. 8, pp. 1735–1780, 1997.

[10] 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.

[11] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," *arXiv preprint arXiv:1409.0473*, 2014.

[12] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in neural information processing systems*, 2017, pp. 5998–6008.

[13] J. Fajtl, H. S. Sokeh, V. Argyriou, D. Monekosso, and P. Remagnino, "Summarizing videos with attention," in *Asian Conference on Computer Vision*. Springer, 2018, pp. 39–54.

[14] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever, "Language models are unsupervised multitask learners," *OpenAI Blog*, vol. 1, no. 8, p. 9, 2019.

[15] Z. Lan, M. Chen, S. Goodman, K. Gimpel, P. Sharma, and R. Soricut, "Albert: A lite bert for self-supervised learning of language representations," *arXiv preprint arXiv:1909.11942*, 2019.

[16] K. Zhang, W.-L. Chao, F. Sha, and K. Grauman, "Video summarization with long short-term memory," in *European conference on computer vision*. Springer, 2016, pp. 766–782.

[17] Z. Ji, K. Xiong, Y. Pang, and X. Li, "Video summarization with attention-based encoder-decoder networks," *IEEE Transactions on Circuits and Systems for Video Technology*, 2019.

[18] B. Mahasseni, M. Lam, and S. Todorovic, "Unsupervised video summarization with adversarial lstm networks," in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2017, pp. 202–211.

[19] K. Zhou, Y. Qiao, and T. Xiang, "Deep reinforcement learning for unsupervised video summarization with diversity-representativeness reward," in *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[20] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[21] M.-T. Luong, H. Pham, and C. D. Manning, "Effective approaches to attention-based neural machine translation," *arXiv preprint arXiv:1508.04025*, 2015.

[22] J. Han and C. Moraga, "The influence of the sigmoid function parameters on the speed of backpropagation learning," in *International Workshop on Artificial Neural Networks*. Springer, 1995, pp. 195–201.

[23] S. E. F. De Avila, A. P. B. Lopes, A. da Luz Jr, and A. de Albuquerque Araújo, "Vsumm: A mechanism designed to produce static video summaries and a novel evaluation method," *Pattern Recognition Letters*, vol. 32, no. 1, pp. 56–68, 2011.

[24] D. Potapov, M. Douze, Z. Harchaoui, and C. Schmid, "Category-specific video summarization," in *European conference on computer vision*. Springer, 2014, pp. 540–555.

[25] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 1–9.

[26] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein *et al.*, "Imagenet large scale visual recognition challenge," *International journal of computer vision*, vol. 115, no. 3, pp. 211–252, 2015.

[27] M. Gygli, H. Grabner, and L. Van Gool, "Video summarization by learning submodular mixtures of objectives," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3090–3098.

[28] C.-Y. Lin, "ROUGE: A package for automatic evaluation of summaries," in *Text Summarization Branches Out*. Barcelona, Spain: Association for Computational Linguistics, Jul. 2004, pp. 74–81. [Online]. Available: https://www.aclweb.org/anthology/W04-1013

[29] H. Wei, B. Ni, Y. Yan, H. Yu, X. Yang, and C. Yao, "Video summarization via semantic attended networks," in *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

IEEE Third International Conference on Data Stream Mining & Processing
August 21-25, 2020, Lviv, Ukraine

A pipeline for detecting and classifying objects in images

Marek Vajgl
Institute for Research and Applications of Fuzzy Modeling
Centre of Excellence IT4Innovations, University of Ostrava
Ostrava, Czech Republic
marek.vajgl@osu.cz

Petr Hurtik
Institute for Research and Applications of Fuzzy Modeling
Centre of Excellence IT4Innovations, University of Ostrava
Ostrava, Czech Republic
petr.hurtik@osu.cz

***Abstract*—With the increased accessibility to neural network frameworks and computation clouds, a wide range of competition websites offer real tasks for the neural network community to join. In this paper, we discuss a problem of object detection and classification. Based on our experience, we describe a general pipeline and necessary steps that may help researchers willing to participate in such competition. We partition the problem into two separate tasks. Firstly, we present state of the art for relevant neural networks concerning accuracy and computation time trade-off. Further, we create a survey of major techniques that leads to accuracy improvement. Namely, we recall image augmentation techniques, demonstrate the impact of various optimizers, and discuss ensemble techniques. The pipeline and techniques reflect our experience with a competition, in which we were able to reach a highly competitive solution and ended in fourth place. The uniqueness of our solution is that we used only free Google Colab computation service and still overperformed many more computation extensive approaches.**

***Index Terms*—object detection, image classification, neural networks, YOLO, optimizers**

I. Problem statement

As neural networks (NNs) become state of the art (SOTA) in image processing, they also spread into the industry and everyday life. They are used for many purposes, such as detecting fake videos, superresolution, optical quality recognition, automatic segmentation of satellite images, or autonomous driving. The spread is also invoked by the growing community, including community-competition websites such as Kaggle, Signate, or DrivenData. These websites accept commercial proposals for competitions and offer them to community participants to compete and win an award. Such competitions are essential as a supplement to a scientific part because this environment creates fair conditions for all solutions, where results have to be reproducible and resistant against scams and cheating.

In this overview paper, we will refer to Signate 'Tobacco detection and classification' competition (https://signate.jp/competitions/159), where we finally ended fourth of more than one thousand registered competitors. The aim of the competition and its scheme was as follows. The competition schedule was divided into two phases, where the first one served as a development phase, i.e., participants were allowed to submit up to three submissions per day to see the score. Then, each participant had to select only one submission uploaded in the first phase to be used in the second one. The second phase, realized after the first one, then determined the final leaderboard using new (unseen) data over the selected submission. Such a scheme prevents data overfitting, manual labeling, or cheating. For the first phase, there were 178 shelf images available for training and 55 images available for testing (for submitting the predictions). Other 54 shelf images were used for the second phase. Each shelf image contained approximately 30 – 200 cigarette boxes that had to be detected and classified. In total, the training dataset included approximately 26000 cigarette boxes, which had to be detected in the shelf images and correctly classified. The classification was aimed to link each single cigarette box with one of the 223 predefined classes.

Fig. 1. The image illustrates the task: an image containing a shelf with boxes is captured, and all the boxes are detected and recognized by an algorithm. The illustrative photo without the detections is taken from github.com/gulvarol/grocerydataset.

In the following sections, we describe the full pipeline leading to our solution. Such an overview helps develop neural network applications, and we recall here the most important and recent techniques in the area. Our contributions are as follows:

- We present a full pipeline of how a real task/competition

should be solved.
- We make an overview of object detection techniques and describe image cropping as a useful preprocessing technique.
- We present benchmark results of various neural networks for an image classification task.
- We collect and describe the main techniques for additional accuracy improvement, including methods for data augmentation, ensembling techniques, or proper selection of an optimizer.

II. The general pipeline

Here, we present the general pipeline, which we split into two parts: object detection and object classification. Such partition allows us to use neural networks that are specialized for the tasks.

A. Object detection

The first task to solve is object detection that relies on determining bounding box coordinates of all objects (tobacco boxes in our case) in an input image. If we omit the standard approaches represented by hand-crafted features such as HOG [1] or Viola-Jones cascades [2] and focus on NN-based approaches, the origin is given by the sliding window approach [3]. Here, multiple overlapping crops are extracted and classified by a NN. The complication is in the proper selection of parameters of the crop size and the degree of overlapping. Also, the whole process suffers from computation speed.

The sliding window can be replaced by an algorithm generating proposal regions, as it is used in the R-CNN approach [4]. Further improvement relies on integrating the region proposal into NN architecture in Faster R-CNN [5]. Because such a solution has still its functionality divided into proposing regions and regions classification, such class of detectors is called two-stage object detection. The current generation of NN focuses on single-stage detectors, where one pass through a single architecture is able to detect the boxes. The major schemes are SSD [6], YOLO [7], RetinaNet [8], and EfficientDet [9]. Based on various comparisons [10], a trade-off between speed and accuracy has to be considered. Generally, two-stage detectors are able to reach higher accuracy, but one-stage detectors have much higher processing speed. Here, YOLOv3 [11] excels as it is able to make an inference in real-time.

Our goal in the competition was to reach the best possible result (according to the competition's measure) on free hardware, accessible for everyone through Google Colab service[1]. Therefore, we have selected YOLOv3, which is one of the fastest detectors, notwithstanding it is not precise as, e.g., RetinaNet. The principle of YOLO is as follows. The training data are firstly pre-processed by k-means [12], which extracts the most prototypical widths and heights of the bounding boxes of detected objects. In detail, YOLOv3 uses nine prototypical boxes. That is apriori information that helps the NN to be trained faster. Then, a batch of images is passed through the backbone (given by Darknet53), which extracts features that are projected into three output scales. Each of the scales consists of a particularly sized grid (13×13, 26×26, or 52×52 cells) where each cell of the grid can detect up to three boxes. For details, see the YOLO architecture visualized in [11]. The advantage of YOLO is that one image is connected with the labels of all objects inside it and trained for them in one step; that is, in contrast with RetinaNet, where one image is connected with the labels separately one-by-one, and multiple steps are necessary, so YOLO training is significantly faster.

The important fact is that YOLO's input resolution is 416×416 natively. Downscaling the images into the lower resolution means the boxes can shrink so much that one side size is below 10 px. Then, their detection will be highly imprecise. This issue can be solved by increasing the input resolution, but this also increases the size of the symbolic tensor used in the learning phase. Such change results in a huge graphic card memory requirement that is not available on Google Colab. Therefore, we proposed to keep the low input resolution, but the training will be realized over image crops. The crops have to fulfill the following:

1) A crop size should be bigger than the biggest object in an image. Otherwise, it would not be possible to make an efficient ensemble of predictions.
2) A crop should include boxes bigger than 13×13 px, which is the coarsest resolution of the YOLO grid. Otherwise, the labels of boxes can overwrite each other.

The disadvantage of cropping is the increase of inference time. Originally, an image is downscaled, and prediction is made in time t. Here, n crops are created and inferred in time of tn. The linear complexity can be further improved by batch inferencing. For a general neural network, the additional speedup varies between approximately 1.5×–10×, for details see [13]. The benefit is a significant improvement in accuracy. In our case, we reached a score of 0.7829 with a vanilla solution and a score of 0.9893 with the cropping. The score has been reached by detecting all the overlapping crops, matching them based on overlapping, and selecting the best ones using non-maxima suppression. More details are described in Sections III-C and IV.

B. Object classification

YOLO is able to classify objects like other detectors. However, the competition dataset includes 224 classes, which creates big output tensors and wastes the network capacity needed for the correct detection. Therefore, we propose to split the task and classify the detected boxes with a separate neural network. Generally, we can split neural network architectures for image classification into two groups: fully connected and convolutional. Fully connected are historically older, tend to overfit more than convolutional networks, and usually reach lower accuracy. Therefore, we will focus on convolutional-based architectures only. For completeness, these architectures

[1] https://colab.research.google.com/

can be combined with functionalities based on fully connected layers, such as squeeze-excitation [14].

Before a classifier can be trained, a dataset has to be analyzed, and a loss function has to be selected. In our case, we reported a high imbalance between classes. That leads to more frequent updates of the gradient for the classes with more samples and finally results in overfitting. For such cases, it is beneficial to use Focal loss [8] instead of commonly used categorical cross-entropy. It has two advantages. First, it is able to work with imbalanced datasets, because well-classified samples have dynamically reduced loss. Second, it is able to avoid over-confident learning [15].

Here, we will use Keras[2], a high level framework for neural networks. There are several SOTA architectures in Keras "Applications" library[3], like ResNet [16], DenseNet [17], InceptionResnetv2 [18], XCeption [19], MobileNet [20], or NasNet [21]. Because our data for classification are cropped cigarette boxes with a reasonable resolution of 204×272 (we create all the crops with the same size), and because 26000 training samples are available, even the bigger networks can be used. The proper way is to create preliminary training with a bigger validation split (0.5) and choose the best network for the final training. The results are shown in Table I, and we have selected XCeption for the final training.

TABLE I
THE TABLE SHOWS THE ACCURACY OF SOTA NNS FOR THE 224-CLASS DATASET OF 26000 IMAGES WITH 0.5 VALIDATION SPLIT.

Network	Accuracy [%]	Epoch time [s]
ResNetV2 101	79.92	585
InceptionResnetV2	81.25	402
XCeption	81.23	230
Nasnet (mobile)	78.42	284
DenseNet 121	80.80	466
EfficientNet B3	80.99	296
MobileNetV2	80.60	230

Based on our experience, datasets always include incorrect labels that may confuse a neural network. One possible technique to deal with this problem is label smoothing [22], that modifies the label y_i into $y_i = (1-\alpha) * y_i + \alpha/K$, where K is a number of classes and α is smoothing factor. Unfortunately, this technique works reasonably only for small K. As our $K = 224$, high smoothing factor such as $\alpha = 0.5$ leads to the label 0.5 for true class and 0.002 for other classes; thus, the true label is decreasing, and the other labels stay almost unchanged. Therefore, we used the second technique: negative mining. A classifier was trained, used for classification the non-augmented training set, and, finally, the samples with the incorrect classification were manually inspected. We observed that most of the incorrect classifications are due to an incorrect label. The solution is to fix the incorrect labels manually and fine-tune the classifier.

[2]https://keras.io/

[3]https://keras.io/applications/

III. METHODS IMPROVING THE ACCURACY

Selecting a proper neural network architecture for a particular task is only beginning. A system for solving tasks has to be designed globally, so adequate augmentation routine, proper optimizer, and/or ensembling scheme have to be involved too. In this section, we describe the techniques in detail.

A. Methods preventing overfitting

Let us recall that neural networks are trained mainly in a data-driven way. Therefore, it is helpful to use regularization techniques that prevent overfitting. Generally, we can partition the techniques used to prevent overfitting into the four following groups: methods that augment data; methods that modify an NN architecture; methods adjusting kernel weights; schemes for NN training.

The first group, data augmentation, can be logically divided into spatial augmentation and intensity augmentation, where the latter consists of statistic-based approaches and visual-based approaches. The spatial augmentation includes flips among vertical/diagonal axis, rotations, shifts, zooms, or elastic transformations. The current technique is also cropping an image into small non-overlapping parts and their shuffling, which is useful mainly in the pixel-level semantic segmentation. Statistic-based intensity approach utilizes data centering to zero value, normalization by standard deviation, or ZCA whitening [23]. Finally, visual intensity augmentations realize (non)linear intensity shifts, blurrings/sharpenings, converting to grayscale, thresholding, or adding JPEG artifacts. For the image classification task, technique Cutout [24], [25] that replaces an arbitrary part of an image by a random value(s) can postpone a NN overfit and increase its accuracy.

The second group includes chiefly Dropout [26] and DropConnect [27]. These techniques remove random neurons/connections during training according to the defined ratio. That forces a neural network to focus on general features. Unfortunately, these techniques are reasonable primarily for fully connected networks.

The major techniques in the third group are kernel regularization and stochastic weight averaging. The former technique uses for a regression problem L_1 or L_2 regularization, which removes the significant values (outliers) in kernels and, therefore, more general features can be extracted [28]. Stochastic Weight Averaging (SWA) [29] helps to find a better local minimum in a space given by the loss function. Because neural network training is a process involving random numbers, there is no guarantee that the training will reach the global minimum, which represents the best solution. In the case of SWA, a neural network is trained regularly, and when a loss value starts to oscillate, SWA begins to save checkpoints (weights) after each epoch. The checkpoints are then averaged. This is similar to a standard ensembling with the difference that the ensembling here is in the sense of kernel weights and not models.

The last group, schemes for training neural networks, includes the techniques of early stopping and batch learning. The former watches validation loss: when it is not decreasing,

training is stopped. The latter smooths the surface of the loss function, which forces a network to generalize [30]. Notwithstanding learning rate warm-up [31] and cyclical learning rate [32] are techniques that do not primarily reduce overfitting, we can mark them also as schemes for training neural networks. The former one skips the instability of the training process by continual increasing of learning rate. The cyclical process reduces the learning rate until a threshold is reached and then increases it rapidly, and the whole process is repeated. Such the learning rate restarting can be viewed as an alternative to weight initialization techniques: weights from the previously-trained cycle are used for further training. Both methods are able to increase accuracy significantly.

B. Optimizers

The most important part of the neural network implementation is the process of training itself – how the NN weights in convolutional kernels are adjusted during the process. The base of the approach is the usage of the well-known scheme – backpropagation – which updates the weights during the learning process. However, there are several approaches used to decide when, how, and how significantly the weights should be updated during the training. The tools making such decisions on the basis of gradient computation are called optimizers.

Generally, optimizer's task is to optimize (minimalize) the value of a loss function by updating the network's internal parameters – the weights and biases. As the point is to optimize the loss function and its course during the training, the optimizers use loss function derivatives:

- First derivative-based optimizers are evaluating the first derivative depending on network parameters, gradient. As the shape of gradient is used to find the global minimum of the loss function according to the parameters, such an approach is called gradient Descent.
- Second derivative-based optimizers are aimed at the second derivative of the loss function. Such derivatives have much higher time complexity to evaluate but are more prone to staying in the local minimum and behaving better. However, due to calculation requirements, they are typically not preferred.

As commonly used, we will aim at the Gradient Descend technique. The main idea is to calculate the gradient, which is the Jacobian matrix containing first-order derivatives for all network parameters of the loss function. Using this matrix, the "direction" of the highest slope is determined. Then, the adjustment of the parameters is made according to the gradient concerning this steepest slope. Moreover, a learning rate (a multiplier), affects the size of the adjustment.

The basic technique (Stochastic Gradient Descent – SGD) has three main weak points: a need to set the learning rate for a sound value, a fact that the learning rate is shared for all parameters, and a weak behavior in the case when the gradient is indecisive. Therefore, there are many improvements done in comparison to SGD. First of them is an addition of a momentum [33], which forces the changes to be consistent with the previous ones. Another extension to this approach is Nesterov Accelerated Gradient (NAG), which tries to estimate the next step to avoid the minimum miss. Another notable improvement is the idea of the learning rate definition per parameter instead of one value for the whole model, which allows specific parameter stress; this approach was implemented in Adagrad [34] technique. As the learning rate in Adagrad is always decreasing, another extension, AdaDelta [35] was introduced, adjusting the learning rate with respect to history given by a floating window.

Another major update is applied in Adam [36] optimizer. It applies the idea that not only learning rates but also momentum is changed for each parameter separately. Another two extension, NAdam [37] and RAdam [38] were evolved from this technique. Slightly other approach is applied in RMSProp [39] technique, which takes the gradient direction only and couples it together with learning rate specification for each parameter, extended for batch processing.

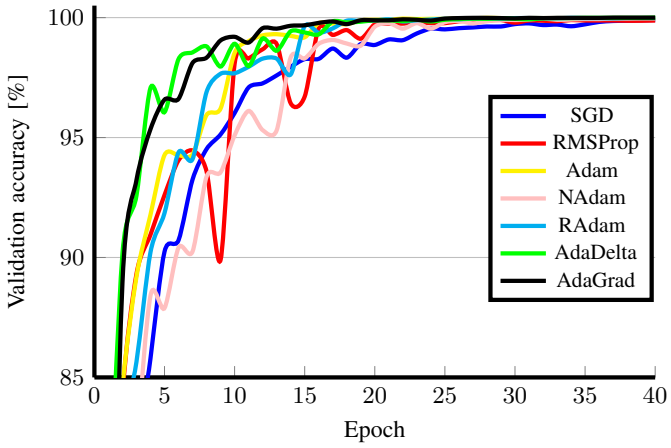


Fig. 2. Comparison of the selected optimizers on the image classification task.

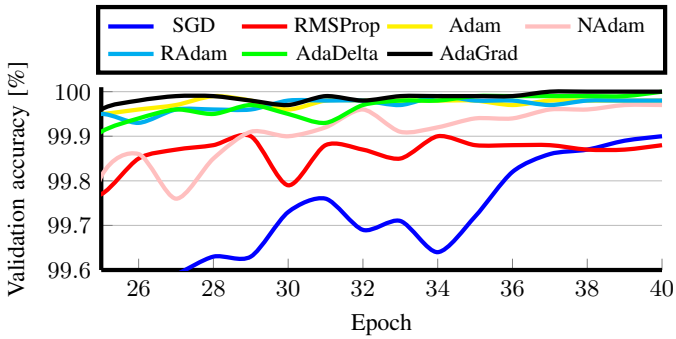


Fig. 3. Zoomed-in comparison of the selected optimizers on the image classification task for the last 15 epochs.

To compare optimizers, we took the XCeption network and a cleaned dataset from Section II-B. We set the validation split to 0.2 and the number of epochs to 40. Figure 2 shows the results for all the epochs, and Figure 3 presents the last

15 epochs in detail. The interpretation of the results is as follows: AdaDelta and AdaGrad converge fastest, and both are able to reach the maximum accuracy of 100 %. Adam-based optimizers converge slower as they use the momentum technique. RMSProp converges fast until it over-jumps the local minimum, which leads to a significant decrease in accuracy. After a learning rate is decreased, the process is repeated in a cycle. SGD is the most stable optimizer, but it converges slowly, and it is not able to reach the highest accuracy. Based on the inspection, we have selected AdaGrad as the best optimizer for the task. Let us note that such an inspection has to be realized for another problem because the finding above is not generally valid for all possible tasks.

C. Ensembling techniques

We consider a network that predicts labels $\mathbf{y_1}$, where the recognized class is determined as argmax($\mathbf{y_1}$). Because of the stochasticity of the learning process, we can observe that the same model trained for the second time gives $\mathbf{y_2}$ where $\mathbf{y_1} \neq \mathbf{y_2}$ and also it is possible that argmax($\mathbf{y_1}$) $\neq$ argmax($\mathbf{y_2}$). Therefore, it is beneficial to create $\mathbf{y}' = \sum_{i=1}^{n} \mathbf{y_i}$ for n models and compute the final classification as argmax($\mathbf{y}'$). Such the ensemble of prediction will lead to an increase in the accuracy if the models are not correlated. There are two main ways how to reach predictions with low correlation. The first way is to use different architectures with a distinct number of parameters, distinct augmentation, various optimizers, etc. The second way is to split data into k folds, train models with the same architecture over particular fold's data, and ensemble them. Because the distribution of classes in the folds may not be equal to the whole dataset's distribution, it is beneficial to use stratified folding [40], which ensures this property. The two ways can also be combined.

The alternative way to ensembling is test time augmentation (TTA) that requires only one model. In TTA, a sample for the classification is augmented n-times, classified n times, and the prediction vectors are aggregated. Obviously, the augmentations used for TTA have to be from the set of augmentations used during training. The same principle of ensembling, as is described for a classifier, is also applicable to a detector. The only difference is that coordinates of detected boxes (or widths/heights) are either weighted according to model confidence, nor a (soft)non-maxima suppression [41] is applied to select the boxes with the highest confidence.

IV. Performance of the proposed pipeline in the competition

Here, we are going to demonstrate the impact of our pipeline and recommendations on the competition[4] score, which was evaluated as follows. Overlaps of detected boxes and labels were computed in the manner of the intersect over union (IOU) coefficient. If a particular box reached IOU bellow 0.5, it was marked as undetected. As a result, it is unnecessary to use as a precise detector as possible (e.g., RetinaNet or EfficientDet), but prefer a detector that will not miss boxes. The predicted and true labels for classes were compared, and based on it, the F_1 score was computed from the detected boxes.

For the evaluation, our baseline submission consisting of the YOLO detector and the MobileNet classifier reached F_1 score of 0.7829. Such a score was reached for non-overlapping crops for the YOLO detector. By modification into overlapping boxes and ensembling the overlapping boxes by non-maxima suppression, the score was improved to 0.9133. Furthermore, we reached 0.9417 by training the networks for more epochs and 0.9519 by selecting the proper optimizer, as we demonstrate in Section III-B. Because here is essential to find the best balance between precision and recall, we elaborated with both YOLO and the classifier confidence thresholds and reached score 0.9692 for the optimal values. Then, we performed an analysis of the optimal classifier, as is shown in Section II-B, where the proper selection led to 0.97314. By visual analysis of the data, we set hyperparameters of augmentations to reflect precisely the data distribution. Also, we involved the CutOut technique into the classifier augmentation. This improved augmentations postponed overfitting and increased the score to 0.9843. By additional training of the models with cyclic learning rate, the score of 0.9887 was achieved. The last improvement to 0.9893 was reached by negative mining.

As can be seen, our first solution based on SOTA techniques reached 0.7829, which was significantly improved to the final score of 0.9893 due to data analysis, problem understanding, and application of hacks. Because we have not used huge models, we have reached low overfitting and ended the second round of the competition at the fourth place with a score of 0.9824, where we were almost on par with the third place, which yielded 0.9828. Because we used limited free cloud service, we did not involve ensembling techniques recommended in Section III-C, such as test time augmentation for the detector and model ensemble for the classifier. With them, we would possibly overperform even the first place that reached 0.9844.

V. Summary

The fast-growing popularity of neural networks, mainly due to commonly available development frameworks and easy-to-access cloud computing, causes the massive usage of neural network-based solutions in all fields. One of the ways how a company can easily find the solution for its task is the usage of the community by awarded competitions. In this paper, a solution for such competition aimed at object detection and classification was discussed. Based on your experience, we have presented a sequence of general steps that would improve the overall performance and should be considered. We have divided the problem into two tasks. The first one solves object detection and the second one classification of the detected objects. Furthermore, we discuss useful image augmentation techniques, optimizer selection importance, and ensemble techniques. All the presented findings reflected our experience with a public competition of object classification and detection, in which we were able to reach fourth place.

[4] https://signate.jp/competitions/159

Moreover, although using only free Google Colab computation service, we were able to beat other, more computation complex approaches.

ACKNOWLEDGMENT

The work is supported by ERDF/ESF "Centre for the development of Artificial Intelligence Methods for the Automotive Industry of the region" (No. CZ.02.1.01/0.0/0.0/17_049/0008414)

For more supplementary materials and an overview of our lab's work, please visit http://graphicwg.irafm.osu.cz/storage/pr/links.html.

REFERENCES

[1] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05)*, vol. 1. IEEE, 2005, pp. 886–893.

[2] C. Zhang, J. C. Platt, and P. A. Viola, "Multiple instance boosting for object detection," in *Advances in neural information processing systems*, 2006, pp. 1417–1424.

[3] T. Wang, D. J. Wu, A. Coates, and A. Y. Ng, "End-to-end text recognition with convolutional neural networks," in *Proceedings of the 21st International Conference on Pattern Recognition (ICPR2012)*. IEEE, 2012, pp. 3304–3308.

[4] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 580–587.

[5] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," in *Advances in neural information processing systems*, 2015, pp. 91–99.

[6] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "Ssd: Single shot multibox detector," in *European conference on computer vision*. Springer, 2016, pp. 21–37.

[7] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 779–788.

[8] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2980–2988.

[9] M. Tan, R. Pang, and Q. V. Le, "Efficientdet: Scalable and efficient object detection," *arXiv preprint arXiv:1911.09070*, 2019.

[10] L. Liu, W. Ouyang, X. Wang, P. Fieguth, J. Chen, X. Liu, and M. Pietikäinen, "Deep learning for generic object detection: A survey," *International Journal of Computer Vision*, vol. 128, no. 2, pp. 261–318, 2020.

[11] J. Redmon and A. Farhadi, "Yolov3: An incremental improvement," *arXiv preprint arXiv:1804.02767*, 2018.

[12] J. MacQueen *et al.*, "Some methods for classification and analysis of multivariate observations," in *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, vol. 1, no. 14. Oakland, CA, USA, 1967, pp. 281–297.

[13] S. Bianco, R. Cadene, L. Celona, and P. Napoletano, "Benchmark analysis of representative deep neural network architectures," *IEEE Access*, vol. 6, pp. 64 270–64 277, 2018.

[14] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7132–7141.

[15] J. Mukhoti, V. Kulharia, A. Sanyal, S. Golodetz, P. H. Torr, and P. K. Dokania, "Calibrating deep neural networks using focal loss," *arXiv preprint arXiv:2002.09437*, 2020.

[16] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[17] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 4700–4708.

[18] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. A. Alemi, "Inception-v4, inception-resnet and the impact of residual connections on learning," in *Thirty-first AAAI conference on artificial intelligence*, 2017.

[19] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1251–1258.

[20] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "Mobilenetv2: Inverted residuals and linear bottlenecks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 4510–4520.

[21] B. Zoph, V. Vasudevan, J. Shlens, and Q. V. Le, "Learning transferable architectures for scalable image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 8697–8710.

[22] R. Müller, S. Kornblith, and G. E. Hinton, "When does label smoothing help?" in *Advances in Neural Information Processing Systems*, 2019, pp. 4696–4705.

[23] A. Kessy, A. Lewin, and K. Strimmer, "Optimal whitening and decorrelation," *The American Statistician*, vol. 72, no. 4, pp. 309–314, 2018.

[24] T. DeVries and G. W. Taylor, "Improved regularization of convolutional neural networks with cutout," *arXiv preprint arXiv:1708.04552*, 2017.

[25] Z. Zhong, L. Zheng, G. Kang, S. Li, and Y. Yang, "Random erasing data augmentation," *arXiv preprint arXiv:1708.04896*, 2017.

[26] 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.

[27] L. Wan, M. Zeiler, S. Zhang, Y. Le Cun, and R. Fergus, "Regularization of neural networks using dropconnect," in *International conference on machine learning*, 2013, pp. 1058–1066.

[28] C. Cortes, M. Mohri, and A. Rostamizadeh, "L2 regularization for learning kernels," *arXiv preprint arXiv:1205.2653*, 2012.

[29] P. Izmailov, D. Podoprikhin, T. Garipov, D. Vetrov, and A. G. Wilson, "Averaging weights leads to wider optima and better generalization," *arXiv preprint arXiv:1803.05407*, 2018.

[30] E. Hoffer, I. Hubara, and D. Soudry, "Train longer, generalize better: closing the generalization gap in large batch training of neural networks," in *Advances in Neural Information Processing Systems*, 2017, pp. 1731–1741.

[31] P. Goyal, P. Dollár, R. Girshick, P. Noordhuis, L. Wesolowski, A. Kyrola, A. Tulloch, Y. Jia, and K. He, "Accurate, large minibatch sgd: Training imagenet in 1 hour," *arXiv preprint arXiv:1706.02677*, 2017.

[32] L. N. Smith, "Cyclical learning rates for training neural networks," in *2017 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 2017, pp. 464–472.

[33] N. Qian, "On the momentum term in gradient descent learning algorithms," *Neural networks*, vol. 12, no. 1, pp. 145–151, 1999.

[34] J. Duchi, E. Hazan, and Y. Singer, "Adaptive subgradient methods for online learning and stochastic optimization," *Journal of machine learning research*, vol. 12, no. Jul, pp. 2121–2159, 2011.

[35] M. D. Zeiler, "Adadelta: an adaptive learning rate method," *arXiv preprint arXiv:1212.5701*, 2012.

[36] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[37] T. Dozat, "Incorporating nesterov momentum into adam," 2016.

[38] L. Liu, H. Jiang, P. He, W. Chen, X. Liu, J. Gao, and J. Han, "On the variance of the adaptive learning rate and beyond," *arXiv preprint arXiv:1908.03265*, 2019.

[39] T. Tieleman and G. Hinton, "Divide the gradient by a running average of its recent magnitude. coursera: Neural networks for machine learning," *Technical Report.*, 2017.

[40] R. Kohavi *et al.*, "A study of cross-validation and bootstrap for accuracy estimation and model selection," in *Ijcai*, vol. 14, no. 2. Montreal, Canada, 1995, pp. 1137–1145.

[41] N. Bodla, B. Singh, R. Chellappa, and L. S. Davis, "Soft-nms–improving object detection with one line of code," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 5561–5569.

IEEE Third International Conference on Data Stream Mining & Processing
August 21-25, 2020, Lviv, Ukraine

Real-Time Automatic Video Inspection System for Piece Products Marking

Nonna Kulishova
Media Systems and Technologies Department
Kharkiv National University of Radioelectronics
Kharkiv, Ukraine
https://orcid.org/0000-0001-7921-3110

Anton Paramonov
Media Systems and Technologies Department
Kharkiv National University of Radioelectronics
Kharkiv, Ukraine
https://orcid.org/0000-0002-2124-064X

Volodymyr Tkachenko
Media Systems and Technologies Department
Kharkiv National University of Radioelectronics
Kharkiv, Ukraine
https://orcid.org/0000-0002-5076-0724

***Abstract*—The paper considers the problem of real time automated video inspection of piece products. The architecture of automatic video inspection system is proposed, the temporal characteristics of its action are investigated. The possibility of analyzing images of objects after transformations in grayscale and Lab space, the use of Hough transforms and second-order moments to frame object localization is considered. The possibility of using machine learning algorithms for real time automatic video inspection of industrial products was studied.**

Keywords—real-time image processing, automatic video inspection, image recognition

I. Introduction

The tasks of processing video streams are quite common in people's practice. One of the most important areas is the technical control of industrial products. In production, video control systems that carry out quality control of semi-finished products and finished products [1-3] became an integral part. Modern video cameras have sufficient rate to carry out non-destructive visual quality control of products in production lines without reducing their speed. Such systems are highly demanded in packaging production, when it is necessary to check the labeling quality on each production unit, and sort the products into two categories - good and defective.

Products are manufactured and marked on equipment, which rate is hundreds of thousands of pieces per hour. Therefore, visual quality inspection system should guarantee high accuracy (up to 2%) of product sorting and maintain high speed equipment. Known examples of such applications confirm high efficiency, low cost and a high degree of automation [4-8].

The high performance of visual inspection systems even allows using of machine learning and artificial intelligence for real-time recognition [9, 10]. However, it is obvious that such methods should be time-aligned with the hardware capabilities of the inspection systems. In this paper, we consider the problem of designing a real-time automatic industrial video control system. The controlled objects are markings on piece products that are manufactured using high-speed equipment.

The architecture of a video inspection system is proposed. Its operation is based on several image analysis algorithms. The system is a hardware-software complex in which the interaction of individual functional parts must be implemented within a single measurement cycle. The main goal of the work is to investigate the time intervals distribution between elements of a video inspection system within one measurement cycle to select image-processing algorithms that provide high speed and analysis accuracy of the entire system.

II. Architecture of The Real-Time Automatic Video Inspection System

Fig. 1 shows the system structure. The video monitoring system can include from one to six control points.

Each control point includes:

- flash and camcorder unit,
- conveyor activity sensors,
- image analysis system,
- device for separating defective products (ejector).

Objects under control enter the control point acting area from the left and move using the conveyor. A camera with flash mounted above the conveyor captures images of objects under control.

Images are transferred to the inspection module. Using conveyor activity sensors, the module associates images with the corresponding object.

The inspection module performs the main function of the system - controlled objects images analysis and deciding whether these objects fit all the requirements specific to the product. If the module rated the object as good, the object can move along the conveyor further to the exit from the system. If the module recognizes an object as defective, the ejector discards this object from the conveyor.

A feature of the product to be checked is its light weight, cylindrical shape; dimensions not exceeding 40 mm. To hold and move objects in the inspection area, it is necessary to use special devices in the form of conveyors. Individual elements of these devices inevitably come into view.

Therefore, image analysis includes procedures for the formation of several areas of control:

1) along the contour of the object for its positioning in the field of view of the camera;

2) along the contour of the checked image;

978-1-7281-3214-3/20/$31.00 ©2020 IEEE

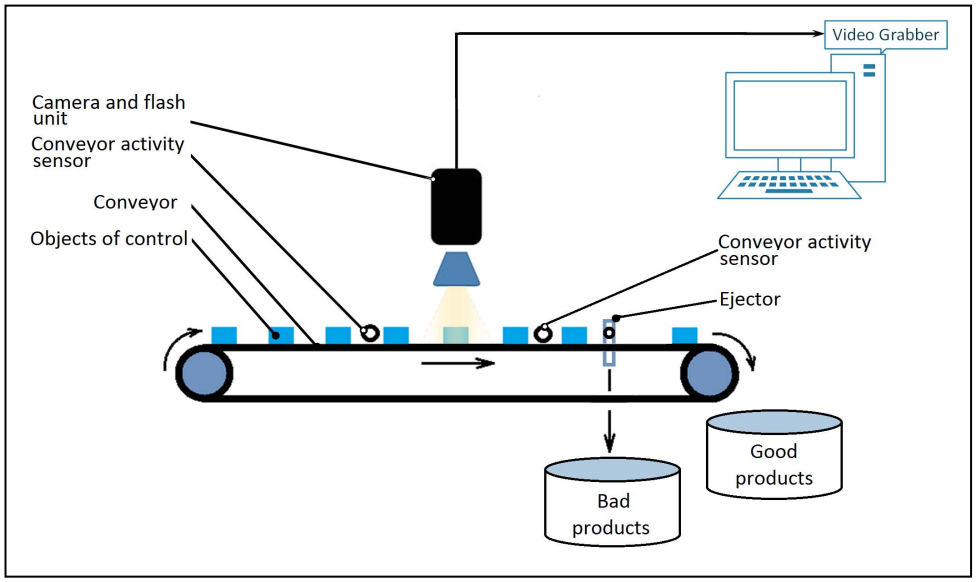


Fig. 1. Architecture of the real-time automatic video inspection system

3) along the contour of the elements excluded from the analysis (sealing rings, stiffeners, etc.).

To implement the control functions, it is proposed to distinguish two stages in the main actions of the system (Fig. 2). During the first stage, a template product sample is introduced into the system. This sample will be a reference for comparison with all products in series. Reference's description contains inspection areas and image characteristics (brightness, contrast) that are set up. This stage is carried out before the main products flow.

The second stage is the actual inspection, which is carried out in products flow with necessary rate.

At the first stage of reference description forming, the operations of brightness and contrast adjusting are quite important: they change the overall image brightness balance, which, in turn, determines the results of subsequent binarization. The system operator performs adjustment manually; settings are saved for inspection in product flow.

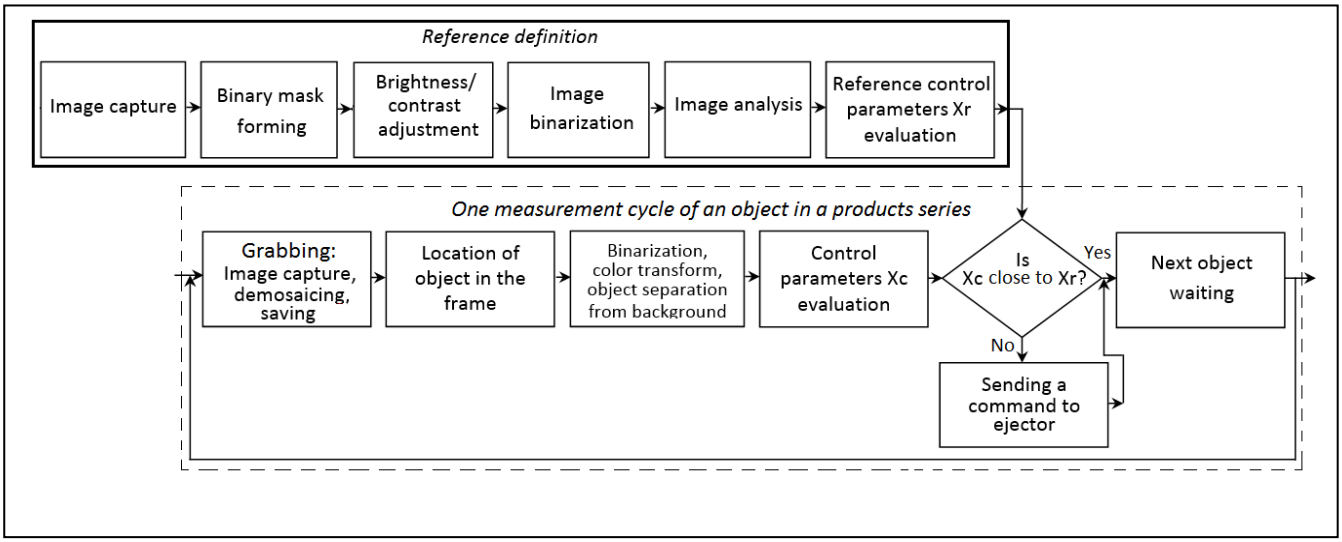


Fig. 2. The algorithm of the video inspection system

It is proposed to use the well-known Otsu algorithm [9, 10] for binarization of images. The binarization threshold should ensure the following conditions:

$$\sigma^2(thresh) = q_1(thresh)\sigma_1^2(thresh) + q_2(thresh)\sigma_2^2(thresh)$$

where

$$q_1(thresh) = \sum_{i=1}^{thresh} P(i) \;\&\; q_2(thresh) = \sum_{i=thresh+1}^{N} P(i);$$

$$\mu_1(thresh) = \sum_{i=1}^{thresh} \frac{iP(i)}{q_1(thresh)} \;\&\; \mu_2(thresh) = \sum_{i=thresh+1}^{N} \frac{iP(i)}{q_2(thresh)};$$

$$\sigma_1^2(thresh) = \sum_{i=1}^{thresh} [i - \mu_1(thresh)]^2 \frac{P(i)}{q_1(thresh)} \;\&$$

& $\sigma_2^2(thresh) = \sum_{i=1}^{thresh} \left[i - \mu_2(thresh) \right]^2 \frac{P(i)}{q_2(thresh)}$,

$P(i)$ - brightness of the *i*-th pixel, i = *1,2,...,N*; $q_1(thresh)$, $q_2(thresh)$ - probabilities, $\mu_1(thresh)$, $\mu_2(thresh)$ - mathematical expectations, and $\sigma_1^2(thresh)$, $\sigma_2^2(thresh)$ - variances of two classes of pixels separated by a threshold.

The binarization results are shown in Fig. 3.

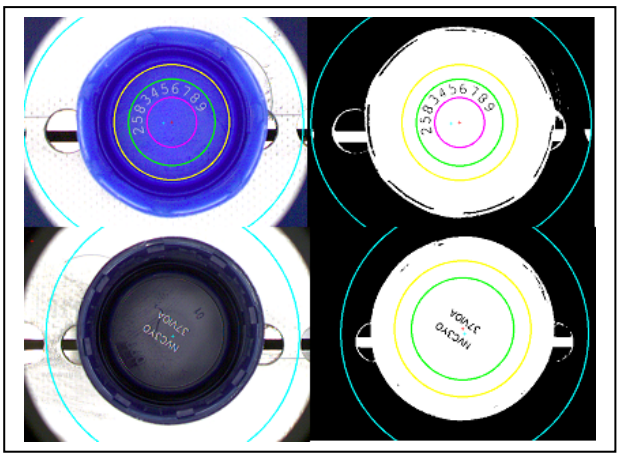


Fig. 3. Binarization of controlled objects images

This approach helps to compensate for conveyor belt uneven surface on which controlled objects are located, and to separate object from background.

Object separation from background could be performed by several approaches:

- by frame center coordinates (this approach is suitable, when object's positioning in the frame is stable enough);
- by 2-nd order moments (their calculating is standard operation in most programming libraries (for example, in Open CV) [11, 13];
- by Hough transform [11, 14].

If necessary, conversions to Lab, CMYK color spaces can be performed, when objects are hard to separate from the background. By default, the original image formed in RGB space is converted to Grayscale.

The use of masks formed in reference description allows limiting the inspection zone to the marking area. Controlled parameters $\mathbf{X}_C$ are calculated based on number of pixels displaying the marking N_M (Fig. 4).

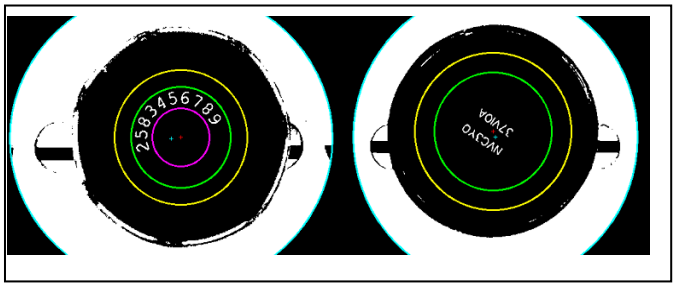


Fig. 4. Marking control zone after image transformations

During reference description creation, operator sets reference values of controlled parameters $\mathbf{X}_r = \{N_{Bj}\}$ and tolerances N_{Tj} (*j* – defect number). Thus, for suitable products, controlled parameter values must be close to reference values in sense ensuring the condition

$$N_{Mj} \in \left[\left(N_{Bj} - N_{Tj} \right), \left(N_{Bj} + N_{Tj} \right) \right]. \quad (1)$$

To control the fact of marking bias, the number of pixels is counted in areas bounded by other masks (color circles shown in Fig. 3 - 4). For this parameter, the fulfillment of conditions similar to condition (1) are also checked.

Thus, the system allows checking products for several indicators:

- presence of spots, defects in the material;
- marking too light or dark;
- missing or incomplete labeling.

The system solves the problem of recognition such defects on each object images in real time. For this purpose the function is used:

$$F\left(O(k)\right) = \sum_{j=1}^{4} f_j\left(O(k)\right) \quad (2)$$

where $O(k)$ - is the current object to be monitored at time *k*,

$$f_1\left(O(k)\right) = \begin{cases} 0, & if \;\; N_M\left(O(k)\right) \in \left[\left(N_{BM} - N_{TM} \right), \left(N_{BM} + N_{TM} \right) \right] \\ 1 & otherwise \end{cases}$$

$$f_2\left(O(k)\right) = \begin{cases} 0, & if \;\; N_{Mat}\left(O(k)\right) \in \left[\left(N_{BMat} - N_{TMat} \right), \left(N_{BMat} + N_{TMat} \right) \right] \\ 1 & otherwise \end{cases}$$

$$f_3\left(O(k)\right) = \begin{cases} 0, & if \;\; N_B\left(O(k)\right) \in \left[\left(N_{BB} - N_{TB} \right), \left(N_{BB} + N_{TB} \right) \right] \\ 1 & otherwise \end{cases}$$

$$f_4\left(O(k)\right) = \begin{cases} 0, & if \;\; N_F\left(O(k)\right) \in \left[\left(N_{BF} - N_{TF} \right), \left(N_{BF} + N_{TF} \right) \right] \\ 1 & otherwise \end{cases},$$

N_{Mat} - number of pixels displaying defects in the material, N_B - number of pixels to control marking brightness, N_F - number of pixels to control marking fullness, N_{BM}, N_{BMat}, N_{BB}, N_{BF} - basic reference values for these parameters, N_{TM}, N_{TMat}, N_{TB}, N_{TF} - corresponding tolerances.

Controlled objects sorting is carried out based on a decisive rule:

$$O(k) \in \begin{cases} \mathfrak{G} & if \;\; F\left(O(k)\right) = 0 \\ \mathfrak{B} & otherwise \end{cases} \quad (3)$$

where $\mathfrak{G}$ - a good objects set, $\mathfrak{B}$ - a defective objects set.

III. Experiment

The proposed video inspection system includes components:

- camera BASLER acA800-510uc, connect USB 3.0;
- objective Basler Lens C125-0818-5M F1.8 f8mm;

  flash Light ALB0804A-W00 / AN;
- computer Intel Core i3-8100 CPU @ 3.600 GHz×4, installed memory 8 GB;
- OS: Debian GNU / Linux 10 64-bit.

The video inspection system was located on the production line, where objects feed rate was up to 60 pieces per second, that is, one measurement cycle duration is 16667 μs. The error of sorting objects into good and defective was approximately 1.5%. Objects series was checked and certain operations duration was measured:

- image capture, demosaicing, conservation (grabbing);
- object localization in the control zone by frame center coordinates;
- object localization in the control zone by second order moments;
- localization of rounded objects on the conveyor belt to exclude them from the background by Hough transform;
- binarization;
- color conversion from RGB space to Grayscale;
- color conversion from RGB to Lab;
- counting the number of pixels in the control zones of marking and offset;
- calculating the value of the function, checking condition (3) and sending the command to ejector.

Measurement results are presented in Table 1.

Estimates of such operations duration were also carried out:

- object separation from background and binarization – approximately 10 μs;
- rounded objects localization using the Hough transform – 1800 ... 51779 μs;
- complete image analysis to calculate the control parameters values – 5419 ... 5911 μs.

The distributions of time intervals between operations within one measurement cycle are shown in Fig. 5.

TABLE I. DURATION OF BASIC OPERATIONS OF VIDEO INSPECTION SYSTEM ALGORITHM

Experiment number	Grabbing, μs	Object localization by moments, μs	RGB to Grey conversion, μs	RGB to Lab conversion, μs	Control parameters evaluation, μs
1	6131	204	124	667	238
2	6350	204	126	666	243
3	6171	203	126	668	245
4	6203	204	123	668	246
5	6233	204	123	657	240
6	6447	204	123	656	242
7	6223	204	123	657	240
8	5653	204	124	657	241
9	6147	204	124	667	236
10	4802	203	125	656	241
11	6161	204	124	656	241
12	5892	204	124	667	238

IV. Conclusions

Experiments showed that use of Hough transform, designed for higher accuracy of object localization in a frame at given measurement speed, is unjustified. Its maximum duration 51799 μs exceeds three measurement cycles in time (Fig. 5, bottom row); this is unacceptable in an industrial environment.

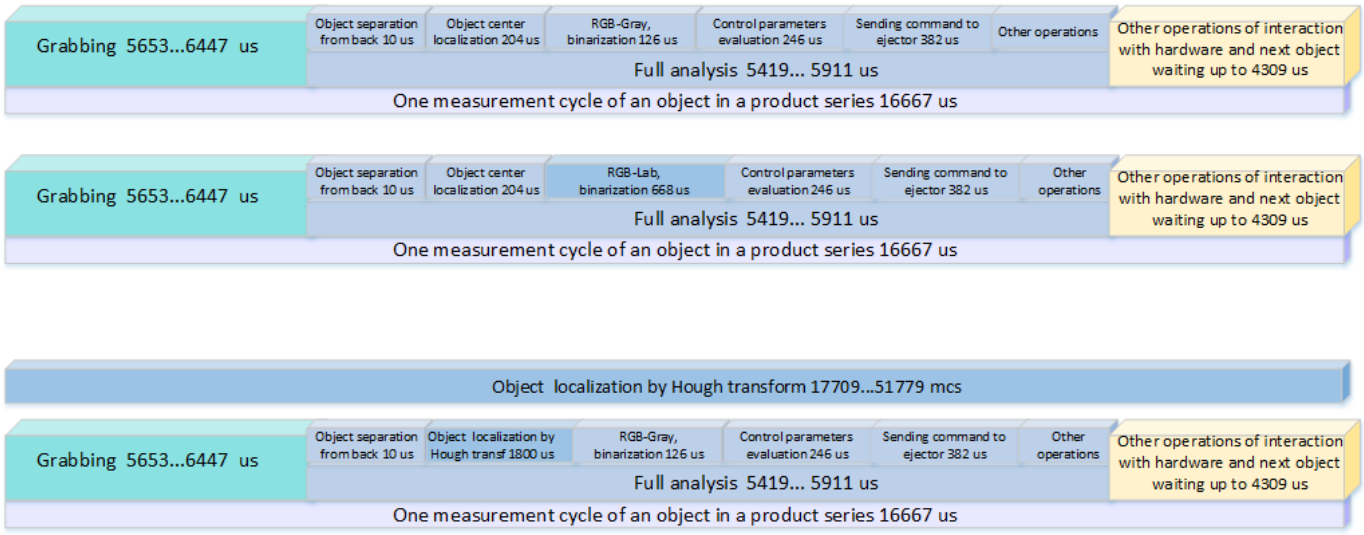


Fig. 5. Duration of basic operations of video inspection system algorithm during one measurement cycle

The use of the Lab color space for the analysis of complex color objects requires an additional about 500 μs of the total analysis duration.

If it is necessary to increase the level of sorting objects accuracy using the proposed system, it may be necessary to use machine-learning algorithms. Obviously, for a given speed for such algorithms, within a single measurement cycle, there remains a duration up to 9000-1000 μs.

The functional diagram of the video monitoring system and the approach to ensuring product quality control, presented in the work, allow sorting products according to specified indicators and ensure high equipment performance.

REFERENCES

[1] A. Mitra, Fundamentals of quality control and improvement, John Wiley & Sons, 2016.

[2] Dr.M.S. Nagmode, R.D. Komati and Aniket Khule, "A Review of Automated and Intelligent Visual Inspection Using Vision System," Computer Science, vol. 7, No. 3, 2015.

[3] N. M. Saad, N. A. Rahman, A. R. Abdullah, A. R. Syafeeza and N. S. M. Noor, "Quadratic Distance and Level Classifier for Product Quality Inspection System," Proc.of the Int. MultiConference of Engineers and Computer Scientists, vol. 1, 2017.

[4] S.-H. Huang and Y.-Ch. Pan, "Automated visual inspection in the semiconductor industry: A survey", Elsevier Computers in Industry,vol.66, p. 110, July 2014

[5] N. M. Duong, M. T. Chew, S. Demidenko,Q. H. Pham and D. K. Pham, "Vision Inspection System for Pharmaceuticals," Sensors Applications Symposium (SAS), IEEE, pp. 201 – 206, 2014

[6] G. Di Leo, C. Liguori, A. Paolillo et al, "Online visual inspection of defects in the assembly of electromechanical parts", Instrumentation and Measurement Technology Conference (I2MTC) Proceedings, Montevideo, IEEE, pp.407 – 411, 2014.

[7] S.H. Huang and Y.C. Pan, "Automated visual inspection in the semiconductor industry: a survey," Comput. Ind., 66, pp. 1–10, 2015.

[8] Q. Y. Li and S.W. Ren, "A real-time visual inspection system for discrete surface defects of rail heads", IEEE Trans. on Instrumentation and Measurement, vol. 61, no. 8, pp. 2189-2199, 2012.

[9] T. Zan, Zh. Su, M. Wang, X. Gao and D. Chen, "Statistical Process Control with Intelligence Based on the Deep Learning Model," Appl. Sci., 10, 308, 2020. doi:10.3390/app10010308

[10] Ye. Bodyanskiy, N. Kulishova and O. Chala, "The Extended Multidimensional Neo-Fuzzy System and its Fast Learning in Pattern Recognition Tasks," Data 3, 63 – 73, 2018. doi:10.3390/data3040063

[11] R.C. Gonzalez, R.E. Woods. Digital Image Processing, 4th edition. – Pearson, 2018, 1168 p.

[12] https://docs.opencv.org/master/d7/d1b/group__imgproc__misc.html#gae8a4a146d1ca78c626a53577199e9c57

[13] https://docs.opencv.org/master/d3/dc0/group__imgproc__shape.html#ga556a180f43cab22649c23ada36a8a139

[14] https://docs.opencv.org/master/d2/d15/group__cudaimgproc__hough.html

IEEE Third International Conference on Data Stream Mining & Processing
August 21-25, 2020, Lviv, Ukraine

The Steganographic Approach to Data Protection Using Arnold Algorithm and the Pixel-Value Differencing Method

Nataliia Kukharska
Department of Information Security Management
Lviv State University of Life Safety
Department of Information Systems and Technologies
Lviv Polytechnic National Universiy
Lviv, Ukraine
ORCID 0000-0002-0896-8361

Andrii Lagun
Department of Information Systems and Technologies
Lviv Polytechnic National Universiy
Lviv, Ukraine
ORCID 0000-0001-7856-9174

Orest Polotai
Department of Information Security Management
Lviv State University of Life Safety
Lviv, Ukraine
ORCID 0000-0003-4593-8601

***Abstract* – Steganographic data transformation is an effective means of ensuring the confidentiality and integrity of information resources, so developing the methods for improving the reliability and authenticity of steganographic systems is a promising research area. In the article we propose two approaches for embedding secret information into a BMP image by changing the difference between its pixel values. Both approaches involve the prior use (before embedding additional information) of Arnold's transformation to rearrange the pixels of the image. In the first approach, the Arnold transform is applied to the entire color matrix of the image, in the second - the image matrix is broken into blocks, and then twice do the Arnold transformation: to change the order of the blocks themselves and the sequence of pixels inside blocks. As a result, the pixels with the embedded information will be located in the image chaotically and evenly. The application of developed algorithms does not change the visual image quality, but complicates the fact of hidden information detection.**

Keywords — information security, computer steganography, digital image, BMP format, pixel-value difference, PVD method, Arnold transformation, scrambling, data hiding.

I. INTRODUCTION

Nowadays, modern information technologies and very effective computer's means open new opportunities to increase the volume and speed of information processing and transmission, facilitate the organization of remote access to global information resources. Because of it the need to development for reliable computer systems for defense the network data exists. One way to ensure the confidence and authenticity of information transmitted through open communication channels is using the computer steganography techniques.

The main advantage of steganographic methods for defending information is hiding the fact of transmitted messages. This is achieved by embedding them in digital data, which is usually analog in nature – images, videos, audio, text files, and even executable application files.

The main principles of computer steganography are:

1) Providing the authenticity and integrity of the file.
2) The adversary's knowledge about techniques of computer steganography.
3) Security is based on the keeping the basic properties of the transmitted file when entering a secret message and some unknown to an adversary information (the key) by steganographic transformation.
4) Extracting a secret message should be a complex computational task.

The theoretical foundations of modern methods of computer steganography have been researched in the works of such scientists as O.V. Ahranovskyi, O.D. Azarov, V.H. Hrybunin, V.K. Zadiraka, H.F. Konakhovych, S.V. Lenkov, I.I. Marakova, V.A. Mukhachev, I.H. Okov, D.O. Prohonov, O.Yu. Puzyrenko, I.V. Turyntsev, V.O. Khoroshko, O.A Smirnov, M. Ye. Shelest, Yu.Ye. Yaremchuk, C. Bergman, J. Davidson, J. Fridrich, M. Goljan, R. Liu, M.J. Medley, D.A. Pados, T. Tan and other.

Every year, the number of scientific publications on the issues of steganography and steganographic analysis increases. At the same time, despite significant results, a number of problems remain unresolved and poorly understood, including those related to improving the quality of the steganographic systems for information protection and their transmission reliability. Because of this, the improvement of existing and developing new robust steganographic algorithms for data transforming is one of the perspective directions of developing computer steganography.

II. OBJECT, PURPOSE AND TASKS OF THE RESEARCH

The research object is the process of information resources steganographic protection.

The main purpose of this research is solving the important scientific and technical problem of improving the quality of information steganographic protection systems. This can be achieved by using the selected element of the container and other algorithm parameters as a private key to increase the security of the message, the variability of algorithms for embedding information and improve the authenticity of its transmission.

In developing an approach to improve the reliability of the steganographic transforming process and the software for it implements, the following tasks should be solved:

978-1-7281-3214-3/20/$31.00 ©2020 IEEE

1) to determine the structure of the steganographic model;

2) to select necessary methods;

3) to develop an algorithm according to the chosen methods;

4) to write program code;

5) to develop a friendly (convenience) user interface.

In addition, each of the steps should be taken to check and test the developed software product and, as a result, to correct, to debug, to document this program code and provide recommendations to users.

III. THE MAIN PART

A. Image as steganographic container

A lot of the researches in digital steganography is devoted to embedding confidential messages into still digital images. It occurs because the digital images due to availability the large amount of redundant information make it possible to provide high bandwidth of hidden channel while maintain the integrity of the image perception. Among other reasons of increasing interest in graphic steganographic systems, there are numerous and diverging methods of image processing. It determines the existence of different possible methods and algorithms that can be applied to the transmission of hidden information. Another important reason for choosing an image as a steganographic container is the presence of the human visual system physiological features, namely the weak sensitivity of the human eye to slight changes in the image brightness.

In this article we consider the process of embedding confidential information in the spatial domain of BMP format digital images. In terms of steganography, the BMP format is the most advantageous graphic data format. The simple structure and large volume of BMP files make it possible to modify their contents without the need for decompression and therefore without the damaging of hidden information, which could occurs during the file compression.

B. Description of the classic Pixel-Value Differencing method

In this article we implement the steganographic transformations by the pixel difference method (Pixel-Value Differencing – PVD) [1, 2]. Unlike the popular steganographic method of least significant bit (LSB) [3-5] which has been early proposed due to its simplicity, the PVD method adapts the number of embedded bits to the grayscale/color changes in consecutive pixels. It enables the possibility the PVD method to provide enough a high value of the hidden bandwidth with maintaining high image quality [6, 7].

Let's consider the sequence of steps that implement the PVD method of embedding a confidential message into a digital image with an 8-bit gray scale. Algorithm of the method is based on the fact that human eyes can easily observe small changes in the gray values of smooth areas in the image but they cannot observe relatively larger changes at the edges areas.

The classic PVD method uses the modify brightness values of two adjacent pixels P_i and P_{i+1}, for which the absolute difference $d_i = |P_i - P_{i+1}|$ $d_i \epsilon [0,255]$ is calculated. The lower and upper limits $[lower_i, upper_i]$ of range R_i and number of bits $t = \lfloor \log_2(upper_i - lower_i + 1) \rfloor$, which can embedded in the pixels, are determined based on the obtained value of d_i in accordance with the table of quantization ranges. The message bits sequence of length t is converted to a decimal value t_d, after which is calculated a new difference value $d_i' = t_d + lower_i$. The brightness values of the pixels P_i and P_{i+1} are modified according to the formula [6]:

$$(P_i', P_{i+1}') = \begin{cases} \left(P_i + \left\lceil \frac{m}{2} \right\rceil, P_{i+1} - \left\lfloor \frac{m}{2} \right\rfloor\right), & if\ P_i \geq P_{i+1}, \& d_i' > d_i \\ \left(P_i - \left\lfloor \frac{m}{2} \right\rfloor, P_{i+1} + \left\lceil \frac{m}{2} \right\rceil\right), & if\ P_i < P_{i+1} \& d_i' > d_i \\ \left(P_i - \left\lfloor \frac{m}{2} \right\rfloor, P_{i+1} + \left\lceil \frac{m}{2} \right\rceil\right), & if\ P_i \geq P_{i+1} \& d_i' \leq d_i \\ \left(P_i + \left\lceil \frac{m}{2} \right\rceil, P_{i+1} - \left\lfloor \frac{m}{2} \right\rfloor\right), & if P_i > P_{i+1} \& d_i' \leq d_i \end{cases}$$

where $m_i = |d_i' - d_i|$, $\lfloor \cdot \rfloor$ – rounding to a smaller integer, $\lceil \cdot \rceil$ – rounding to a larger integer.

When the resulting values of pixel brightness are outside the accepted range [0,255], they corrected as follows [8]:

$$(P_i'', P_{i+1}'') = \begin{cases} (255, 255 - d - m), & if\ P_i' > 255; \\ (0, d + m), & if\ \ P_i' < 0; \\ (255 - d - m, 255), & if\ P_{i+1}' > 255; \\ (d + m, 0\ ), & if\ \ P_i' < 0. \end{cases}$$

Errors while extracting secret information from images can be avoided due to using the brightness a correcting procedure.

A typical setting of the ranges is that [0, 7], [8, 15], [16, 31], [32, 63], [64, 127] and [128, 255]. You can use several other variants of the quantization tables, which differ in the size of the ranges, and therefore the number of bits that can be embedded [8].

Fig. 1 shows an example of hiding information in two adjacent pixels, the brightness values of which are 50 and 65 and give an absolute difference 15. This value belongs to the range [8, 23] whose width is 23–8+1=16=2^4. The number of bits that can be placed in the pixels considered $t = \log_2 16 = \log_2 2^4 = 4$. From the sequence of the secret message bits we choose four consecutive bits – 1010, the decimal value of which is $t_d = 10$. We calculate the new value of the absolute difference in the pixels brightness $d' = 8 + 10 = 18$. Finally we get modified pixel brightness values using the formula above: 48 and 66.

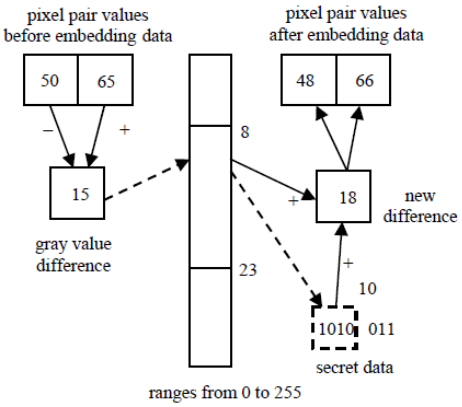


Fig. 1. Secret data embedding by PVD [1].

The majority of developments using PVD method implement embedding information into adjacent color matrix pixels [4, 6-8]. In this case, the pairs of pixels which use for embedding information follow one another that shown in Fig. 2.

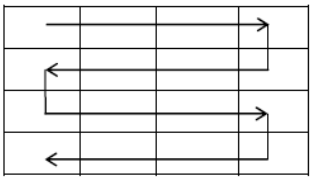


Fig. 2. PVD zigzag scan of an image [1].

To enhance the security of the hidden information, it would be advisable to place this information in the image not in sequentially but in pseudo randomly pixel pairs, since sequential posting allows the attacker to easily detect hidden information in the intercepted image by separating the bit sequence by PVD method.

C. *Arnold transformation*

To increase the robustness of steganographic transforms, we use the image scrambling technique [9]. Image scrambling techniques scramble the pixels of an image in such a manner that the image becomes chaotic and indistinguishable. These scrambling techniques generally use several keys and without the correct keys and an appropriate method the third party users cannot access the secret information even if they will intercept the medium.

In this article for image scrambling we use the simple but powerful Arnold transformation [10-11] which is periodic in nature and is very much popular in spatial domain applications:

$$\begin{pmatrix} x' \\ y' \end{pmatrix} = \begin{pmatrix} 2 & 1 \\ 1 & 1 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} (mod\ N),$$

where $x, y \in \{0, 1, \ldots . N-1\}$ and N is the size of a digital image.

The period of Arnold transformation depends on the size of the image to which it is applied. For example for 512x512 gray scale Lena image (Fig. 3) it is equal 384.

Fig. 3. Empty steganographic container Lena.bmp.

D. *The algorithm for steganographic hiding of secret information in the image using the PVD method and scrambling by Arnold transformation*

Now we consider the steps of the confidential information embedding algorithm by a PVD method in pixels of a immoveable image, which with application the Arnold transformation are chaotically disposed in the filled container.

Step 1. Firstly, we scramble the images using the Arnold transformation.

Step 2. We stop the converting image process by Arnold transformation at the some step for order to hide the message in the resulting modified image. The Arnold transformation series stop number during decoding process uses as the secret key. It is clear that during the converting process, pixels change their locations. The original image is visually distorted.

As example on Fig. 4 you can see Lena image (Fig. 3) after 90 times scrambled by Arnold Transforms.

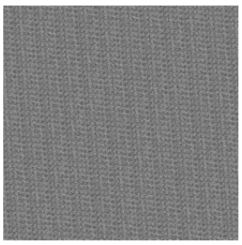


Fig. 4. Empty steganographic container Lena.bmp, scrambled by Arnold transformation.

We embed information in the transforming image sequentially, according to the standard simple scheme with modifying the brightness values of pixels pairs arranged in the scrambled image one by one (Fig. 2).

Step 3. We are implementing the rest (up to a full period) of Arnold transformation steps to get an image that will not visually different from the original, but will contain a hidden secret information.

E. *The algorithm for steganographic hiding of secret information in the image divided in blocks with using the PVD method and scrambling by Arnold transformation*

In this article we also consider another approach to using the PVD method, according to which the confidential information hides in the individual image blocks. Generally, blocks of sizes 3×3 [12-13] or 2×2 [14] were considered in researching.

We divide the original image into square blocks. The only requirement for blocks is their size: it must necessarily be a multiple of two, since we hide the secret information in them using brightness modification of arranged in pairs pixels. A digital container with a $V \times U$ dimension can hold up to $\frac{2VU}{m}$ pieces of information message, where m – is the number of elements in a single block.

Embedding message pieces in a container should be mostly uniform and equally likely, which is regarded by the uncomprehending person as transfer low quality images.

There are two possibilities to achieve that purpose.

1) To divide messages into pieces with their rearrangement and sequential embedding into container blocks.

2) To rearrange the image blocks of the container, followed by the sequential hiding of the message fragments.

If you have necessary to hide long messages and provide the ease of implementation of embedding algorithms at both the hardware and software levels, it is advisable to use the second option, which gives a possibility to create a variable-

structure code converter depending on key element of the container.

The code converter implements the rearrangement of container element numbers to embedding the message and also to extraction the message. An image with hidden information (stego) transmits through a communication channel that controls by the attacker. The main task of the attacker is to identify the information which was embedding in the intercepted digital object.

To the stability of the steganographic system critically is influencing by the choice of the steganographic container elements that are subjects to modification in the process of embedding information. In order to increase the reliability of the steganographic system, we propose to apply the Arnold transform twice in the first step of the information embedding algorithm described above. The first time it is used to rearrangement blocks in an image, and the second time – to rearrangement pixels within individual image blocks. Moreover, the number of Arnold transformation series (for blocks and for pixels) may be different. In this case, the steganographic system key is complex and consists on such three parameters:

- the size of the blocks to which the color matrix of images was broken;
- the step number to stop the Arnold transformation applied to the image when we operate with block coordinates;
- the step number to stop the Arnold transformation applied to individual image blocks when we use pixel coordinates.

Let's consider one of the rearrangement variants (Fig. 5) based on the Arnold transformation for blocks size 8×8 and theirs elements of immovable image (Fig. 3)

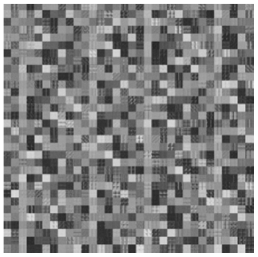


Fig. 5. Empty steganographic container Lena.bmp, scrambled by Arnold transformation.

IV. CONCLUSION

In this article we have developed the steganographic approaches to embed confidential information into BMP digital images, based on the PVD method and Arnold transformation which is used for image scrambling. Due to the equally likely distribution of the steganographic container blocks, equally likely distribution of the message elements and using keys, the security of the steganographic message is enhanced during its hidden transmission by open communication channels.

In case of using the steganographic algorithm for hiding information with PVD method in the image scrambled by Arnold Transform the power of the keys space depends on the size of the image container. It equals the Arnold transformation period value reduced by one. For the considered example Lena.bmp power of key space equals 383. In the case of information hiding by the PVD steganographic method in an image previously divided into blocks, with twice application of the Arnold transformation the power of the keys space we can count using formula: $\lfloor \log_2 V \rfloor \times (P_{im} - 1) \times (P_{bl} - 1)$, where P_{im} – is the period of the Arnold transformation, which uses the coordinates of matrix and blocks, and P_{bl} – is the period of the Arnold transformation, which uses a matrix with pixel brightness values of different image blocks. For the container Lena.bmp, divided into blocks of size 8 × 8, the power of the keys space is $\lfloor \log_2 512 \rfloor \times 47 \times 5 = 2\,115$. It is greater than the value in previous case, what allows making a conclusion that the level of protection of the secret message is improved.

The advantages of the considered approaches are:

1) high throughput;
2) high resistance to unauthorized access;
3) high resistance to frequency detection;
4) high resistance to destruction the least significant bits of container;
5) resistance to trimming edges.

REFERENCES

[1] D. C. Wu, and W. H. Tsai, "A steganographic method for images by pixel-value differencing," Pattern Recognition Letters, vol. 24, no. 9–10, 2003, pp. 1613–1626.

[2] El-Sayed M. El-Alfy, and Azzat A. Al-Sadi," Pixel-Value Differencing Steganography: Attacks and Improvements", ICCIT, 2012, pp. 757–762.

[3] B. Li, J. He, J. Huang, and Y. Q. Shi, "A survey on image steganography and steganalysis," Journal of Information Hiding and Multimedia Signal Processing, vol. 2, no. 2, 2011, pp. 142–172.

[4] C. K. Chan, and L. Cheng, "Hiding data in images by simple LSB substitution," Pattern Recognition, vol. 37, no. 3, 2004, pp. 469–474.

[5] H. C. Wu, N. I. Wu, C. S. Tsai, and M. S. Hwang, "Image steganographic scheme based on pixel-value differencing and LSB replacement methods," IEE Proceedings on Vision, Image and Signal Processing, vol. 152, 2005, pp. 611–615.

[6] Hsien-Wen Tseng, and Hui-Shih Leng, "A Steganographic Method Based on Pixel-Value Differencing and the Perfect Square Number", Journal of Applied Mathematics, vol. 2013, pp.1–8.

[7] Khalid A. Darabkh, Ahlam K. Al-Dhamari, and Iyad F. Jafar, "A New Steganographic Algorithm Based on Multi Directional PVD and Modifed LSB", Journal of Information Technology and Control, vol. 46, no. 1, 2017, pp. 16–36.

[8] A. V. Akhmametieva, and V. V. Kovalenko, "Development of the steganographic method of embedding of additional information into the spatial domain of color images", Informatics and Mathematical Methods in Simulation, vol. 8, no. 2, 2018, pp. 110–120.

[9] M. Mishra, P. Mishra, M. C. Adhikary, and S. Kumar, "Image encryption using Fibonacci-Lucas transformation", International Journal on Cryptography and Information Security, vol. 2, no. 3, 2012, pp. 131–141.

[10] V. I. Arnold, and A. Avez, Ergodic Problems in Classical Mechanics. New York: Benjamin, 1968.

[11] Z. G. Ma, and S. S. Qiu, "An image cryptosystem based on general cat map", Journal of China Institute of Communications, no. 24, 2003, pp. 51–57.

[12] A. Pradhan, K. Raja Sekhar, and G. Swain, "Digital image steganography based on seven way Pixel Value Differencing" Indian Journal of Science and Technology, vol. 9, 2016, pp. 1–11.

[13] O. Hosam, and N. Ben, "Halima adaptive block-based pixel value differencing steganography", Security Comm. Networks, no. 9, 2016, pp. 5036–5050.

[14] G. Swain, "A steganographic method combining LSB substitution and PVD in a block", International Conference on Computational Modeling and Security (CMS 2016), Procedia Computer Science, pp. 39–44.

IEEE Third International Conference on Data Stream Mining & Processing
August 21-25, 2020, Lviv, Ukraine

A New Approach to Image Enhancement by Non-Linear Contrast Stretching

Sergei Yelmanov
Special Design Office of Television Systems
Lviv, Ukraine
sergei.yelmanov@gmail.com

Yuriy Romanyshyn[1, 2]
[1]Lviv Polytechnic National University
[2]University of Warmia and Mazury
[1]Lviv, Ukraine, [2]Olsztyn, Poland
yuriy.romanyshyn1@gmail.com

***Abstract*— This paper looks at the challenge of creating new, effective ways to improve images that are intended for use in mobile applications. The purpose of this study is to improve the efficiency of enhancing images in a fully automatic mode by parameter-free image intensity transformation. This paper proposes a new approach to enhancing the image by adaptively transforming its intensity based on the analysis of brightness distribution at the boundaries of the objects of this image. To demonstrate the possibilities of this approach, new techniques of intensity transformation are proposed. The proposed techniques provide effective image enhancement without the appearance of distortion and artifacts and are focused on use in mobile applications.**

Keywords*— *image enhancement, intensity transformation, contrast stretching.

I. INTRODUCTION

The major trend in virtually all areas of human activity is the ever-increasing use of mobile applications based on the real-time analysis of video information. Intensive use of technologies of image analysis requires preliminary enhancing (pre-processing) the raw images [1, 2, 3].

The need for pre-processing of source images is due to some objective reasons, the main of which are the disadvantages of the observed scenes (e.q., uneven illumination), the unfavorable conditions of shooting, the not high enough technical characteristics of the image sensor, and others [1, 2]. The effectiveness of image processing and analysis at all stages depends crucially on the quality of the initial image [2, 3, 4].

Image processing in real-time applications has several specific distinguishing features that are very important. Techniques of image pre-processing in real-time must meet the following primary requirements, namely:

- there should be a high efficiency of enhancing images for different scenes, observed objects, and any possible conditions of observation;
- pre-processing should be carried out in a fully automatic mode, without any additional interactive settings;
- computational costs should be minimal to real-time implement pre-processing in mobile gadgets with low performance.

The challenge of creating new, effective technologies to improve images through their real-time pre-processing is now particularly urgent [1, 2]. A huge amount of research has been devoted to address the task of enhancing images, but its final decision is still a very long way off [5, 6]. The most of the existing methods have significant drawbacks that limit their use to improve images in mobile applications [2, 3, 4].

This work addresses the issue of improving images in mobile applications. The purpose of this study is to improve the efficiency of enhancing images in a fully automatic mode by parameter-free image intensity transformation. To meet this challenge in this paper a new approach to enhancing the image by adaptively transforming its intensity based on the analysis of brightness distribution at the boundaries of the objects of this image is proposed.

II. RELATED WORKS

The issue of improving images has always received particular attention, and this problem is the subject of a large number of works [1, 2, 3, 5, 6]. There are now different approaches to improving images among which the most popular and widely used are ways to transform images in the spatial domain, due to the simplicity of their implementation.

The methods based on the non-inertial statistical transformations of image intensity are most often used to enhance images in real-time applications. Intensity transformation is the simplest type of conversion in a spatial domain, has low computational costs, and is very simple to implement. Intensity transformation is the simplest type of conversions in a spatial domain, highly efficient and as simple to implement, has low computational costs, and therefore is very much in demand for improving images in real-time [2, 3, 5, 6].

The most well-known and generally accepted generalized description for the technique of intensity transformation is described as [4]:

$$r = r_{\text{low}} + \left(r_{\text{upp}} - r_{\text{low}}\right) \cdot T(b), \tag{1}$$

$$0 \le r_{\text{low}} < r_{\text{upp}} \le 1, \text{ and } 0 \le b_{\min} \le b \le b_{\max} \le 1, \tag{2}$$

where b is the brightness for the current pixel in the source image B; $T(b)$ is the intensity transformation; r is the result of transforming in the current pixel; R is the transformed image; r_{low} and r_{upp} are the boundaries for the dynamic range $[r_{\text{low}}, r_{\text{upp}}]$ of possible brightness values for R.

$T(b)$ is a monotonically non-decreasing function:

$$\forall b_1, b_2 \in [b_{\min}, b_{\max}]: \text{ if } b_1 \le b_2 \Rightarrow T(b_1) \le T(b_2), \tag{3}$$

which, as assumed, is to satisfy the following conditions:

$$0 \le T(b) \le 1 \; \forall b \in [b_{\min}, b_{\max}], \; T(b_{\min}) = 0 \wedge T(b_{\max}) = 1. \tag{4}$$

978-1-7281-3214-3/20/$31.00 ©2020 IEEE

In doing so, it is most often assumed that $r_{low} = 0$ and $r_{upp} = 1$, and definition (1), respectively, takes the form:

$$r = \begin{cases} 0, & \text{if } b < b_{\min}, \\ T(b), & \text{if } b_{\min} \le b \le b_{\max}, \\ 1, & \text{if } b_{\max} < b. \end{cases} \quad (5)$$

Without loss of generality, taking into account (3), suppose that T(b) is an integral transformation that has the form:

$$T(b) = k \cdot \int_{b_{\min}}^{b} t(x)\,dx, \quad (6)$$

where $t(b)$ is the distribution density (the increment) for $T(b)$ $(t(b) \ge 0)$; k is scaling factor, which is determined from conditions (4):

$$k = \left[\int_{b_{\min}}^{b_{\max}} t(x)\,dx \right]^{-1}. \quad (7)$$

The most popular intensity transformations usually belong to three best-known groups [3, 4], namely to core transformations of intensity, piecewise linear stretching techniques, and histogram matching-based techniques.

The core transformations of the image intensity are widely known [2, 3, 4], and their capabilities and features are well studied. Their effectiveness depends on the distribution of brightness and the parameters of conversion, which are typically set interactively [4].

Known techniques of linear stretching [3, 5] are based on the assumption that the increment $t(b)$ is piecewise constant function (or step function) with finitely many pieces and can be presented as a finite linear combination of indicator functions of intervals of brightness range:

$$t(b) = \sum_{i=1}^{N} \alpha_i \chi_i(b) \; \forall b \in [0,1], \; N \ge 1, \quad (8)$$

where α_i are constants, real numbers; N is the finite number of sub-intervals of brightness scale; $\chi_i(b)$ are indicator functions, where:

$$\chi_i(b) = \begin{cases} 1, & if \; b \in S_i \\ 0, & if \; b \notin S_i \end{cases}, \; i = \overline{1,N} \quad (9)$$

where S_i is the i-th sub-interval (sub-range) of brightness scale ($1 \le i \le N$).

Suppose that $N = 3$ and the increment $t(b)$ has the form:

$$t(b) = \begin{cases} 0, & \text{if } b \le b_{low} \\ \dfrac{1}{b_{upp} - b_{low}}, & \text{if } b_{low} < b < b_{upp} \\ 0, & \text{if } b \ge b_{upp} \end{cases}, \quad (10)$$

where b_{low} and b_{upp} the boundaries for the interval $[b_{low}, b_{upp}]$ of dynamic range.

Based on (10), the transformation $T(b)$ (6) can be defined as:

$$r = T(b) = \begin{cases} 0, & \text{if } b \le b_{low} \\ \dfrac{b - b_{low}}{b_{upp} - b_{low}}, & \text{if } b_{low} < b < b_{upp} \\ 1, & \text{if } b \ge b_{upp} \end{cases}. \quad (11)$$

The best-known technique from this group is the min-max contrast stretching, in which $b_{low}=b_{min}$ and $b_{upp}= b_{max}$ [4].

Another well-known and popular technique is the percentile contrast stretching in which the values b_{low} and b_{upp} are determined from the conditions [3, 4]:

$$F(b_{low}) = \int_{0}^{b_{low}} f(x)dx = \varepsilon, \qquad F(b_{upp}) = \int_{0}^{b_{upp}} f(x)dx = 1 - \varepsilon. \quad (12)$$

where $f(b)$ is probability density function (pdf); $F(b)$ is cumulative distribution function (CDF).

At known techniques of piecewise-linear stretching [4, 5], the brightness range is subdivided into a higher number of intervals ($N > 3$). In the case when $N = 4$, the increment $t(b)$ can be presented as:

$$t(b) = \begin{cases} 0, & \text{if } b < b_{low} \\ \dfrac{\alpha_2}{b_{tr} - b_{low}}, & \text{if } b_{low} \le b \le b_{tr} \\ \dfrac{1 - \alpha_2}{b_{upp} - b_{tr}}, & \text{if } b_{tr} \le b \le b_{upp} \\ 0, & \text{if } b > b_{upp} \end{cases}, \quad (13)$$

where b_{tr} is the threshold value of brightness ($b_{low} < b_{tr} < b_{upp}$); α_2 is gain factor, parameter ($0 < \alpha_2 < 1$) [6].

For distribution (11), $T(b)$ is equal to :

$$T(b) = \begin{cases} 0, & \text{if } b < b_{low} \\ \alpha_2 \cdot \dfrac{b - b_{low}}{b_{tr} - b_{low}}, & \text{if } b_{low} \le b \le b_{tr} \\ \alpha_2 + (1 - \alpha_2) \cdot \dfrac{b - b_{tr}}{b_{upp} - b_{tr}}, & \text{if } b_{tr} \le b \le b_{upp} \\ 1, & \text{if } b > b_{upp} \end{cases}. \quad (14)$$

Most often, it is assumed that $b_{tr} = b_{mean}$ and $\alpha_2 = ½$ [6].

The critical issue in implement the piecewise linear stretching is to choosing boundaries of intervals and the values of constant α_i [1, 2].

To implement the piecewise linear stretching, we propose a new technique based on the analysis of the existing brightness values that are present in the image.

To that end, we propose to define the increment $t(b)$ as:

$$t(b) = \begin{cases} 0, & \text{if } f(b) = 0 \\ 1, & \text{if } f(b) > 0 \end{cases}. \quad (15)$$

In general, given the presence of noise in the image, the increment can be defined as:

$$t(b) = \begin{cases} 0, & \text{if } f(b) < \theta \\ 1, & \text{if } f(b) \ge \theta \end{cases}, \quad (16)$$

where θ is the threshold value.

Definitions (1), (6), and (16) are a description of the proposed technique of piecewise linear stretching.

The histogram matching (or mapping) is a statistical no-inertial transformation in which the distribution of brightness is converted into a given shape [7]. Equalization [4, 6, 7] is a particular case of the procedure of matching, in which the brightness distribution becomes uniform.

For histogram equalization we have that:

$$t(b) = f(b) \cdot \tag{17}$$

In this case (17), we can define the histogram equalization as [4]:

$$r = T(b) = \int_0^b t(x)\,dx = \int_0^b f(x)\,dx = \Pr\{B \le b\}, \tag{18}$$

where $\Pr\{\cdot\}$ is probability of an event.

Histogram equalization is a high efficiency, a simplicity to implement, and has a low computational cost. However, it has several disadvantages.

The well-known and very significant drawbacks of histogram equalization based techniques are a decrease in the contrast of objects with small sizes, an excessive increase in the contrast of extended objects, and, as a consequence of this, the possible appearance of unwanted artifacts and distortions in the image.

To address these disadvantages, the equalization of the clipped histogram (clipped HE) is applied [3, 4], where:

$$t(b) = \min(f(b), \theta) = \begin{cases} f(b), & if\ f(b) \le \theta \\ \theta, & if\ f(b) > \theta \end{cases}. \tag{19}$$

Another known technique [3, 4] is the power-law intensification of the histogram (power-law HE), for which:

$$t(b) = f(b)^{\gamma} / \int_0^1 f(x)^{\gamma}\,dx, \tag{20}$$

where γ is the exponent, parameter.

The challenge of developing new, effective technologies to improve images through non-linear intensity transformations is now particularly urgent.

III. Proposed Technique

This paper looks at the challenge of creating new, effective ways to improve images that are intended for use in mobile applications.

The purpose of this study is to improve the efficiency of enhancing images in a fully automatic mode by parameter-free image intensity transformation. To meet this challenge in this paper proposes a new approach to enhancing the image by adaptively transforming its intensity based on the analysis of brightness distribution at the boundaries of the objects of this image.

The proposed approach is based on the assumption that $T(b)$ is an integral transformation that has the form:

$$T(b) = \beta \cdot \int_{b_{\min}}^{b} \mu_t(b_{\min}, x) \cdot \mu_t(x, b_{\max})\,dx, \tag{21}$$

where $\mu_t(b_1, b_2)$ is the assessment of the density of increment $t(b)$ in the interval $[b_1, b_2]$ $(b_1 < b_2)$:

$$\mu_t(b_1, b_2) = \begin{cases} \int_{b_1}^{b_2} t(x)\,dx \Big/ \int_{b_1}^{b_2} dx, & \text{if } b_1 < b_2 \\ 0, & \text{otherwise} \end{cases}, \tag{22}$$

and β is the normalizing factor, which is defined as:

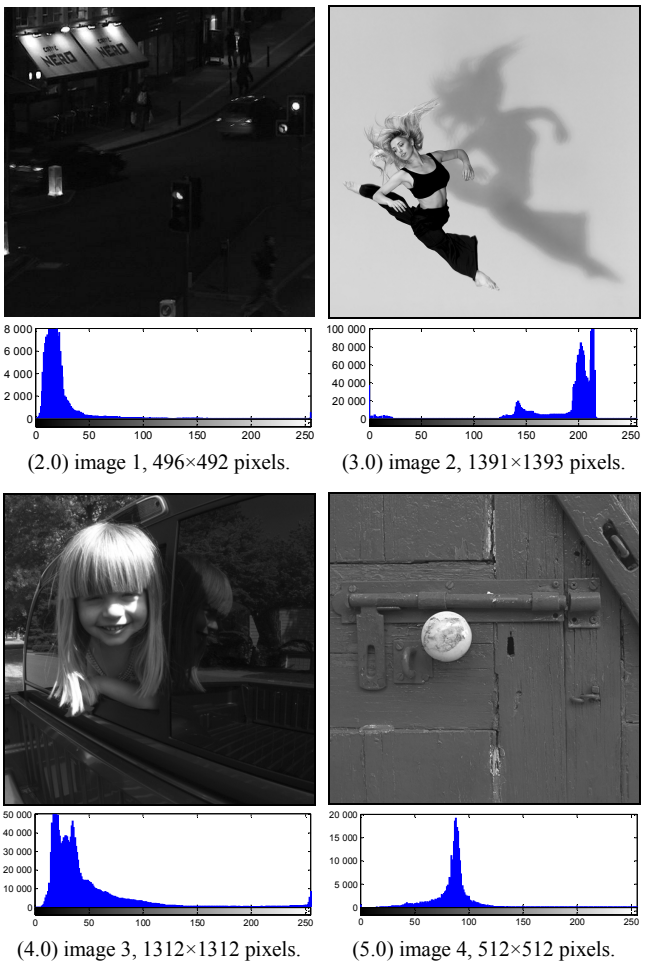
$$\beta = \left[\int_{b_{\min}}^{b_{\max}} \mu_t(b_{\min}, y) \cdot \mu_t(y, b_{\max})\,dy \right]^{-1}. \tag{23}$$

Definitions (21)-(23) describe the proposed approach to enhancing the image by adaptively transforming its intensity.

In the proposed approach (21)-(23), as the estimates of increment $t(b)$, may use the assessments (16)-(20).

IV. Research

This study is based on the results of assessing contrast for four groups of images using different metrics. Each from the four groups is formed by converting the corresponding source image using selected methods of processing it. Four raw source images and their histograms are shown in Fig. 1.



(2.0) image 1, 496×492 pixels. (3.0) image 2, 1391×1393 pixels.

(4.0) image 3, 1312×1312 pixels. (5.0) image 4, 512×512 pixels.

Fig. 1. The appearance of test images and their histograms.

To improve the images by converting their intensity, the known methods and proposed techniques were used:

(a) min-max stretching (11) [4, 5];

(b) percentage linear stretching (11), (12) [3, 4];

(c) proposed stretching for nonzero brightness (16);
(d) piecewise-linear stretching (14) [3, 4, 5];
(e) global histogram equalization (18) [6, 7] ;
(f) clipped histogram equalization (19) [3, 4];
(g) power-law histogram equalization (20) [3, 4];
(h) proposed technique (21)-(23) using (17);
(i) proposed technique (21)-(23) using (19);
(j) proposed technique (21)-(23) using (20).

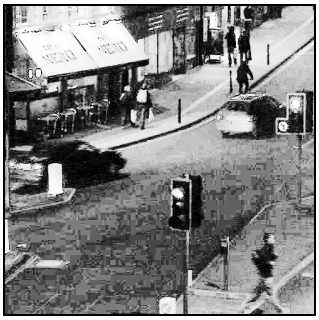
The results of processing the raw source images (Figure 1) are shown in Figures 2, 3, 4 and 5.

(2.1) min-max stretching.

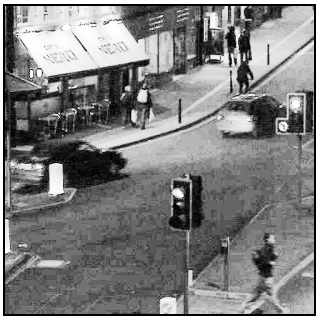
(2.2) percentile stretching.

(2.3) proposed stretching (16).

(2.4) piecewise linear stretching.



(2.5) global HE.



(2.6). clipped HE.

(2.7). power-law HE.

(2.8) proposed (21)-(23) using (17).

(2.9) proposed (21)-(23) using (19).

(2.10) proposed (21)-(23) using (20).

Fig. 2. Image processing results

(3.1) min-max stretching.

(3.2) percentile stretching.

(3.3) proposed stretching (16).

(3.4) piecewise linear stretching.

(3.5) global HE.

(3.6). clipped HE.

(3.7). power-law HE.

(3.8) proposed (21)-(23) using (17).

(3.9) proposed (21)-(23) using (19).

(3.10) proposed (21)-(23) using (20).

Fig. 3. Image processing results.

(4.9) proposed (21)-(23) using (19).

(4.10) proposed (21)-(23) using (20).

Fig. 4. Image processing results.

(4.1) min-max stretching.

(4.2) percentile stretching.

(4.3) proposed stretching (16).

(4.4) piecewise linear stretching.

(4.5) global HE.

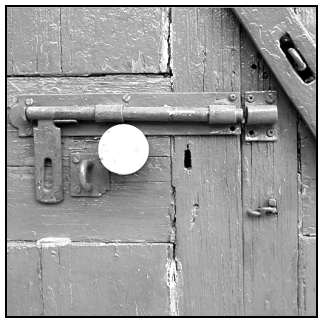
(4.6). clipped HE.

(4.7). power-law HE.

(4.8) proposed (21)-(23) using (17).

(5.1) min-max stretching.

(5.2) percentile stretching.



(5.3) proposed stretching (16).

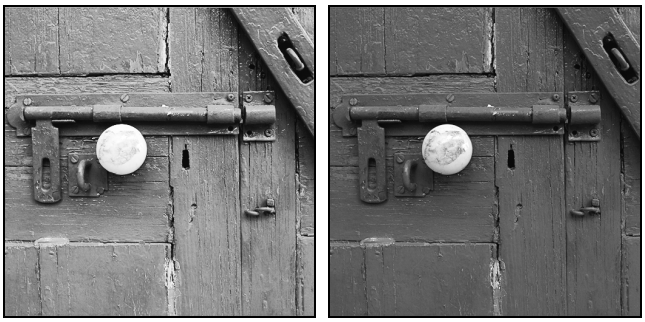
(5.4) piecewise linear stretching.

(5.5) global HE.

(5.6). clipped HE.

(5.7). power-law HE.

(5.8) proposed (21)-(23) for (17).



(5.9) proposed (21)-(23) for (19). (5.10) proposed (21)-(23) using (20).

Fig. 5. Image processing results.

To measure the global image contrast, the following metrics were used:

1) generalized contrast [8]:

$$C_{gen} = \int_0^1 \int_0^1 |x-y| f(x) f(y)\, dxdy\,, \qquad (24)$$

2) incomplete integral contrast [8]:

$$C_{inc} = \int_0^1 |x - b_{mean}| f(x)\, dx\,, \qquad (25)$$

3) root mean square (RSM) [4],

4) squared deviations between brightness values (DEV):

$$\mathrm{DEV} = \left[\int_0^1 \int_0^1 (x-y)^2 f(x) f(y)\, dxdy \right]^{1/2}. \qquad (26)$$

The results of the studies using the above metrics of contrast are shown in Figures 6, 7, 8 and 9.

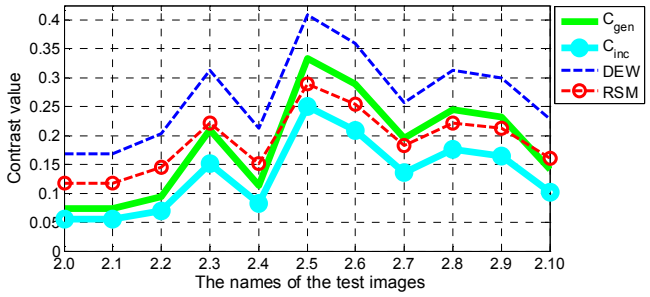


Fig. 6. Contrast of images from the first group (Figures 1 and 2).

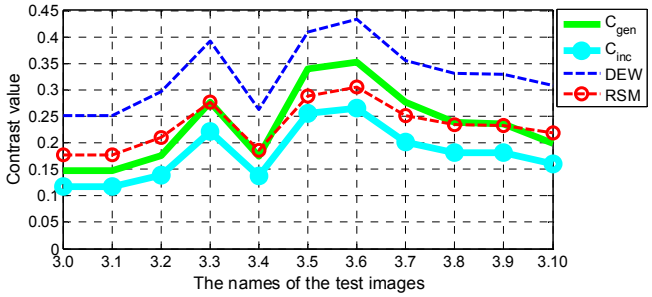


Fig. 7. Contrast of images from the second group (Fig. 1 and Fig. 3).

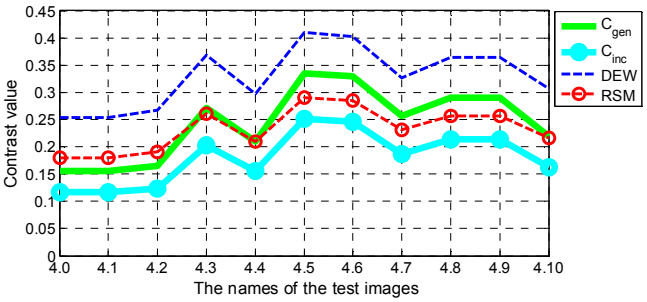


Fig. 8. Contrast of images from the third group (Fig. 1 and Fig. 4).

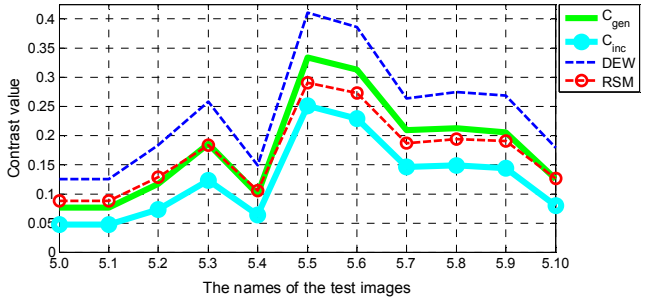


Fig. 9. Contrast of images from the fourth group (Fig. 1 and Fig. 5).

V. Discussion

The results of the studies confirm the well-known fact that methods of min-max and percentile stretching are ineffective to enhance images with a full dynamic range (Figures 2.1, 2.2, 4.1, 4.2).

The results of piecewise linear stretching (14), (16) and gamma correction are determined by the distribution of brightness and the values of the parameters of transformation (Figures 2.4, 3.3, 3.4, 5.4).

Histogram equalization based methods are most effective for increasing image contrast (Figures 2.5 - 2.7 and 4.5 - 4.7). The well-known and very significant drawbacks of these methods are a decrease in the contrast of objects with small sizes, an excessive increase in the contrast of extended objects, and, as a consequence of this, the possible appearance of unwanted artifacts and distortions in the image (Figures 3.5, 3.6, 5.5 and 5.6).

Research shows that the proposed techniques allow us to increase contrast by an average of 64-122% for all the raw images (Figures 6, 7, 8 and 9) without the appearance of unwanted artifacts and distortions (Figures 2.8, 2.9, 3.8, 3.9, 4.8, 4.9, 5.8, 5.9).

VI. Conclusions

The purpose of this study is to improve the efficiency of enhancing images in a fully automatic mode by parameter-free image intensity transformation.

In this work, a new approach to enhancing the image by adaptively transforming its intensity was proposed based on the analysis of brightness distribution at the boundaries of the objects of this image. To demonstrate the possibilities of this approach, new techniques of intensity transformation were proposed. Various approaches to assessing the increment of the function of intensity transformation had been considered.

The proposed techniques provide effective image enhancement without the appearance of distortion and artifacts and are focused on use in mobile applications.

These techniques can be considered as an alternative to traditional histogram equalization.

REFERENCES

[1] L. Xu and D. Doermann, Computer Vision and Image Processing Techniques for Mobile Application, Center for Automation Research, University of Maryland, LAMP-TR-151, 2008.

[2] W.K. Pratt, Introduction to Digital Image Processing, 1st edn., CRC Press, Taylor & Francis Group, Boca Raton, NW, USA, 2014, eBook ISBN 9780429170522.

[3] R.C. Gonzalez, R.E. Woods, Digital Image Processing. 4th Edition. Pearson Education, New Jersey (2018), ISBN 978-0-13-335672-4.

[4] W.K. Pratt, Digital image processing: PIKS Scientific inside, 4th ed., PixelSoft Inc., Los Altos, California, 2017.

[5] N.R. Mokhtar, et al., Image enhancement techniques using local, global, bright, dark and partial contrast stretching for acute leukemia images. In: Proc. of the World Congress on Engineering (WCE), vol. 1, pp. 807–812, London, U.K. (2009).

[6] O. Patel, Y. Maravi and S. Sharma, "A Comparative Study of Histogram Equalization Based Image Enhancement Techniques for Brightness Preservation and Contrast Enhancement", in Signal & Image Processing : An International Journal (SIPIJ) vol.4, no.5, pp. 11-25, Oct. 2013.

[7] R.A. Hummel, Histogram modification techniques, Computer Graphics and Image Processing, vol.4, issue 3, pp. 209–224, 1975.

[8] S. Yelmanov, Y. Romanyshyn, "A New Approach to Measuring Perceived Contrast for Complex Images", in book: "Advances in Intelligent Systems and Computing III", Springer Nature America Inc., AISC vol. 871, pp. 85–101, 2019.

IEEE Third International Conference on Data Stream Mining & Processing
August 21-25, 2020, Lviv, Ukraine

A Quick No-Reference Quantification of the Overall Contrast of an Image

Sergei Yelmanov
Special Design Office of Television Systems
Lviv, Ukraine
sergei.yelmanov@gmail.com

Yuriy Romanyshyn[1, 2]
[1]Lviv Polytechnic National University
[2]University of Warmia and Mazury
[1]Lviv, Ukraine, [2]Olsztyn, Poland
yuriy.romanyshyn1@gmail.com

***Abstract*—This study examines the issue of quickly quantifying the overall contrast of the image at the lowest required cost. The main goal of this study is to achieve better the accuracy and reliability of the quick assessment of global contrast for multi-element images. To meet this challenge, in this work, we propose a new approach to quickly quantify the contrast of the image by measuring its incomplete integral contrast. At the heart of the proposed approach is the assessment of contrast for each object in the image relative to the adaptation level given the sizes (area) of these objects. To demonstrate the possibilities of such an approach, new metrics proposed for measuring incomplete integral contrast with the use of different definitions of contrast kernels. Such an approach allows us to a more full and accurate evaluate the impact of objects on forming the estimate of image overall contrast.**

Keywords— image quality, global contrast, quantifying.

I. INTRODUCTION

Over the past ten years, there has been a steady trend of the ever-increasing use of technologies based on the analysis of video information to address a wide range of challenges in real-time applications. Widespread use of video information in real-time applications requires accurate and reliable assessing the quality of images in the tempo of their receiving [1, 2].

The problem of assessing the quality of images includes many aspects and, in the vast majority of cases, comes down to expert estimates or to quantify the most important characteristics of an image [1, 2, 3, 4].

Contrast is the most critical objective characteristic of the image, which has a significant impact on other indexes of its quality [3, 4, 5]. But, assess the global contrast of the image with a large number of objects with different characteristics is not so easy [4, 5] and this problem is now very relevant [6]. There are now many different ways to assess the global contrast of the image, but this task is still extremely far from its final solution [1, 2, 4, 7].

Known methods for no-reference assessing global contrast are, in their vast majority, integral estimates and are based on measuring the contrast of pairs of image elements (objects and background) [4, 8, 9]. The main drawbacks of these methods are that they based on heuristic estimates of the brightness distribution and that the sizes of the objects disregarded at assessing global image contrast [8, 9].

The objective of this study is to improve the reliability and accuracy of quickly no-reference assessing the global contrast of multi-element images.

To achieve the intended objective, in this work, a new approach to quantify the incomplete integral contrast is proposed based on analyzing the results of assessing the contrast of objects of this image relative to the image adaptation level, given the sizes of these objects.

II. RELATED WORKS

A large number of different aspects need to be considered when assessing the image quality. Contrast is a key factor that significantly affects the perception of image quality.

The global contrast (also often referred to as overall or generalized contrast) is the most crucial characteristic of the image, which largely determines its visual perception.

Different methods of quantifying the contrast of images are now known. The basis for the vast majority of known methods of no-reference assessing contrast in the image is quantifying the contrast of the pairs of its elements (objects and background) [8].

It is generally accepted that the overall contrast C_{ovr} is identical to the mean for the values of the contrast for all pairs of objects in an image [8]:

$$C_{\text{ovr}} = \int_{-\infty}^{\infty} \left| C_{ij} \right| \cdot p(C_{ij})\, dC_{ij} \qquad (1)$$

where C_{ij} is the contrast of two objects in an image; $p(C_{ij})$ is the distribution of contrast.

But assessing the distribution of contrast is in itself a rather daunting task. Given this, the overall contrast in an image is usually defined as [8]:

$$C_{\text{ovr}} = \int_0^1 \int_0^1 \left| C(b_i, b_j) \right| \cdot f(b_i, b_j)\, db_i\, db_j, \qquad (2)$$

where b_{i} and b_{j} are the values of image brightness; $C(b_{\text{i}}, b_{\text{j}})$ is a contrast of two image elements (objects), which is also referred to as the contrast kernel; $f(b_{\text{i}}, b_{\text{j}})$ is the joint probability density function.

However, this approach disregarded the level of adaptation, the magnitude of which has a strong impact on the overall contrast of the image.

To address this shortcoming, in [8] was proposed a concept of a complete integral contrast C_{com}:

$$C_{\text{com}} = \int_0^1 \int_0^1 \left| C_{\mu}(b_i, b_j) \right| \cdot f(b_i, b_j)\, db_i\, db_j, \qquad (3)$$

where $C_\mu(x_i, x_j)$ is the contrast between two image elements relative to a given adaptation level μ.

In [8], the notion of the contrast of two image elements relative to a given level μ of adaptation was suggested:

$$C_\mu(b_i, b_j) = \frac{C(b_i, \mu) + C(b_j, \mu)}{1 + C(b_i, \mu) \cdot C(b_j, \mu)}. \quad (4)$$

As the level μ of adaptation, the mean of image brightness is most often used [8], i.e.:

$$\mu = b_{mean} = \int_{L_{min}}^{L_{max}} b \cdot f(b)\, db, \quad (5)$$

where $f(b)$ is a probability density function (pdf).

At the heart of most of the existing no-reference metrics (for example, (2)-(4)) is the assessment of the contrast between the two image objects, which is also referred to as the contrast kernel.

The best-known definition of the kernel of contrast is the so-called weighted contrast, which is most often described as [8, 10]:

$$C^{wei}(b_1, b_2) = \mathrm{sgn}(b_2 - b_1) \cdot \frac{|b_2 - b_1|}{b_2 + b_1}, \quad (6)$$

where sgn(·) is the sign function:

$$\mathrm{sgn}(x) = \begin{cases} -1, & \text{if } x < 0 \\ 0, & \text{if } x = 0 \\ 1, & \text{if } x > 0 \end{cases} \quad (7)$$

Based on this estimate (4), the definition (3) for weighted contrast (6) can be described as [8]:

$$C^{wei}_{com} = \int_0^1 \int_0^1 \frac{|b_i \cdot b_j - \mu^2|}{b_i \cdot b_j + \mu^2} \cdot f(b_i, b_j)\, db_i\, db_j. \quad (8)$$

The major problem for (2) and (3) is to assess the joint distribution $f(b_i, b_j)$ of brightness. To solve this problem, the following assessment is most often used [8]:

$$f(b_i, b_j) = f(b_i) \cdot f(b_j). \quad (9)$$

Based on (9), a definition (8) of the complete contrast takes the form [8]:

$$C^{wei_1}_{com} = \int_0^1 \int_0^1 \frac{|b_i \cdot b_j - \mu^2|}{b_i \cdot b_j + \mu^2} \cdot f(b_i) \cdot f(b_j)\, db_i\, db_j. \quad (10)$$

For (9), the overall contrast (2) based on kernel (6) is defined as:

$$C^{wei}_{ovr} = \int_0^1 \int_0^1 \frac{|b_i - b_j|}{b_i + b_j} \cdot f(b_i) \cdot f(b_j)\, db_i\, db_j. \quad (11)$$

To measure contrast, the definition of linear kernels is also widely used [9]:

$$C^{lin}(b_1, b_2) = k \cdot \mathrm{sgn}(b_2 - b_1) \cdot |b_2 - b_1|, \quad (12)$$

where k is the scale factor.

For (12), most often assume that $k = (b_{max} - b_{min})^{-1}$ [9], or $k = 1$, where b_{min} and b_{max} are the extreme values of brightness in the image.

The definition (2) for linear contrast (12) using (9) can be described as [9]:

$$C^{lin}_{ovr} = k \cdot \int_0^1 \int_0^1 |b_i - b_j| \cdot f(b_i) \cdot f(b_j)\, db_i\, db_j. \quad (13)$$

In (2), to assess the joint distribution of image brightness, other approaches can also be used.

In [8], to simplify the calculations, the following assessment for distribution $f(b_i, b_j)$ was proposed:

$$f(b_i, b_j) = \delta(b_i - b_j) \cdot f(b_i) \quad (14)$$

where δ(·) is delta function.

In [8], based on assessments (3), (4) and (14), the concept of incomplete integral contrast was also proposed:

$$C_{inc} = 2 \int_0^1 \frac{C(b, \mu)}{1 + C(b, \mu)^2} \cdot f(b)\, db. \quad (15)$$

Incomplete contrast (15) for weighted kernel (6) has the form [8]:

$$C^{wei_1}_{inc} = 2 \int_0^1 \frac{|b^2 - \mu^2|}{b^2 + \mu^2} \cdot f(b)\, db. \quad (16)$$

The best-known and most used assessment of incomplete weighted contrast is:

$$C^{wei_2}_{inc} = 2 \int_0^1 \left| \frac{b - \mu}{b + \mu} \right| \cdot f(b)\, db. \quad (17)$$

It is also known the definition of incomplete linear contrast, which can be in general case defined as:

$$C^{lin_1}_{inc} = 2k \cdot \int_0^1 \min(|b - \mu|, T) \cdot f(b)\, db, \quad (18)$$

where T is the threshold value, parameter.

For a case where T = ½, the incomplete linear contrast has the form:

$$C^{lin_2}_{inc} = \int_0^1 \left| \frac{b - \mu}{BMAX} + \frac{1}{2} - \left| \frac{b - \mu}{BMAX} - \frac{1}{2} \right| \right| \cdot f(b) db, \quad (19)$$

where $BMAX$ is the maximum possible value of brightness.

For $T = 1$, the incomplete linear contrast takes the form [9]:

$$C_{\text{inc}}^{\text{lin}_3} = 2k \cdot \int_0^1 |b-\mu| \cdot f(b)\, db \,. \quad (20)$$

The above definitions (10), (11) and (16), (17), (19), (20) are no-reference metrics and widely used to assess the overall image contrast at a given level of adaptation.

However, these metrics have several disadvantages.

Existing metrics of incomplete integral contrast are inherently heuristic, as the basis for them is an approximation (9). These metrics also do not take into account the size of the object and its surrounding background when assessing the contrast between them.

The main goal of this study is to improve the reliability and accuracy of quick estimating the contrast of images.

To meet this challenge in this work a new approach to rapidly assessing the contrast of the image by quantifying its incomplete integral contrast is proposed.

III. Proposed approach

In this work, we propose a new approach to quickly assess the contrast of the image by measuring its incomplete integral contrast. At the heart of the discussed approach is the assessment of contrast for each object in the image relative to the adaptation level given the sizes (area) of these objects.

Suppose that the set of objects in the image is finite, and these objects have different brightness.

To evaluate the joint distribution $f(b_i, b_j)$ of brightness, we propose the following assessment:

$$f(b_i, b_j) = \delta(b_i, b_j) \cdot f_{\text{trg}}(b_i) \cdot f_{\text{bcg}}(b_j) \quad (21)$$

where $f_{\text{trg}}(b_i)$ is the density of brightness distribution for a single object (target) with brightness b_i; $f_{\text{bcg}}(b_j)$ is the density of distribution for the background that surrounds this single target with brightness b_j, where:

$$f_{\text{trg}}(b_i) = \Pr\{\omega \in \Omega : B(\omega) = b_i\} = f(b_i) \quad (22)$$

$$f_{\text{bcg}}(b_j) = \Pr\{\omega \in \Omega : B(\omega) \neq b_j\} = 1 - f(b_j) \quad (23)$$

where $\Pr\{\cdot\}$ is the probability of an event; Ω is a set of image pixels; ω is image pixel.

It follows from (22)-(23) that estimate (21) takes the form

$$f(b_i, b_j) = \delta(b_i, b_j) \cdot f(b_i) \cdot [1 - f(b_j)], \quad (24)$$

and is equal to:

$$f(b_i, b_j) = f(b_i) \cdot [1 - f(b_i)]. \quad (25)$$

Based on (3) and the proposed assessment (25), incomplete contrast we propose defined as:

$$C_{\text{prp}} = \alpha \cdot \int_{b_{\min}}^{b_{\max}} C_\mu(b) \cdot f(b) \cdot [1 - f(b)]\, db \quad (26)$$

where α is the normalizing factor, which is equal to:

$$\alpha = \left[\int_{b_{\min}}^{b_{\max}} f(b) \cdot (1 - f(b))\, db \right]^{-1} \quad (27)$$

Expressions (26)-(27) are a generalized description for the proposed approach to quantify the incomplete integral contrast.

Consider a simple two-element image that consists of one object (target) and the background surrounding it.

Let's also assume for the simple image that the brightness of the target and surrounding background are equal to b_i and b_{bg}, and the average brightness values for simple and source images are equal to each other. Assume further that the magnitude of adaptation level μ is equal to the average brightness b_{mean} (5), and b_{bg} is equal to b_j.

In this case for simple image, the brightness b_{bg} of the surrounding background takes the value:

$$b_{bg}(b_i, \mu) = \frac{\mu - b_i \cdot f(b_i)}{1 - f(b_i)} \quad (28)$$

Based on the above assumptions, the definition (6) of weighted contrast for a simple image can be defined as:

$$C_\mu^{\text{wei}}(b_i) = \text{sgn}(b_i - b_{bg}(b_i, \mu)) \cdot \frac{|b_i - b_{bg}(b_i, \mu)|}{b_i + b_{bg}(b_i, \mu)}, \quad (29)$$

$$C_\mu^{\text{wei}}(b_i) = \frac{b_i - \mu}{b_i + \mu - 2b_i \cdot f(b_i)} \quad (30)$$

Based on (26) and (30), we propose the definition of incomplete contrast for weighted kernel:

$$C_{\text{prp}}^{\text{wei}} = \alpha \cdot \int_{b_{\min}}^{b_{\max}} \left| \frac{b - \mu}{b + \mu - 2 \cdot b \cdot f(b)} \right| \cdot f(b) \cdot [1 - f(b)]\, db. \quad (31)$$

The linear contrast of a single object relative to its surrounding background at a given adaptation level defined as:

$$C_\mu^{\text{lin}}(b_i) = \beta \cdot \text{sgn}(b_i - b_{bg}(b_i, \mu)) \cdot |b_i - b_{bg}(b_i, \mu)|, \quad (32)$$

where β is a scaling factor that is determined from the condition:

$$C_{1/2}^{\text{lin}}(1) = -C_{1/2}^{\text{lin}}(0) = 1, \text{ for case where } f(1) = 1/2 \quad (33)$$

From (33) it follows that $\beta = 1$ and:

$$C_\mu^{\text{lin}}(b_i) = b_i - b_{bg}(b_i, \mu) = \frac{b_i - \mu}{1 - f(b_i)} \quad (34)$$

Expressions (30) and (34) describe the proposed definitions for contrast of two image elements relative to a given adaptation level.

Based on (26) and (34), we propose the definition of incomplete contrast for linear kernel:

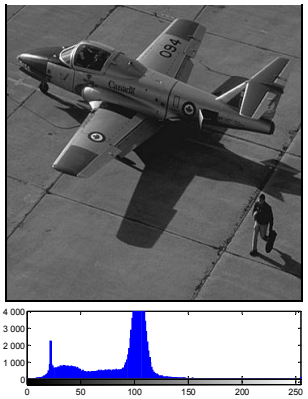
$$C_{\text{prp}}^{\text{lin}} = \int_{b_{\min}}^{b_{\max}} |b - \mu| \cdot f(b)\, db \Big/ \int_{b_{\min}}^{b_{\max}} f(b) \cdot (1 - f(b))\, db. \quad (35)$$

The definitions (31) and (35) are proposed no-reference metrics for incomplete contrast based on kernels (30) and (34).

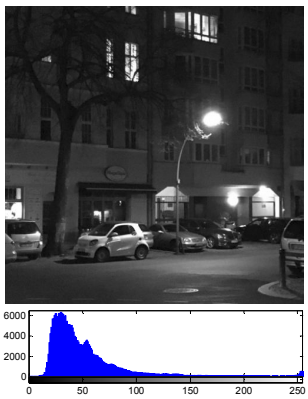
IV. Research

The study is based on the results of quantifying contrast for three groups of images using different metrics.

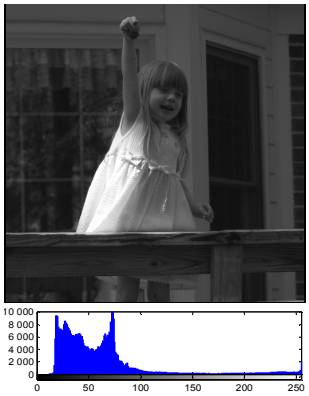
The each group consists of eight complex images, which, together with their histograms, are given in Figures 1, 2 and 3.



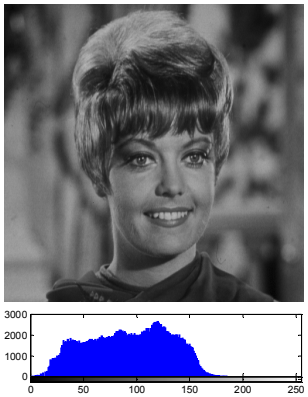
(1.1) test image 1.



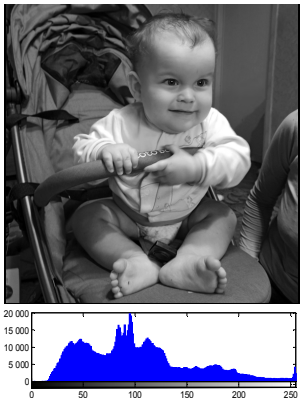
(1.2) test image 2.



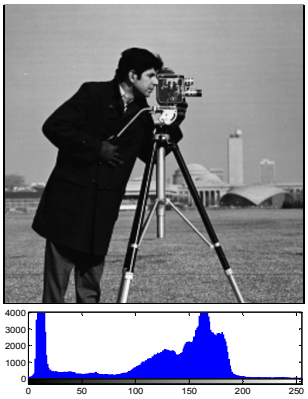
(1.3) test image 3.



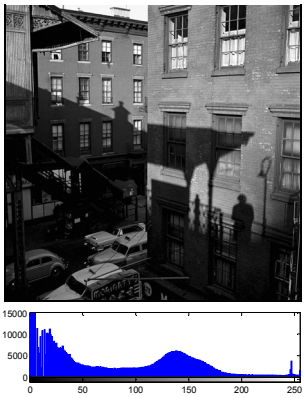
(1.4) test image 4.



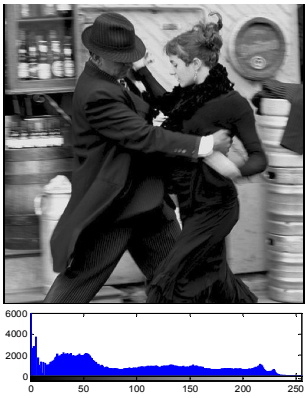
(1.5) test image 5.



(1.6) test image 6.

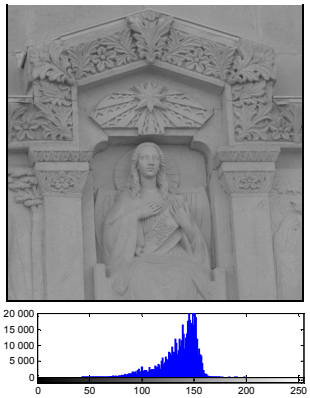


(1.7) test image 7.

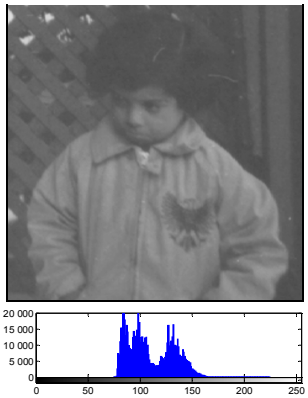


(1.8) test image 8.

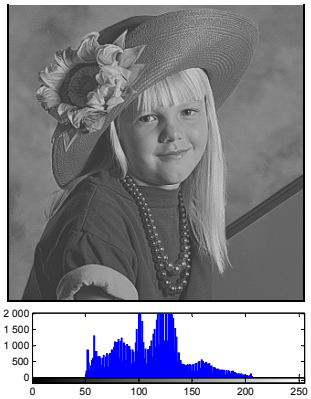
Fig.1. Test images from the first group and their histograms.



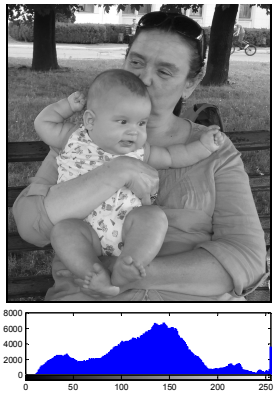
(2.1) test image 9.



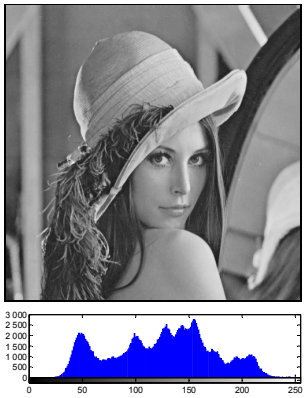
(2.2) test image 10.



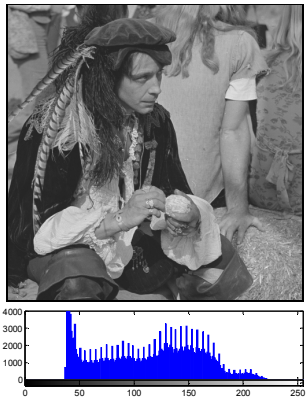
(2.3) test image 11.



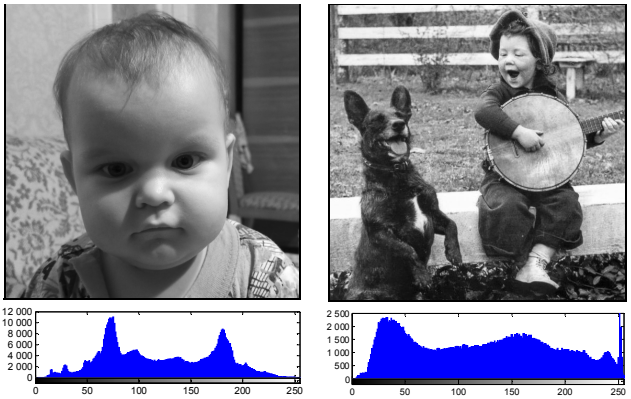
(2.4) test image12.



(2.5) test image13.

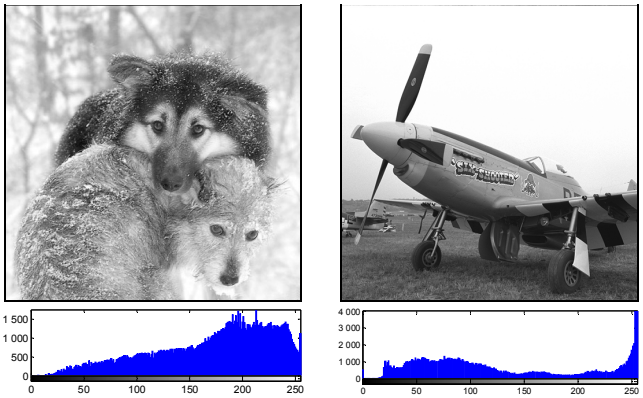


(2.6) test image14.



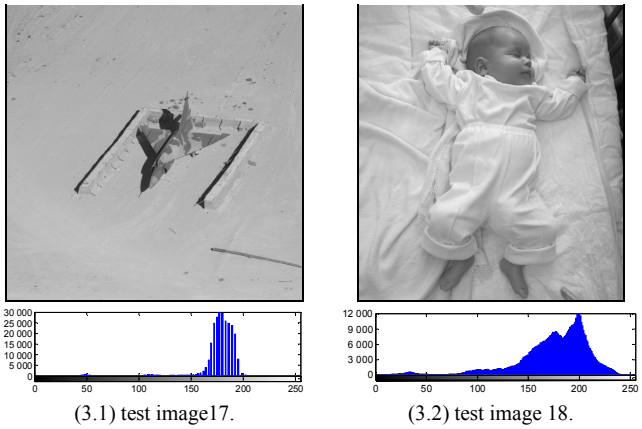
(2.7) test image15. (2.8) test image16.

Fig. 2. Test images from the second group and their histograms.

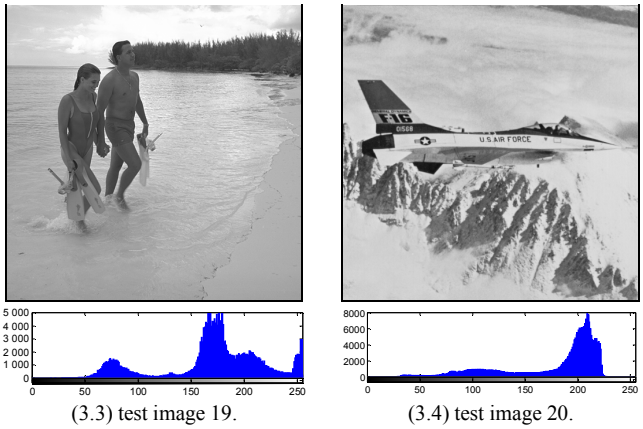


(3.7) test image 23. (3.8) test image 24.

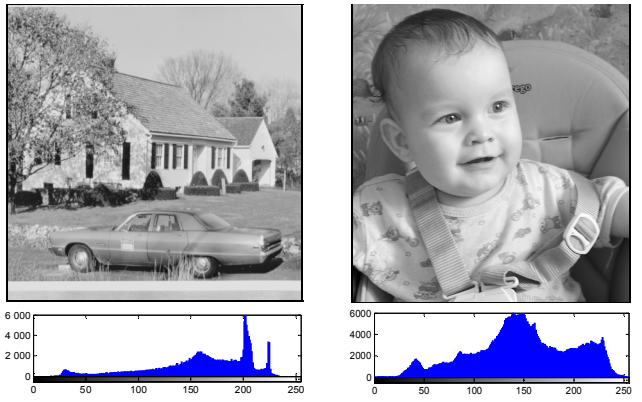
Fig. 3. Test images from the third group and their histograms.



(3.1) test image17. (3.2) test image 18.



(3.3) test image 19. (3.4) test image 20.



(3.5) test image 21. (3.6) test image 22.

The study was carried out by assessing the contrast of images using two groups of metrics.

The first group includes metrics of linear contrast:

1) overall linear contrast $C_{\text{ovr}}^{\text{lin}}$ (13);

2) squared deviations between brightness values (DEV);

3) incomplete integral linear contrast $C_{\text{inc}}^{\text{lin}_3}$ (20);

4) root mean square (RSM);

5) proposed incomplete lineal contrast $C_{\text{prp}}^{\text{lin}}$ (35);

The second group includes metrics of weighted contrast:

1) complete integral weighted contrast $C_{\text{com}}^{\text{wei}_1}$ (10);

2) overall weighted contrast $C_{\text{ovr}}^{\text{wei}}$ (11);

3) incomplete integral weighted contrast $C_{\text{inc}}^{\text{wei}_1}$ (16);

4) incomplete integral weighted contrast $C_{\text{inc}}^{\text{wei}_2}$ (17);

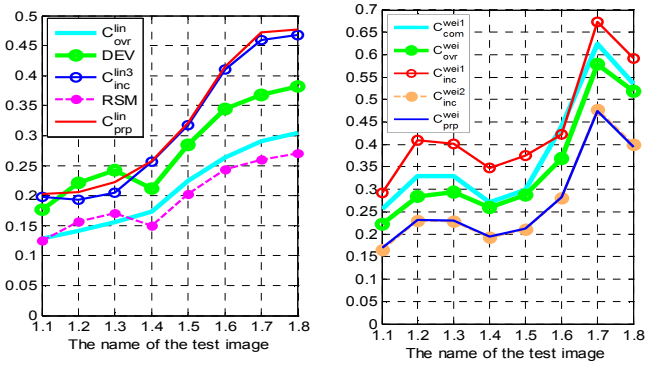
5) proposed incomplete weighted contrast $C_{\text{prp}}^{\text{wei}}$ (31).

THE RESULTS OF QUANTIFYING IMAGE CONTRAST

Name of test image	Used no-reference metrics of image contrast.									
	The metrics of linear contrast.					The metrics of weighted contrast.				
	$C_{\text{ovr}}^{\text{lin}}$	DEV	$C_{\text{inc}}^{\text{lin}_3}$	RSM	$C_{\text{prp}}^{\text{lin}}$	$C_{\text{com}}^{\text{wei}_1}$	$C_{\text{ovr}}^{\text{wei}}$	$C_{\text{inc}}^{\text{wei}_1}$	$C_{\text{inc}}^{\text{wei}_2}$	$C_{\text{prp}}^{\text{wei}}$
1.1	0.128	0.177	0.198	0.125	0.203	0.257	0.222	0.293	0.166	0.169
1.2	0.141	0.221	0.193	0.157	0.206	0.329	0.284	0.409	0.230	0.233
1.3	0.156	0.242	0.205	0.171	0.222	0.330	0.294	0.402	0.229	0.231
1.4	0.173	0.212	0.256	0.150	0.258	0.271	0.260	0.347	0.193	0.194
1.5	0.225	0.285	0.317	0.202	0.321	0.298	0.288	0.375	0.211	0.213
1.6	0.263	0.344	0.410	0.243	0.415	0.441	0.368	0.422	0.281	0.282
1.7	0.290	0.368	0.459	0.260	0.473	0.624	0.579	0.673	0.477	0.474
1.8	0.305	0.382	0.468	0.270	0.477	0.535	0.518	0.591	0.400	0.396
2.1	0.073	0.096	0.106	0.068	0.108	0.076	0.072	0.102	0.052	0.053
2.2	0.098	0.123	0.153	0.087	0.156	0.115	0.113	0.175	0.089	0.090
2.3	0.139	0.174	0.197	0.123	0.199	0.164	0.161	0.222	0.116	0.117
2.4	0.222	0.279	0.312	0.198	0.314	0.265	0.255	0.314	0.179	0.180
2.5	0.215	0.265	0.312	0.188	0.313	0.249	0.238	0.316	0.173	0.174
2.6	0.215	0.264	0.323	0.187	0.326	0.278	0.262	0.363	0.200	0.201
2.7	0.239	0.295	0.368	0.209	0.370	0.273	0.266	0.374	0.206	0.207
2.8	0.309	0.379	0.464	0.268	0.466	0.374	0.350	0.456	0.270	0.271

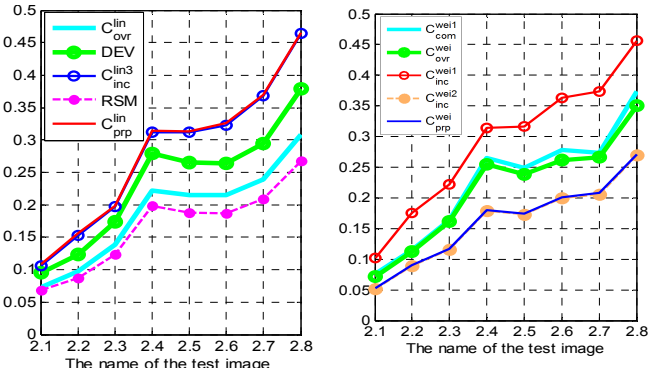
3.1	0.067	0.123	0.089	0.087	0.097	0.062	0.057	0.068	0.037	0.040
3.2	0.142	0.193	0.200	0.136	0.202	0.120	0.116	0.148	0.079	0.080
3.3	0.206	0.269	0.276	0.190	0.279	0.183	0.172	0.211	0.114	0.115
3.4	0.183	0.255	0.293	0.181	0.297	0.182	0.154	0.218	0.117	0.119
3.5	0.201	0.260	0.289	0.184	0.292	0.194	0.184	0.232	0.128	0.129
3.6	0.223	0.278	0.313	0.197	0.315	0.214	0.210	0.266	0.146	0.147
3.7	0.250	0.315	0.369	0.223	0.371	0.223	0.211	0.276	0.152	0.153
3.8	0.370	0.482	0.640	0.341	0.688	0.366	0.321	0.463	0.270	0.292

The results of measuring contrast are presented in Figures 4, 5 and 6.



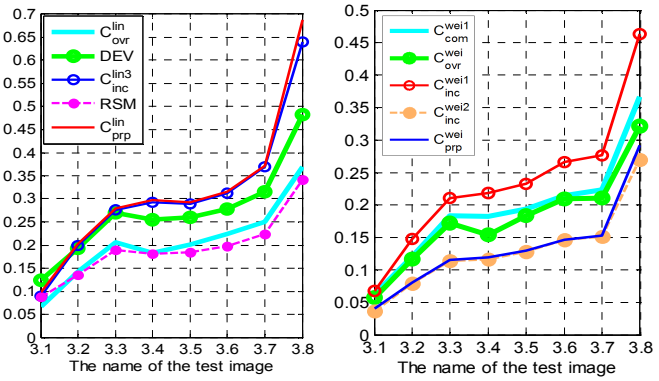
(4.1) Assessments for linear kernel. (4.2) Assessments for weighted kernel.

Fig. 4. Assessments of contrast for the first group of images (Figure 1).



(5.1) Assessments for linear contrast (5.2) Assessments for weighted kernel

Fig. 5. Assessments of contrast for the second group of images (Figure 2).



(6.1) Assessments for linear contrast (6.2) Assessments for weighted kernel

Fig. 6. Assessments of contrast for the third group of images (Figure 3).

V. Discussion

In this paper new definitions (31) and (35) of the incomplete integral contrast and their implementation for weighted (30) and linear (34) kernels are considered.

Studies show that the assessments (31) and (35) of the incomplete integral contrast (26) are close enough to the results of measuring using (17) and (20) for images with a uniform distribution of brightness (Table 1, Figures 4, 5 and 6). The proposed definition (26) for incomplete contrast allows improving the accuracy of quantifying the contrast for images with uneven distribution and large number of modes (Figures 1.2, 1.3, 1.7, 1.8 and 3.8). Research shows that the results of assessing the integral contrast using the proposed definition (31) are the closest to expert estimates of perceived contrast.

VI. Conclusion

The purpose of this study is to achieve better the accuracy and reliability of the quick assessment of global contrast for multi-element images. In this work a new approach to quickly quantify the contrast of the image by measuring its incomplete integral contrast was proposed.

At the heart of the proposed approach is the assessment of contrast for each object in the image relative to the adaptation level given the sizes (area) of these objects. The proposed approach allows us to a more full and accurate evaluate the impact of objects on forming the estimate of image overall contrast. To demonstrate the possibilities of this approach, new metrics (31) and (35) were proposed for measuring incomplete integral contrast with the use of different definitions of contrast kernels (30) and (34). The proposed approach (26)-(27), (29), (32) allows improving the accuracy and validity of assessing the contrast of complex images with uneven distribution and large number of modes.

References

[1] R. Soundararajan, A. Bovik, “Survey of information theory in visual quality assessment”, Signal, Image Video Process, vol. 7, no. 3, pp. 391-401, (2013).

[2] S. Chikkerur, V. Sundaram, M. Reisslein, and L.J. Karam, “Objective video quality assessment methods: a classification, review, and performance comparison”, in IEEE Trans. Broadcasting 57 (2), pp. 165 182 (2011).

[3] Sheikh and A. Bovik, “Image information and visual quality”, Image Processing, IEEE Transactions, vol.15, no.2, pp. 430–444, 2006.

[4] Z. Wang, A.C. Bovik, “Modern Image Quality Assessment.” in: Synthesis Lectures on Image, Video, and Multimedia Processing, Vol. 2, No. 1, pp. 1-156. Morgan and Claypool Publishers, New York (2006), vol. 2, no. 1, pp. 1-156.

[5] W. K. Pratt, Digital image processing: PIKS Scientific inside, 4th ed., Pixel Soft Inc., Los Altos, California, 2017.

[6] Z. Wang, A. Bovik and L. Lu, “Why is image quality assessment so difficult?”, in IEEE International Conference on Acoustics, Speech, and Signal Processing, vol. 4, pp. IV.3313-IV.3316, 2002.

[7] Z. Wang, H.R. Sheikh, and A.C. Bovik, “Objective video quality assessment”, in Furht The Handbook of Video Databases: Design and Applications, CRC Press, Austin, USA, (2003), vol.41, pp.1041-1078.

[8] 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, vol. 5, pp. 11-13.

[9] S. Yelmanov, Y. Romanyshyn, “A New Approach to Measuring Perceived Contrast for Complex Images”, in book: "Advances in Intelligent Systems and Computing III", Springer Nature America Inc., AISC vol. 871, pp. 85–101, 2019.

[10] A.A. Michelson, Studies in Optics, University of Chicago Press, Chicago (1927).

IEEE Third International Conference on Data Stream Mining & Processing
August 21-25, 2020, Lviv, Ukraine

Deep Feature Extractor with Information-Extreme Decision Rules for Visual Classification of Sewer Pipe Defects and its Training Method

Viacheslav Moskalenko
dept. of computer science
Sumy State University
Sumy, Ukraine
v.moskalenko@cs.sumdu.edu.ua,
https://orcid.org/0000-0001-6275-9803

Alona Moskalenko
dept. of computer science
Sumy State University
Sumy, Ukraine
a.moskalenko@cs.sumdu.edu.ua,
http://orcid.org/0000-0003-3443-3990

Zaretskyi Nikolay
dept. of computer science
Sumy State University
Sumy, Ukraine
n.zaretskij@gmail.com,
https://orcid.org/0000-0001-9117-5604

Viktor Lysyuk
Molfar.AI sp. z o.o
Gdansk, Poland
viktor.lysyuk@molfar.tech
https://orcid.org/0000-0003-1646-3336

***Abstract*— CCTV inspections are frequently used to diagnose defects in underground sewer pipes. A model for classification of sewer pipe defects in video frames and a corresponding five-phase training method are proposed. The model is based on deep convolutional neural network with a sigmoid output layer and information-extreme decision rules. The training method includes augmentation, feature extractor training resemblant of a Siamese network with triplet mining and softmax, computation of binary code for each class, training with joint binary cross-entropy loss using binary codes for each class as label and optimization of a hyper-spherical container for each class in Hamming space based on information criterion. Information criterion, expressed in the paper as logarithmic function of accuracy characteristics, provides a robust and reliable model for the most difficult case in the statistical sense. Results of a simulation with the optimal model on a dataset provided by *Ace Pipe Cleaning, Inc* confirm the efficiency of the proposed model and method and their suitability for practical usage.**

Keywords*— *sewer pipe inspection, convolutional neural network, Siamese network, information-extreme learning, classification

I. INTRODUCTION

Sewer pipes are critical infrastructure items which require frequent monitoring. A conventional method of analysis of sewer pipe conditions involves a CCTV sewer pipe inspection with the faults and defects inside the pipe identified and coded in a report in accordance with the applicable standards. Among the widely used standards are British MSCC5 and US PACP6 and PACP7. Reporting on the sewer condition in accordance with these standards requires careful examination and detailed analysis of collected CCTV footage. Application of machine vision and machine learning methods to CCTV inspections allows to increase the productivity and reduce inspection costs [1].

First attempts to use machine vision toolkit for pipe inspections involved edge detection and contour processing algorithms to identify the structural defects [2]. This method required manual hyper-parameter tuning and did not account for the contextual information, which reduced the reliability of such approach. Also, in this case no detailed analysis of the areas of interest for future coding to the applicable standards of pipe inspection was undertaken. An application of Gabor Filter for feature extraction during the CCTV sewer pipe inspections offered a more flexible and theoretically sound approach [3]. However, the shallow models of this type have low capacity and are not suitable for effective analysis of complex defects with contextual information. Besides, the automated CCTV inspection data analysis task has many unique challenges owing to the presence of a large number of artefacts and noise. Low visibility during inspection can occur for many reasons – due to steam, dense gas, position of the light source, substantial change in water level, blurring due to movement and low camera resolution.

Hierarchical feature extractors are a popular contemporary choice of a tool used to improve visual data analysis models. They are an integral part of deep data analysis models and are characterized by a higher information capacity in comparison with the shallow models [4, 5]. However, certain pipe defect classes appear infrequently and have high variability, which leads to imbalances in the data and low quantity of available labeled samples covering complex defects. This limits the application of the traditional deep learning methods, which typically require large quantities of labeled set.

Applying information theory ideas and methods and principles of decision rule synthesis in a geometric approach reduces the requirement for training data quantity and makes the process more robust. Siamese neural networks, often used for training of invariant feature embedding, are an example of such an approach [6]. Siamese neural networks show considerable promise under limited training data constraint and often form the base for the few-shot learning algorithms. Error-correcting output codes and information-extreme learning methods combine the principles of information theory and geometric approach [7, 8]. These algorithms are most useful for constructing the decision rules with the optimal feature dictionary or for improving the efficiency of the existing data analysis models. Thus the most promising path to further improvement of CCTV sewer

978-1-7281-3214-3/20/$31.00 ©2020 IEEE

inspection data analysis models is a combination of ideas and methods of Siamese neural networks, error-correcting output codes and information-extreme learning.

To improve the efficiency of the visual classification of sewer pipe defects models under the limited labeled training data constraint a Deep feature extractor with information-extreme decision rules and its new training method are proposed. The proposed algorithms are based on the ideas and methods of Siamese neural networks, error-correcting output codes and information-extreme learning. The current article considers all the training stages and their corresponding results obtained on the dataset provided by *Ace Pipe Cleaning, Inc.*

II. Model and Training Method Design

Sewer defect analysis is not performed on the whole of the frame but only on the parts of the frame located along the frame boundaries – where the distance counter value is known, which eases the defect localization. To this end a series rectangular images corresponding to 15% of the image area are cut from the sides, top and bottom of the frame. These images are then resized to 160x160 pixels. For experimentation and proof of concept a feature extractor based on the MobileNet general purpose convolutional network is proposed. The capacity coefficient of such convolutional network is set to 0.25 and only a convolutional backbone without fully connected layers is used [9]. The classifier model structure is depicted on Fig. 1.

MobileNet backbone
Global Average Pooling 2D Layer
Dropout Layer (rate=0.5)
Dense Layer (128 nodes)
Sigmoid Layer
Rounding Layer
RBF Layer

Fig.1 Image Classifier Model Structure

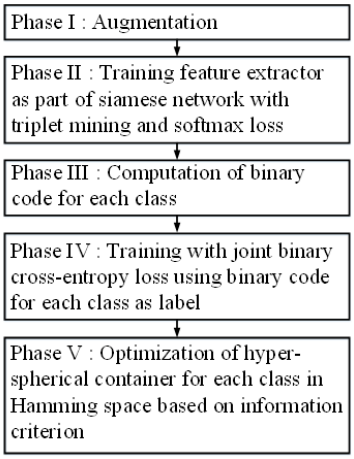
Global Average Pooling is used for dimensionality reduction and a Dropout pseudo-ensemble with 50% of the input features dropping is used for regularization [5, 9]. Fully connected and sigmoid layers form the output feature set.

Image classifier model's decision rules contain the rounding layer which produces the binary coded representation and radial-basis function defining the object's belonging to a certain class, with classes separated by hyper-spherical containers in binary Hamming space. Each hyper-spherical container is defined by the binary reference vector (container center) and container radius in Hamming distance units. In this case radial-basis membership function $\mu_z(b)$ for *N*-dimensional binary vector b is

$$\mu_z(b) = 1 - \sum_{i=1}^{N} b_i \oplus b^*_{z,i} / d^*_z, \qquad (1)$$

where b^*_z – binary reference vector (center of optimal container) for class X^o_z; d^*_z – radius of optimal container for class X^o_z in Hamming distance units.

The proposed method consists of 5 phases (Fig. 2). The first learning phase involves data augmentation, including brightness, scale and turn variations and salt and pepper noise overlay. Phases II-IV entail end-to-end learning and in the last phase the decision rules accounting for the dispersion of observations in each class in binary space are constructed. Second phase of image analysis model training is performed with the use of adaptive back propagation algorithms, of which Adam is the most popular [5, 6].



bat
fun
min

$$L = -\log \frac{\exp(\|f(x_a) - f(x_{ep})\|)}{\exp(\|f(x_a) - f(x_{ep})\|) + \exp(\|f(x_a) - f(x_{shn})\|)}, \qquad (2)$$

where $f(x)$ – function describing the feature extractor and establishing the dependency between the incoming image and sigmoid output layer vector; x_a – anchor image randomly chosen from the mini-batch; x_{ep} – easy positive example which is the most similar image with the same label as the anchor image, e.g.

$$x_{ep} = \arg \min_{x:C(x)=C(x_a)} \|f(x_a) - f(x)\|, \qquad (3)$$

where $C(x)$ – function returning a class label for sample x, x_{shn} – semi-hard negative example which is further from anchor than the selected positive example, which has a different label from the anchor image

$$x_{shn} = \arg \min_{\substack{C(x)\neq C(x_a) \\ x: \|f(x_a)-f(x)\| > \|f(x_a)-f(x_p)\|}} \|f(x_a) - f(x)\| \qquad (4)$$

The next training phase is necessary for the definition of the binary code for each class according to error-correcting output codes principle but accounting for the internal class structure and relationship between the different class samples. Training sample $\{x_{z,s} \mid z = \overline{1,Z}, s = \overline{1,n_z}\}$, containing n_z samples of z class is coded by a discrete representation $\{b_{z,s,i} \mid z = \overline{1,Z}, s = \overline{1,n_z}, i = \overline{1,N}\}$ with dimensionality N and a reference binary vector for each class is calculated. Binary coding is performed by providing images from the set $\{x_{z,s}\}$ as model inputs and rounding of the output of the sigmoid layer to the nearest integer

$$b_{z,s,i} = \begin{cases} 1, & \text{if } f(x_{z,s,i}) > 0.5; \\ 0, & \text{otherwise.} \end{cases} \quad (5)$$

A binary reference vector b_z for X_z^o class can be defined by bitwise comparison of frequency of binary ones in class X_z^o with background frequency of ones in the training set

$$b_{z,i} = \begin{cases} 1, & \text{if } \frac{1}{n_z}\sum_{s=1}^{n_z} b_{z,s,i} > \frac{1}{Z}\sum_{c=1}^{Z}\frac{1}{n_c}\sum_{s=1}^{n_c} b_{c,s,i}; \\ 0, & \text{otherwise.} \end{cases} \quad (6)$$

Binary reference vector of z-th class, b_z, can be used as a label for further training with joint binary cross-entropy loss, which for every input sample x can be calculated as

$$L = -\sum_{i=1}^{N}\left(b_i \log f_i(x) + (1-b_i)\log(1-f_i(x))\right), \quad (7)$$

where $f_i(x)$ – value of i-th sigmoid output for input image x; b_i – value of i-th bit of the reference vector for the class which image x belongs to.

The last machine learning phase involves the container radius optimisation by information criterion to establish the deviation boundaries of the observation binary representation of each class from their representative reference vectors

$$E_z^* = \max_{\{d\}} E_z(d), \quad (8)$$

where $\{d\} = \{0,1,...,\left(\sum_i b_{z,i} \oplus b_{c,i} - 1\right)\}$ is a set of concentric radiuses with center b_z (reference vector) of data distribution in class X_z^o, which is recomputed using rule (6); E_z – information criterion for X_z^o class, which is computed as function of accuracy characteristics [8, 10]

$$E_z = 1 + \frac{1}{2}\left(\frac{\alpha_z}{\alpha_z + D_{2,z}} log_2 \frac{\alpha_z}{\alpha_z + D_{2,z}} + \right.$$
$$+ \frac{\beta_z}{D_{1,z} + \beta_z} log_2 \frac{\beta_z}{D_{1,z} + \beta_z} + \frac{D_{1,z}}{D_{1,z} + \beta_z} log_2 \frac{D_{1,z}}{D_{1,z} + \beta_z} +$$
$$\left. + \frac{D_{2,z}}{\alpha_z + D_{2,z}} log_2 \frac{D_{2,z}}{\alpha_z + D_{2,z}}\right), \quad (9)$$

where α_z – false positive rate of decision rule for z-th class; β_z – false negative rate; $D_{1,z}$ – true positive rate (sensitivity); $D_{2,z}$ – true negative rate (specificity).

The model output layer has an information theory basis. Binary reference class vectors are analogous to self-correcting Hamming codes and container radius is characteristic of the maximum quantity of errors which can be corrected after the receipt of message via a noisy channel. Information criterion defines the measure of reduction in uncertainty after the receipt of the message from a given alphabet.

Thus hyper-spherical container radius optimisation allows to increase the decision-making efficiency under the conditions of overlapping classes with different size of observations' distribution. It also allows detecting the newness in the data in case of appearance of observations which do not fit either of the existing classes. The optimisation criterion in this case is a logarithmic information measure which increases the models ability to generalise and increases its robustness in statistically complicated cases.

III. Training Results and Discussion

Although a multitude of sewer pipe materials exists, only Concrete and Vitrified Clay pipes are considered [1]. These are amongst the most common pipe materials and have a wide variety of characteristic defects. The [3] class alphabet contains 9 most frequent defects. Class X_1^o denotes a «roots» defect, corresponding to tree roots penetrating the pipe. Class X_2^o corresponds to «deposits» defect, causing the reduction in the pipe cross-section. X_3^o, X_4^o, X_5^o, X_6^o and X_7^o classes characterize various structural pipe defects such as «surface damage», «crack», «fracture», «broken» and «hole» respectively. Class X_8^o denotes «infiltration», a defect occurring when water enters the pipe from outside through a structural defect or defective pipe joint or seeps through the porous area in the sewer pipe wall. Class X_9^o corresponds to «Obstruction» defect, a foreign object in the pipe which substantially obstructs the free flow of water or reduces the flow rate. Class X_{10}^o corresponds to the normal pipe and is required for regularisation. 80 samples were collected for each class. Naturally, each of the chosen classes has a considerable number of further sub-classes; however the further classification image analysis is outside the scope of this research. Each of the images was used to expand the training set 4X by applying slight random augmentation such as brightness and scale variation, turning the image up to 10% of the full range and applying salt and pepper noise mask.

The feature extractor is first modified by replacing the sigmoid layer for softmax layer to implement a classic training method with categorical cross-entropy loss and Adam optimiser [4, 5]. Results thus obtained are used as a baseline for comparison with the proposed approach. Fig. 3 depicts the changes in model loss (test_loss and train_loss) and accuracy (test_acc and train_acc) on the test and training sets at each epoch. Each mini-batch contains 32 images and learning rate is set to 0.0001. Analysis of Fig. 3 shows that the loss reduction effectively stopped after 40 epochs with corresponding accuracy of the resulting model equal to 86% on test set.

Fig. 4 depicts the training result for the feature extractor with sigmoidal output layer, where the first 30 epochs occur with the use of loss function (2) and the following 30 epochs with the use of loss function (7). Decision rules are constructed and the class-wise averaged information efficiency criterion (9) are calculated every 5 epochs.

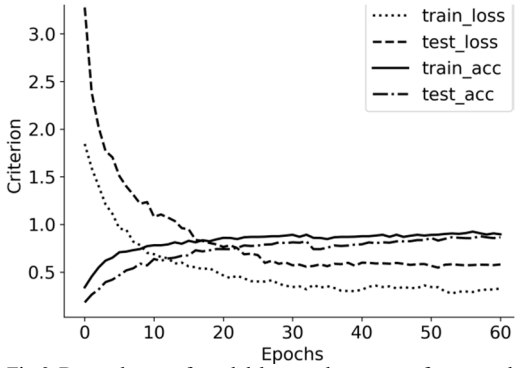


Fig.3 Dependency of model loss and accuracy from number of epochs on test and train sets for conventional approach

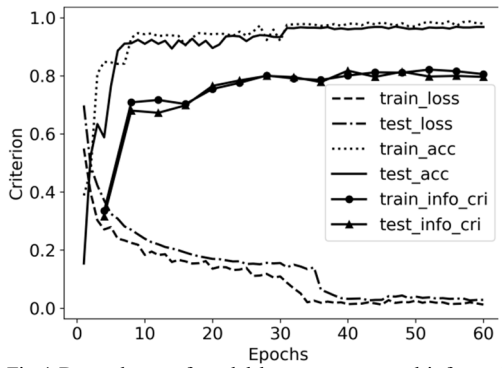


Fig.4 Dependency of model loss, accuracy and information criterion from number of epochs on test and train sets for proposed approach

Analysis of Fig. 4 shows that, beginning from the 10th epoch, reduction in loss on test set (test_loss) slows down considerably and from 31st epoch an increase in the decision rule accuracy by more than 2.5% takes place. However, after the 40th learning epoch, further improvement de-facto stops, with accuracy reaching 98% on the test dataset and information criterion values for the test dataset and training dataset (train_info_cri, test_info_cri) virtually the same at 0.858. The average distance between centres of the hyper-spherical containers is 9 units of Hamming distance whilst the average radius of containers is 8 units. This evidences high robustness of information-extreme decision rules [8, 10] to noise. Thus a multi-phase learning process with information-extreme decisions rules allows to improve the sewer pipe defect detection model accuracy by 12% in comparison with the traditional training scheme with softmax output layer.

CONCLUSIONS

The scientific novelty of the results is as follows:

– the trainable model for visual classification of sewer pipe defects based on multilayer convolutional feature extractor with information-extreme radial-basis decision rules is proposed for the first time. The model is characterized by better computational efficiency and generalisation abilities;

– the multi-phase training method, incorporating feature extractor training as a part of Siamese network, binary code computation for each class, feature extractor training with joint binary cross-entropy loss using the class binary code as a label and radial-basis optimisation decision rules based on information criterion, which allows to increase generalization abilities in case limited size of training set, is proposed;

– the effectiveness of the proposed approach is compared with the conventional one based on a convolution network with a soft-max output layer.

2. The practical value of the obtained results for sewer pipe inspection is the formation of a modern methodological basis for designing the defect detector in case of limited availability of labeled training examples per class. At the same time, the results of model testing on dataset provided by *Ace Pipe Cleaning, Inc* confirm the high efficiency of proposed approach and its suitability for practical application.

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] Moradi, Saeed et al. "Review On Computer Aided Sewer Pipeline Defect Detection And Condition Assessment". Infrastructures, vol 4, no. 1, 2019, p. 10. MDPI AG, doi:10.3390/infrastructures4010010.

[2] Syahrian, Nur Mutiara et al. "Vision-Based Pipe Monitoring Robot For Crack Detection Using Canny Edge Detection Method As An Image Processing Technique". Kinetik, vol 2, no. 4, 2017, p. 243. Universitas Muhammadiyah Malang, doi:10.22219/kinetik.v2i4.243.

[3] Myrans, Joshua et al. "Automated Detection Of Fault Types In CCTV Sewer Surveys". Journal Of Hydroinformatics, vol 21, no. 1, 2018, pp. 153-163. IWA Publishing, doi:10.2166/hydro.2018.073.

[4] Cheng, Jack C.P., and Mingzhu Wang. "Automated Detection Of Sewer Pipe Defects In Closed-Circuit Television Images Using Deep Learning Techniques". Automation In Construction, vol 95, 2018, pp. 155-171. Elsevier BV, doi:10.1016/j.autcon.2018.08.006.

[5] F. Panella, et al. "Deep learning and image processing for automated crack detection and defect measurement in underground structures". ISPRS - International Archives Of The Photogrammetry, Remote Sensing And Spatial Information Sciences, XLII-2, 2018, pp. 829-835. Copernicus Gmbh, doi:10.5194/isprs-archives-xlii-2-829-2018.

[6] Zhan, Huijing et al. "Deepshoe: An Improved Multi-Task View-Invariant CNN For Street-To-Shop Shoe Retrieval". Computer Vision And Image Understanding, vol 180, 2019, pp. 23-33. Elsevier BV, doi:10.1016/j.cviu.2019.01.001.

[7] Joutsijoki, Henry et al. "Error-Correcting Output Codes In Classification Of Human Induced Pluripotent Stem Cell Colony Images". Biomed Research International, vol 2016, 2016, pp. 1-13. Hindawi Limited, doi:10.1155/2016/3025057.

[8] V. Moskalenko et al. "The Model And Training Algorithm Of Compact Drone Autonomous Visual Navigation System". Data, vol 4, no. 1, 2018, p. 4. MDPI AG, doi:10.3390/data4010004.

[9] Michele, Aurelia et al. "Mobilenet Convolutional Neural Networks And Support Vector Machines For Palmprint Recognition". Procedia Computer Science, vol 157, 2019, pp. 110-117. Elsevier BV, doi:10.1016/j.procs.2019.08.147.

[10] A. Dovbysh et al. "Information Synthesis Of Adaptive System For Visual Diagnostics Of Emotional And Mental State Of A Person". Eastern-European Journal Of Enterprise Technologies, vol 4, no. 9(82), 2016, p. 11. Private Company Technology Center, doi:10.15587/1729-4061.2016.756

IEEE Third International Conference on Data Stream Mining & Processing
August 21-25, 2020, Lviv, Ukraine

Handwriting Styles Clustering: Feature Selection and Feature Space Analysis based on Online Input

Karyna Korovai
Samsung R&D Institute Ukraine
57, L'va Tolstogo Str., Kyiv, 01032, Ukraine
Faculty of Computer Science and Cybernetics
Taras Shevchenko National University of Kyiv
k.korovai@samsung.com

Oleksandr Marchenko
Samsung R&D Institute Ukraine
57, L'va Tolstogo Str., Kyiv, 01032, Ukraine
ol.marchenko@samsung.com

***Abstract*—In this paper, we perform a detailed analysis of feature extraction approaches and existing solutions for the problem of handwriting styles clustering. We consider a number of handcrafted features for handwritten text, perform a feature space analysis for finding the best set of extracted features and build an expert system for identifying and clustering handwriting styles using unsupervised methods based on this set. We observed an improvement in clustering results by analyzing the described in this paper metrics for clustering evaluation (such as the Dunn index, silhouette, within- and between-cluster sums of squared errors), therefore we can conclude that the clustering based on such a subset is much more efficient. The results outlined below can be used for handwriting style determination, which is an important step in designing systems for text recognition and localization, authentication and verification tasks, and also can be applied to the problem of the detection of the mental or nervous disorders.**

***Index Terms*—clustering, online handwriting, feature selection, feature engineering, feature space analysis.**

I. Introduction

Over the last few decades, along with the evolution of computerized handwriting recognition systems, a pressing need for research in handwriting has emerged. Handwriting style classification and clustering problems have been widely studied for both online [1] and offline [2] handwriting clustering, recognition systems [3], authentication and verification tasks [4]–[6], writer identification [7] problem, character segmentation [8], text/non-text classification [9], [10], sentiment analysis (emotional context recognition) [11], and even for human-like handwriting generation [12]. It is also can be used for the detection of mental or nervous disorders such as Parkinson's disease [13], [14] and other illnesses [15].

Like other biometric data, handwriting demonstrates remarkable *variation* and *variability*. Variation refers to the difference in handwriting between different people, whereas variability refers to the ability of one person's handwriting to change over time and under different circumstances. That is, handwriting may be different for different writers (interwriter difference) and even for the same writer (intrawriter difference). Variability and variation are mostly explained by personal, emotional, internal, and other side factors [16]. In addition, previous studies have shown that the style of handwriting can vary significantly depending on geographical location, cultural background, age, gender, education, profession, etc. [17]–[22].

Since there are no clear rules for establishing the style (i.e. clear marking) of handwriting, this task does not have a clear solution. For the same reason, clustering algorithms are usually considered for this problem rather than classification methods. In the related works Gaussian mixture model [2], K-means algorithm [7], self-organizing maps [1], and agglomerative hierarchical clustering [23] are usually used. Many approaches have been proposed for style extraction on handwriting stroke, symbol, word, and line-of-text levels; all of them were aimed at getting characteristics (features) that reflect a variety of handwriting styles and are specific to one cluster and as different as possible for other clusters.

In this paper, we propose a method for feature space analysis and feature selection for time series clustering problem in the case of handwritten text styles.

II. Data Analysis and Processing

A. Data Description

The research was based on a dataset that contains handwritten data in an online format, where the data is presented as a sequence of points. Thus, we can describe the style of the handwriting using not only visual characteristics but the dynamics of the writing process (such as speed, acceleration, and jerk) as well. It also allows us to consider handwriting as a time series, so data can be written as a sequence of three temporal functions (1), that represent the trajectory of the pen movement (x and y coordinates) and the pen pressure p when writing on the electronic device.

$$[x(t), y(t), p(t)]. \tag{1}$$

The stroke, which is the minimal unit of the electronic path, in this case, is a basic data structure. The stroke can be described as a continuous sequence of points:

$$s = \{x(t_i), y(t_i), p(t_i)\}_{i=1}^{n}, \tag{2}$$

where n is the length of the stroke s.

One of the most popular datasets with the structure described above is the IAM On-Line Handwriting Database

978-1-7281-3214-3/20/$31.00 ©2020 IEEE

(IAM-OnDB) [24], which is an open dataset that contains handwritten English text samples. In addition to the recorded data and the corresponding text tag, IAM-OnDB also stores information about authors' native language and country, other languages they speak, their age, sex, whether they are right- or left-handed. This provides a high variation in the dataset.

B. Data Processing

A number of studies [25], [26] demonstrate that resampling (to obtain a constant number of regularly spaced points on the trajectory) allows getting more accurate results of handwriting recognition, authentication, etc. In particular, it improves the classification and clustering quality of handwritten characters.

Thus, we first design features possessing temporal information based on the analysis of the text writing process dynamics. Before proceeding to the extraction of spatial and geometric characteristics of handwriting we perform a resampling to remove the temporal noise from the data.

Given that the points in every stroke are now equidistant (as opposed to the input sequence where the points were evenly spaced in time) and can be represented as follows:

$$[x(n), y(n), p(n)], \tag{3}$$

where the values of the parameters at each point become equally representative, regardless of the speed of writing.

Symbols are also shifted to the origin and scaled in order to make the representation invariant to displacement and scale distortion. For every point (x_i, y_i):

$$x_i = \frac{x_i - x_{min}}{y_{max} - y_{min}}, y_i = \frac{y_i - y_{min}}{y_{max} - y_{min}}, \tag{4}$$

where x_{min} is the smallest x-coordinate, y_{min} and y_{max} is the smallest and the largest y-coordinate respectively (both over the entire syllable), and all $y_i \in [-1; 1]$. Also note that after completing the transformations described above, no information is lost about the slant of the handwriting, the curvature in points, the cursivity level, the height-to-width ratio of characters, and other important characteristics, since these transformations do not deform the text as we normalize samples only by height.

III. Feature Engineering

This section describes the process of extracting features from handwriting. As a result we get a global space-time vector that is a basis for handwriting styles clustering.

Feature extraction is the process of transforming data into a form that is easier to interpret; in this case, we are interested in making the data representation clearer for the machine learning model. Our goal is to extract the representative information from the raw data, which would minimize the variability of objects within one cluster while increasing the variation between the objects of different clusters. It is also important that the constructed features ignore variability as much as possible, but take into account variation. In addition, the selected feature set should characterize the style of the writer regardless of the semantic content of the words.

There are the following types of features:

1) dynamic (include local dynamic information such as speed, acceleration, and jerk [23]);
2) spatial or geometric (describe the static shape of local pen trajectories using local angles, curvature [26], connectivity, or inter-character spaces).

This section outlines methods for extracting these features for further handwriting styles clustering.

A. Dynamic Features

Note that during the feature extraction process, the points in the input sequence are evenly spaced over time, and therefore the value of Δt is constant for each point, so these values can be neglected, which facilitates the extraction of dynamic features and reduces the number of calculations required for this purpose.

As dynamic parameters, we consider the horizontal ($V_x(n)$) and vertical ($V_y(n)$) velocities calculated locally at some point n:

$$V_x(n) = \frac{\Delta x(n)}{\Delta t(n)}, V_y(n) = \frac{\Delta y(n)}{\Delta t(n)}, \tag{5}$$

where

$$\begin{aligned} \Delta x(n) &= x(n+1) - x(n-1), \\ \Delta y(n) &= y(n+1) - y(n-1), \\ \Delta t(n) &= t(n+1) - t(n-1). \end{aligned} \tag{6}$$

A similar process is performed to extract local acceleration ($A_x(n)$ and $A_y(n)$) and jerk ($J_x(n)$ and $J_y(n)$) values associated with horizontal and vertical velocity and acceleration derivatives, respectively:

$$A_x(n) = \frac{\Delta V_x(n)}{\Delta t(n)}, A_y(n) = \frac{\Delta V_y(n)}{\Delta t(n)}, \tag{7}$$

$$J_x(n) = \frac{\Delta A_x(n)}{\Delta t(n)}, J_y(n) = \frac{\Delta A_y(n)}{\Delta t(n)}, \tag{8}$$

where

$$\begin{aligned} \Delta V_x(n) &= V_x(n+1) - V_x(n-1), \\ \Delta V_y(n) &= V_y(n+1) - V_y(n-1), \\ \Delta A_x(n) &= A_x(n+1) - A_x(n-1), \\ \Delta A_y(n) &= A_y(n+1) - A_y(n-1). \end{aligned} \tag{9}$$

B. Spatial Features

Different global and local character properties can be represented by geometric and topological features with a high tolerance for distortion and style variations. This type of presentation may also encode some knowledge about the structure of an object.

For spatial features, the data preprocessing process described in section II-B is first performed to ensure that the consecutive points in a given word are equidistant since the parameter values at each point become equally representative regardless of velocity. After processing, information about the

direction and curvature of trajectories at a point (a given time point), which are important signs for identifying handwriting style in clustering and classification tasks, can be retrieved.

An important geometric feature is a direction diagram [27]. The direction of the stroke at the point is estimated as follows:

$$\begin{aligned} cos\theta(n) &= \frac{\Delta x(n)}{\Delta s(n)}, \\ sin\theta(n) &= \frac{\Delta y(n)}{\Delta s(n)}, \end{aligned} \tag{10}$$

where

$$\begin{aligned} \Delta x(n) &= x(n+1) - x(n-1), \\ \Delta y(n) &= y(n+1) - y(n-1), \\ \Delta s(n) &= \sqrt{\Delta x(n)^2 + \Delta y(n)^2}. \end{aligned} \tag{11}$$

These values, as in the case of dynamic features, are calculated throughout the strokes and are used to construct a histogram of 8 bins. The main advantage of this method is the fact that we do not need to perform many complex calculations as the points in the input sequence are equidistant, so $\Delta s(n)$ is constant for each value and can be neglected. The geometric interpretation of the above method is shown in Fig. 1. A visualization of the constructed diagram is shown in Fig. 2.

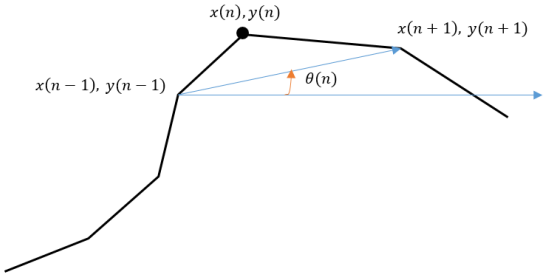


Fig. 1: Finding the angle of inclination at point n.

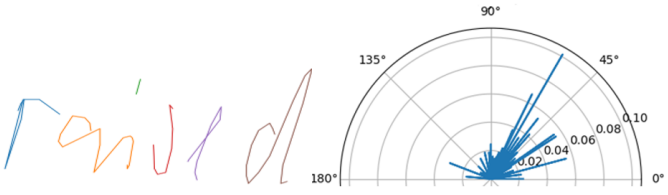


Fig. 2: The direction diagram for the word "raised" (from 0° to 180°).

Also, it is reasonable to include information related to the curvature at the point. Therefore, we measure the local curvature angle between two elementary segments (Fig. 3):

$$\phi(n) = \theta(n+1) - \theta(n-1). \tag{12}$$

This angle can be represented by its cosine and sine, which can be calculated by the following formulas:

$$\begin{aligned} cos\phi(n) &= cos\theta(n-1) \cdot cos\theta(n+1) + \\ &+ sin\theta(n-1) \cdot sin\theta(n+1), \\ sin\phi(n) &= cos\theta(n-1) \cdot sin\theta(n+1) + \\ &+ sin\theta(n-1) \cdot cos\theta(n+1). \end{aligned} \tag{13}$$

All these values can be calculated using formulas (10).

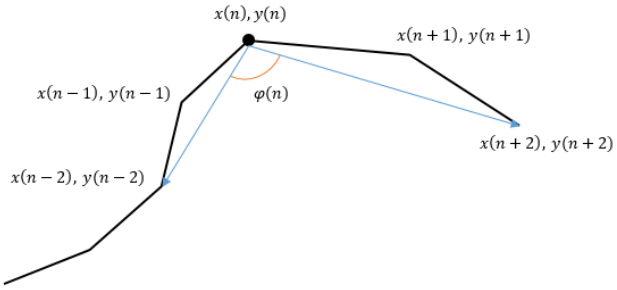


Fig. 3: Finding the curvature at point n.

In addition, it is also important to take into account handwriting conjoints, therefore we introduce a cursivity index. The idea is to associate the number of letters in a word with the number of strokes, that is, the index of cursivity is essentially a measure of the connectedness of the handwriting. The cursivity index for given n words can be calculated by the following formula [28]:

$$C_n = \frac{1}{n} \sum_{w-1}^{n} \left(\frac{N_{l,w} - N_{pd,w} + 1}{N_{l,w}} \right), \tag{14}$$

where $N_{l,w}$ is a number of letters in word w, $N_{pd,w}$ is a number of strokes in word w and C_n is a cursivity index.

It is assumed that for conjoint handwriting C_n will be close to 1, since each word is written with a single stroke. Conversely, if the handwriting is close to typed, it means that each letter is written in a separate stroke, and therefore C_n is about 0. The mixed style gives intermediate values between 0 and 1. If multiple strokes are used to spell each letter (such as when writing hieroglyphs), C_n will be negative.

IV. Feature Selection and Feature Space Analysis

Handwriting style clustering is a complex problem that lacks a priori knowledge of the style labels and categories. The styles identified by any clustering algorithm depends on static and dynamic characteristics (see section III) extracted from the input samples.

Due to the physiological features of the text-writing process, the handwriting is not standardized, prone to distortion, is substantially individual, and re-writing the same sample text is not identical to the previous variant of writing. As a product of the human activity, handwriting styles cannot be clearly separated from each other, since there will always be an "intermediate" style for any two styles.

This fact can be illustrated with the following experiment. Consider the clustering results that are shown in Fig. 4. In this case, we perform a two-tier clustering: at first data were initially clustered by slant, so the objects were split into three clusters:

- clusters 1, 2, 3 - left-slanted handwriting;
- clusters 7, 8, 9 - right-slanted handwriting;
- clusters 4, 5, 6 - handwriting with no slant at all.

After that, each of the obtained clusters was split into 3 groups according to the cursivity index:

- $1 \geq C^1 > C^2 > C^3 \geq -1$;
- $1 \geq C^4 > C^5 > C^6 \geq -1$;
- $1 \geq C^7 > C^7 > C^9 \geq -1$,

where C^i is an average cursivity index for the i-th cluster.

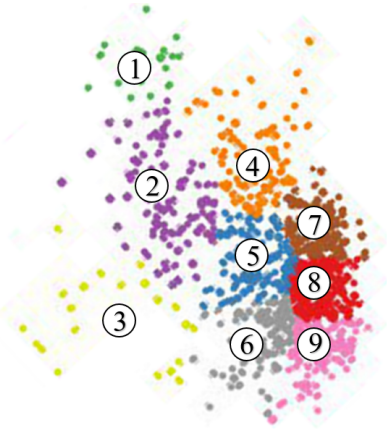
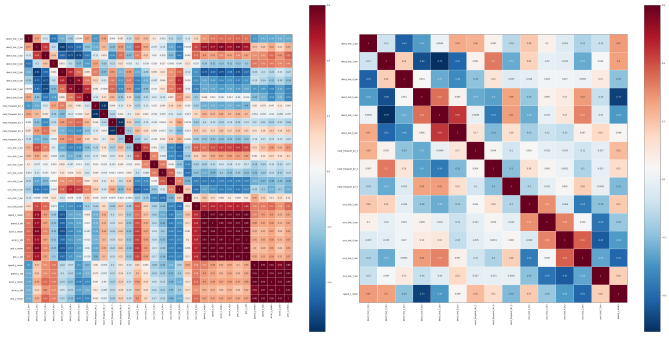


Fig. 4: Two-tier clustering by slant and cursivity index.

Feature selection is a process of automatic or manual selection attributes as the basis for the further machine learning process. Our goal is to identify a group of the most effective features for clustering. The main prerequisite for feature selection technique is features irrelevance or/and redundancy so that some of them can be deleted without much information loss. There are following advantages of feature selection:

- overfitting reduction: less excess data means less noise-based decision-making;
- accuracy improving: modeling accuracy is enhanced by training on more relevant data;
- learning time reduction: the lower dimension of the feature vector reduces the complexity of the algorithm, which results in the faster training process;
- avoiding the curse of dimensionality.

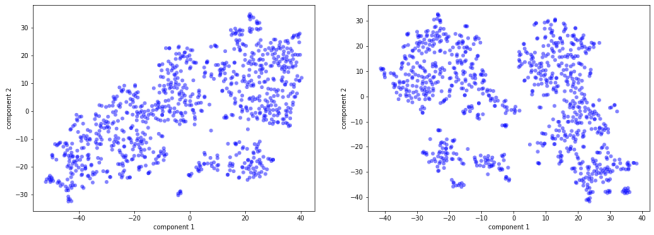
In this section, the feature selection process is based on the correlation matrix analysis. The idea of correlation-based feature selection is that good feature sets contain features that are highly correlated with the target variable but are independent of each other [29], [30]. Fig. 5 shows the correlation matrix for the constructed features. As we can see, some of the constructed features are redundant and should be removed in order to increase the quality of clustering. To improve the quality of clustering, we remove the most correlative features from the data set: 2nd and 6th bins of the diagram of directions; 2nd and 3rd most frequent angles; 1st, 6th and 8th curvature diagram bins; dynamic features other than the average abscissa velocity.



(a) (b)

Fig. 5: Correlation matrix for IAM-onDB dataset based on handcrafted features described above: a) before feature selection; b) after feature selection.

Fig. 6 demonstrates the distribution of projected features before and after the selection process. The projection of features on a two-dimensional plane was performed using the t-SNE algorithm. We can see an improvement in cluster separation and compactness after feature selection. Also, as a result of selection, the dimension of the feature vector decreased from 34 to 15, which significantly reduces the computational complexity of the clustering algorithms.

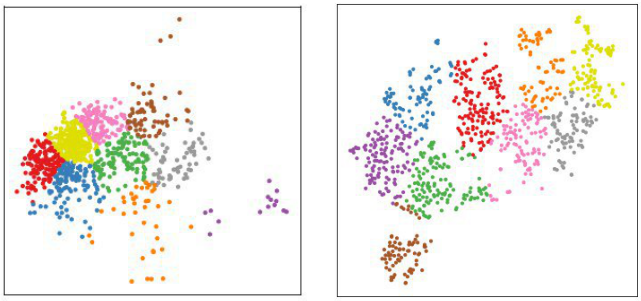


(a) (b)

Fig. 6: Feature distribution: a) before feature selection; b) after feature selection.

For further experiments, we fix the clustering algorithm as K-means to show the improvements of clustering only based on the feature selection technique. Since the handwriting styles are not separated from each other, it is obvious that the quality of the clustering will be improved while splitting into more clusters. But a large number of clusters do not satisfy the conditions of the task, and therefore we choose a *relatively* optimal value. After the elbow method application, we found out that the optimal number of clusters, in this case, is 9. The results of clustering are shown in Fig. 7.

It is necessary to determine the metrics for the evaluation of the quality of the clustering. We propose to use internal metrics that reflect the quality of the clustering only based on the information contained in the data since there is no



(a) (b)

Fig. 7: The result of clustering ($K = 9$): a) - before feature selection; b) - after feature selection.

known subdivision into classes (in this case, handwriting styles) in the existing dataset. Among the internal metrics are the following: cluster compactness; cluster separation; Dunn index; silhouette [31]–[35].

Denote by M - number of clusters, C_j - j-th cluster, $|C_j|$ - number of objects in cluster C_j, x_{ij} - the i-th object of the j-th cluster, $\overline{x}_j$ - centroid of the j-th cluster, $\overline{x}$ - average across all objects, ρ - distance function.

Cluster compactness. The idea behind this method is that the closer the objects are inside the clusters, the better the division into clusters is. Thus, it is necessary to minimize the intracluster distance, for example, the sum of squares of deviations.

$$WSS = \sum_{j=1}^{M} \sum_{i=1}^{|C_j|} (x_{ij} - \overline{x}_j)^2 . \tag{15}$$

Cluster separation. The idea is the opposite - the farther away the objects from different clusters are, the better the clustering is. Therefore, it is necessary to maximize the intercluster distance (for example, the sum of squares of deviations).

$$BSS = \sum_{j=1}^{M} |C_j| \cdot (\overline{x}_j - \overline{x})^2. \tag{16}$$

The Dunn Index [34], [35] is another metric for assessing the quality of clustering. A larger index value means better clustering, meaning that the clusters are compact and well separated from other clusters. In fact, the Dunn index shows the relationship between the minimum intercluster distance and the maximum cluster size. Denote by Δ_i the size of the i-th cluster and by $\delta(C_i, C_j)$ the distance between the clusters C_i and C_j. Then the Dunn index is determined by the following formula:

$$DI = \frac{\min_{1 \le i \le j \le M} \delta(C_i, C_j)}{\max_{1 \le k \le M} \Delta_k}. \tag{17}$$

Silhouette. The value of the silhouette shows how similar the object is to the object of its cluster compared to other clusters. The estimate for the whole cluster structure can be calculated by the following formula:

$$SIL(C) = \frac{1}{N} \sum_{C_k \in C} \sum_{x_i \in C_k} \frac{b(x_i, C_k) - a(x_i, C_k)}{\max\{a(x_i, C_k), b(x_i, C_k)\}}, \tag{18}$$

where $a(x_i, C_k)$ is the average distance from $x_i \in C_k$ to other objects in the cluster C_k (compactness), $b(x_i, C_k)$ is the average distance from $x_i \in C_k$ to objects from another cluster $C_l : k \neq l$ (separateness).

Note that $-1 \le SIL(C) \le 1$, and the clustering quality improves as x goes to 1.

TABLE I: Comparative analysis of clustering before and after feature selection

Metrics	6 clusters			9 clusters		
	Before	After	Δ	Before	After	Δ
WSS	15.342	8.858	-6.484	13.263	7.736	-5.527
BSS	14.766	7.142	-7.624	16.846	8.264	-8.528
DI	0.071	0.087	+0.016	0.059	0.085	+0.026
SIL	0.243	0.316	+0.073	0.248	0.335	+0.087
Time (ms)	1594	1252	-342	2273	1827	-446

As we can see, all the results except BSS have improved. This can be explained by the experiment described at the beginning of section IV: being a product of human activity handwriting styles cannot be clearly separated from each other, and for any two handwriting styles, there will always be an "intermediate" style. Despite this fact, the clusters became more compact, and also the overall quality of clustering has improved.

We also evaluated improvements of clustering results in terms of time consumption: after feature selection, it takes 1.3 times less of time on average to divide data into clusters.

V. Conclusion

In this work, a detailed analysis of existing approaches to feature extraction and feature selection was performed, and an expert system for the handwriting styles clustering based on the designed features was constructed.

We selected and fixed one of the most popular clustering algorithms (K-means) to demonstrate quality improvement only by selecting features.

We performed a handcrafted feature engineering, and as a result, a number of characteristics of the handwriting were designed: 12 dynamic features and 22 spatial (geometric) ones. We considered the following dynamic features (standard deviation and mean for both x- and y-axis) are: speed; acceleration; jerk. And also the spatial features: direction diagram (8 bin); curvature diagram (8 bin); frequency angles (the 8 most preferred angles); cursivity index.

We also performed feature space analysis, and feature selection in order to improve the quality of clustering. Since the values of all metrics described above are comparable, we can also see an improvement in clustering results by analyzing the obtained metrics values. Clusters became more compact (since WSS decreased), moreover, the Dunn Index and Silhouette

values are increased, which means that the clustering quality improves.

Thus, we found a group of the most effective features for handwriting style clustering, and we also can conclude that clustering based on such a subset is much more efficient.

Considering the lack of marking in the data grid, fuzzy clusters, inability to divide the samples into non-overlapping groups, and the complexity and problems present in the subject area, the above results completely satisfy the requirements set out above.

This system can be used to extract the features of handwriting in order to determine its style, which is important for designing systems for recognition or localization of text, which are quite relevant issues nowadays. It can also be used for authentication tasks. In addition, this system can be used in areas such as criminology and graphology, where much attention is paid to identifying features of handwriting, as well as being used to identify psychological abnormalities and diseases.

REFERENCES

[1] V. Vuori, "Clustering writing styles with a self-organizing map," in *Proceedings Eighth International Workshop on Frontiers in Handwriting Recognition*. IEEE, 2002, pp. 345–350.

[2] P. Sarkar and G. Nagy, "Style consistent classification of isogenous patterns," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 1, pp. 88–98, 2005.

[3] D. Zhelezniakov, V. Zaytsev, and O. Radyvonenko, "Acceleration of online recognition of 2d sequences using deep bidirectional lstm and dynamic programming," *Advances in Computational Intelligence. Lecture Notes in Computer Science*, vol. 11507, pp. 438–449, 2019.

[4] A. Bensefia, T. Paquet, and L. Heutte, "Handwriting analysis for writer verification," in *Ninth International Workshop on Frontiers in Handwriting Recognition*. IEEE, 2004, pp. 196–201.

[5] H. Hu, D. Chen, and J. Zheng, "Online handwriting signature verification based on template clustering," in *Proceedings of the 2019 3rd International Workshop on Education, Big Data and Information Technology*, 2019, pp. 129–135.

[6] S. Rashidi, A. Fallah, and F. Towhidkhah, "Feature extraction based dct on dynamic signature verification," *Scientia Iranica*, vol. 19, no. 6, pp. 1810–1819, 2012.

[7] S. K. Chan, Y. H. Tay, and C. Viard-Gaudin, "Online text independent writer identification using character prototypes distribution," in *2007 6th International Conference on Information, Communications & Signal Processing*. IEEE, 2007, pp. 1–5.

[8] V. Volkova, I. Deriuga, V. Osadchyi, and O. Radyvonenko, "Improvement of character segmentation using recurrent neural networks and dynamic programming," in *2018 IEEE Second International Conference on Data Stream Mining & Processing (DSMP)*. IEEE, 2018, pp. 218–222.

[9] I. Degtyarenko, O. Radyvonenko, K. Bokhan, and V. Khomenko, "Text/shape classifier for mobile applications with handwriting input," *International Journal on Document Analysis and Recognition (IJDAR)*, vol. 19, no. 4, pp. 369–379, 2016.

[10] V. Khomenko, A. Volkoviy, I. Degtyarenko, and O. Radyvonenko, "Handwriting text/non-text classification on mobile device," in *The Fourth International Conference on Artificial Intelligence and Pattern Recognition (AIPR)*, 2017, pp. 42–49.

[11] J. Han, G. Chernyshov, D. Zheng, P. Gao, T. Narumi, K. Wolf, and K. Kunze, "Sentiment pen: Recognizing emotional context based on handwriting features," in *Proceedings of the 10th Augmented Human International Conference 2019*, 2019, pp. 1–8.

[12] A. Graves, "Generating sequences with recurrent neural networks," *arXiv preprint arXiv:1308.0850*, 2013.

[13] J. Walton, "Handwriting changes due to aging and parkinson's syndrome," *Forensic science international*, vol. 88, no. 3, pp. 197–214, 1997.

[14] P. Drotár, J. Mekyska, I. Rektorová, L. Masarová, Z. Smékal, and M. Faundez-Zanuy, "Evaluation of handwriting kinematics and pressure for differential diagnosis of parkinson's disease," *Artificial intelligence in Medicine*, vol. 67, pp. 39–46, 2016.

[15] O. Hilton, "Influence of age and illness on handwriting: Identification problems," *Forensic Science*, vol. 9, pp. 161–172, 1977.

[16] J.-P. Crettez, "A set of handwriting families: style recognition," in *Proceedings of 3rd International Conference on Document Analysis and Recognition*, vol. 1. IEEE, 1995, pp. 489–494.

[17] G. Marzinotto, J. C. Rosales, M. A. El-Yacoubi, S. Garcia-Salicetti, C. Kahindo, H. Kerhervé, V. Cristancho-Lacroix, and A.-S. Rigaud, "Age-related evolution patterns in online handwriting," *Computational and mathematical methods in medicine*, vol. 2016, 2016.

[18] F. Holekian, M. G. Lavasani, and Y. Madani, "Handwriting analysis: the role of age and education," *International Journal of Modern Management and Foresight*, vol. 1, no. 6, pp. 208–221, 2014.

[19] S. Al Maadeed and A. Hassaine, "Automatic prediction of age, gender, and nationality in offline handwriting," *EURASIP Journal on Image and Video Processing*, vol. 2014, no. 1, p. 10, 2014.

[20] S.-H. Cha and S. N. Srihari, "A priori algorithm for sub-category classification analysis of handwriting," in *Proceedings of Sixth International Conference on Document Analysis and Recognition*. IEEE, 2001, pp. 1022–1025.

[21] R. K. Powalka, N. Sherkat, and R. J. Whitrow, "Recognizer characterisation for combining handwriting recognition results at word level," in *Proceedings of 3rd International Conference on Document Analysis and Recognition*, vol. 1. IEEE, 1995, pp. 68–73.

[22] G. Marzinotto, J. C. Rosales, M. A. El-Yacoubi, and S. Garcia-Salicetti, "Age and gender characterization through a two layer clustering of online handwriting," in *International Conference on Advanced Concepts for Intelligent Vision Systems*. Springer, 2015, pp. 428–439.

[23] A. Bharath, V. Deepu, and S. Madhvanath, "An approach to identify unique styles in online handwriting recognition," in *Eighth International Conference on Document Analysis and Recognition (ICDAR'05)*. IEEE, 2005, pp. 775–778.

[24] M. Liwicki and H. Bunke, "Iam-ondb-an on-line english sentence database acquired from handwritten text on a whiteboard," in *Eighth International Conference on Document Analysis and Recognition (ICDAR'05)*. IEEE, 2005, pp. 956–961.

[25] R. Plamondon and S. N. Srihari, "Online and off-line handwriting recognition: a comprehensive survey," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 22, no. 1, pp. 63–84, 2000.

[26] I. Guyon, P. Albrecht, Y. Le Cun, J. Denker, and W. Hubbard, "Design of a neural network character recognizer for a touch terminal," *Pattern recognition*, vol. 24, no. 2, pp. 105–119, 1991.

[27] A. Jameel, "Perspectives of pattern recognition in handwritten character recognition." 1996.

[28] L. Vuurpijl and L. Schomaker, "Coarse writing-style clustering based on simple stroke-related features," *Progress in Handwriting Recognition*, pp. 37–44, 1996.

[29] M. A. Hall, "Correlation-based feature selection for machine learning," 1999.

[30] B. Senliol, G. Gulgezen, L. Yu, and Z. Cataltepe, "Fast correlation based filter (fcbf) with a different search strategy," in *2008 23rd international symposium on computer and information sciences*. IEEE, 2008, pp. 1–4.

[31] N. R. Pal and J. Biswas, "Cluster validation using graph theoretic concepts," *Pattern Recognition*, vol. 30, no. 6, pp. 847–857, 1997.

[32] M. Halkidi, Y. Batistakis, and M. Vazirgiannis, "On clustering validation techniques," *Journal of intelligent information systems*, vol. 17, no. 2-3, pp. 107–145, 2001.

[33] D. L. Davies and D. W. Bouldin, "A cluster separation measure," *IEEE transactions on pattern analysis and machine intelligence*, no. 2, pp. 224–227, 1979.

[34] J. C. Dunn, "A fuzzy relative of the isodata process and its use in detecting compact well-separated clusters," 1973.

[35] ——, "Well-separated clusters and optimal fuzzy partitions," *Journal of cybernetics*, vol. 4, no. 1, pp. 95–104, 1974.

IEEE Third International Conference on Data Stream Mining & Processing
August 21-25, 2020, Lviv, Ukraine

The Neural Network Technologies Effectiveness for Face Detection

Kirill Smelyakov

Department of Electronic Computers
Kharkiv National University of Radio Electronics
Ukraine, Kharkiv, Nauky Ave, 14
kyrylo.smelyakov@nure.ua

Anastasia Chupryna

Department of Software Engineering
Kharkiv National University of Radio Electronics
Ukraine, Kharkiv, Nauky Ave, 14
anastasiya.chupryna@nure.ua

Oleksandr Bohomolov

Department of Software Engineering
Kharkiv National University of Radio Electronics
Ukraine, Kharkiv, Nauky Ave, 14
oleksandr.bohomolov@nure.ua

Igor Ruban

Department of Electronic Computers
Kharkiv National University of Radio Electronics
Ukraine, Kharkiv, Nauky Ave, 14
ihor.ruban@nure.ua

***Abstract* – The paper is devoted to estimation of the effectiveness of the use of modern convolutional neural networks for face detection. On standard open datasets, learning of neural networks and comparison of the effectiveness of their functioning are carried out. Conclusions are drawn regarding the practical application of the neural networks for detecting faces on digital photographs.**

Keywords – Convolutional Neural Network, Face Detection, Architecture, Model, Effectiveness.

I. Introduction

The problem of face recognition has become very important in the field of biometric identification. Facial detection in digital images has many applications in solving real-life problems, including entertainment, healthcare, security, etc. [1]. To this day, there are a lot of approaches and models for solving these tasks. Although older algorithms, such as the Viola-Jones algorithm perform facial recognition tasks fast and accurately in close to optimal conditions, their performance diminishes when input images are distorted, too dark or too bright, or when faces appear in the image under different angles [2]. The development of deep learning and deep convolutional neural networks (NN) allowed researches to invent complex neural network-based machine learning algorithms which, when properly trained, perform much better and reliably on images with faces in different environments, angles and lightning conditions [2]. Also, each of the emerged neural network-based algorithms has its features and characteristics which determine its efficiency when applied to different types of applications.

The problem of facial recognition is a subset of a wider problem, object detection [3, 4]. There are two families of neural network-based algorithms that are used most often for object detection: R-CNN algorithms [5], that use Regional Proposal Networks (RPNs) to identify areas in the image where an object may be, and YOLO (an acronym You Only Look Once), which is faster, however less accurate, algorithm than R-CNN [6]. There are other approaches and algorithms to object detection, however, the two approaches above are the ones that saw wide usage and application in the industry. Each of these approaches is also incrementally enhanced to improve accuracy and address their respective shortcomings.

There is also a NN that was developed specifically to solve the facial recognition problem, MTCNN (Multi-task Cascaded Convolutional Network). This model consists of three separate NN models, the P-Net, R-Net, and O-Net, which determine not only the bounding box around the face but also the location of eyes, nose, and mouth [7].

In this paper, we compare the productivity of fast NN models: YOLOv3 and MTCNN. We use transfer learning to train the YOLOv3 model on the Wider Face dataset using pre-trained weights and compare it with the fully trained MTCNN model. For comparison, we use various object detection metrics, such as Intersection over Union (IOU), precision, recall, and others, which are described in the following sections. We do not include R-CNN family algorithms in the comparison because the training of these algorithms is computationally expensive and we lack computational resources to fully train them. GPU on which the experiment is run only has 6 gigabytes of memory, which is not enough to train R-CNN family models.

The dataset we are using is the Wider Face dataset [8]. This dataset was assembled to benchmark different face detection algorithms. It consists of images of people in different situations, the faces on these images are located in different areas, angles and lighting conditions.

The reliable performance of facial recognition algorithms is important for the functioning of modern informational and communicational systems [9-12].

II. Neural Network Models

So, in this paper, we compare two NN architectures for face detection – YOLOv3 [13] and MTCNN [7].

978-1-7281-3214-3/20/$31.00 ©2020 IEEE

YOLOv3 is the further development of the original YOLO architecture. Original YOLO was developed with speed and efficiency in mind, because at that time the state-of-the-art model for object detection, R-CNN, was too slow in training and inference to use it in such tasks as real-time detection. However, original YOLO, while being faster and more resource-efficient, was less accurate than R-CNN. Because of this, further iterations of YOLO development improved accuracy but became significantly slower.

Before YOLO, R-CNN models used a pipeline of two different algorithms, one for generating regional proposals, i.e. areas where an object might be, and the other one to determine whether the proposed region includes an object the network is trying to find.

Original YOLO introduced the way to unify two of these algorithms into a single neural network [6]. It was done by dividing the image into a square grid $S \times S$. Each cell is responsible for predicting B bounding boxes and a confidence score for each of those boxes.

If there is no object in the cell, the confidence score should be zero. Otherwise, it should be equal to the intersection over union (IOU) between the predicted box and the ground truth. Each bounding box consists of 5 predictions: x, y, w, h and a confidence score, which was described earlier.

The x and y coordinates are relative to the cell, and w and h are relative to the width and height of the image respectively. Each cell also predicts C class probabilities, where C equals to the number of object classes the network tries to detect.

Original YOLO network architecture consists of 24 convolutional layers and 2 fully connected layers. Next improvement, YOLO v2, introduced a custom architecture darknet-19, a 19-layer network supplemented with additional 11 levels for object detection. However, even with these additions, the improved architecture still lacked important elements that are widely used in modern state-of-the-art neural networks for computer visions, such as residual blocks, skip-connections and upsampling. The next generation of YOLO, YOLOv3, incorporates all of these.

YOLOv3 uses a variant of Darknet architecture (Fig. 1), with a 53-layered network, together with 53 more layers for detection. This greatly improves the accuracy of the trained models, however, this also substantially increases the time needed for training and inference. One of the most interesting features of YOLOv3 is that it makes detection on 3 different scales. Because of this, YOLOv3 detects small objects much accurately than older versions of YOLO.

YOLO is a neural network architecture for performing general object detection, i.e., detecting as many different object classes as needed. However, in this paper, we compare different neural network architectures for face detection. For this task, there is a specialized neural network architecture, MTCNN [7].

MTCNN stands for Multi-task Cascaded Convolutional Network (MTCNN). Instead of being a single neural network model as YOLO, MTCNN includes three neural networks, called P-Net, R-Net and O-Net (Fig. 2). It returns not only the bounding boxes of detected faces but also 5 facial landmarks: 2 for eyes, 1 for a nose and 2 for a mouth.

MTCNN works the following way. Firstly, it creates an image pyramid of multiple scaled versions of an input image.

	Type	Filters	Size	Output
	Convolutional	32	3 × 3	256 × 256
	Convolutional	64	3 × 3 / 2	128 × 128
1×	Convolutional	32	1 × 1	
	Convolutional	64	3 × 3	
	Residual			128 × 128
	Convolutional	128	3 × 3 / 2	64 × 64
2×	Convolutional	64	1 × 1	
	Convolutional	128	3 × 3	
	Residual			64 × 64
	Convolutional	256	3 × 3 / 2	32 × 32
8×	Convolutional	128	1 × 1	
	Convolutional	256	3 × 3	
	Residual			32 × 32
	Convolutional	512	3 × 3 / 2	16 × 16
8×	Convolutional	256	1 × 1	
	Convolutional	512	3 × 3	
	Residual			16 × 16
	Convolutional	1024	3 × 3 / 2	8 × 8
4×	Convolutional	512	1 × 1	
	Convolutional	1024	3 × 3	
	Residual			8 × 8
	Avgpool		Global	
	Connected		1000	
	Softmax			

Fig. 1. Darknet 53 architecture [13]

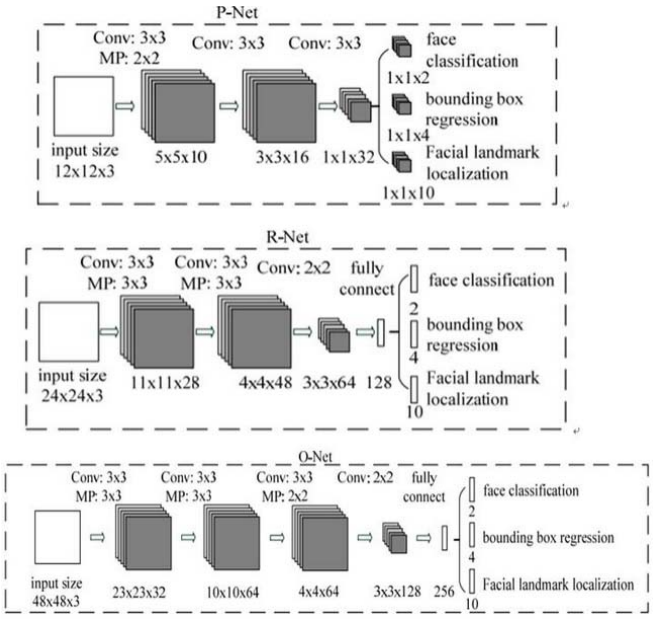


Fig. 2. MTCNN model architecture [7]

Then for of scaled images, it applies a sliding 12x12 kernel with a stride of two. The portion of an image under the kernel is passed to the first neural network, the Proposal Network (P-Net). P-Net returns the bounding box if it finds the face. It also returns the confidence score for each of the bounding boxes.

After that, all P-Net outputs are collected and the bounding boxes with low confidence scores are removed. Then all the coordinates of bounding boxes are being unscaled to match the original input size. However, at this stage, there is still a lot of bounding boxes, so the Non-Maximum Suppression (NMS) method is applied to remove redundant bounding boxes and merge several closely located bounding boxes into one.

After that, the remaining bounding boxes are fed into another convolutional neural network, the Refine Network (R-Net). This network rejects a large number of region proposals from remaining candidates and performs further calibration of bounding boxes. After that, NMS is applied to the result once again, rejecting false bounding boxes.

The final stage is similar to the second stage. The output of the second stage is fed to the last network, the Output Network (O-Net). This time, in addition to the bounding boxes, this network outputs also locations of facial landmarks.

The neural networks mentioned above perform 3 tasks: face / not face classification, bounding box detection and facial landmark detection. For each of these tasks, there is a corresponding loss function

$$L_i^{det} = -\left(y_i^{det}\log(p_i) + \left(1 - y_i^{det}\right)(1 - \log(p_i))\right). \quad (1)$$

In the formula above p_i is the probability with which the input is a face, and y_i^{det} is the ground truth label. This is a binary cross-entropy loss

$$L_i^{box} = \left\|\hat{y}_i^{box} - y_i^{box}\right\|_2^2. \quad (2)$$

The bounding box determination is a regression problem, and for the loss function, we use simple Euclidean loss. In formula 2, $\hat{y}_i^{box}$ is the ground truth bounding box, and y_i^{box} is the bounding box which is the output of the neural network

$$L_i^{landmark} = \left\|\hat{y}_i^{landmark} - y_i^{landmark}\right\|_2^2. \quad (3)$$

The facial landmark detection is also the regression problem, so the loss function for facial landmark detection (3) is the same as the formula (2).

Both YOLO and MTCNN models are capable of doing face detection. The main advantage of YOLO is that together with face detection it can perform detection of any other types of objects. This is useful if, for instance, you need to detect faces together with road signs or automobiles. It is also useful if you have to detect different types of faces, for instance, faces with hats or sunglasses on. The main drawback of YOLO in comparison with MTCNN is that YOLO requires much more computational resources to be trained efficiently.

On the other hand, MTCNN can detect only faces. However, it also performs facial landmark detection, which YOLO doesn't do. Although MTCNN uses 3 neural networks instead of a single one in YOLO, its neural networks are less complex and therefore require much less time and resources to train. Also, MTCNN already ships with pre-trained weights, so unless the problem requires fine-tuning the weights to the specific dataset, no additional training is required.

The advantages and disadvantages of both models are listed above. The rest of the paper will describe the experiment and its results that show how both networks compare in terms of face detection efficiency.

III. Experimental Results Analysis

For models comparison (YOLO and MTCNN), we will use the Wider Face dataset [8]. It is a face detection benchmark dataset which consists of 32,203 selected images from publicly available WIDER image dataset. The images in the Wider Face dataset are diverse. As well as portrait images, the dataset also contains group images, images from sporting events, etc. The diversity of images helps us to compare the performance of both algorithms under less-than-optimal conditions.

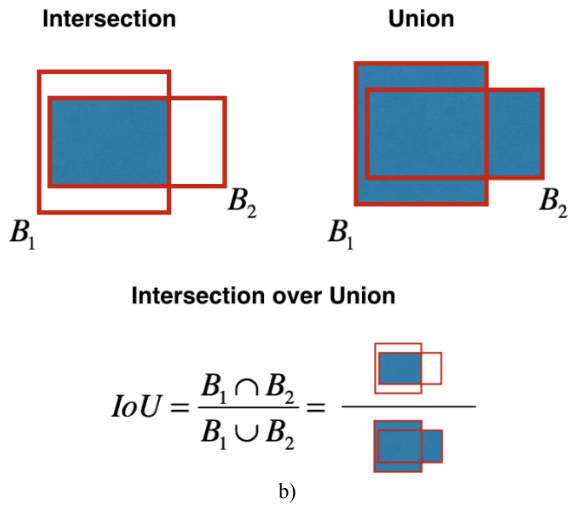
Due to the hardware limitations and lack of computational resources, in the experiment, we will use only the subset of the dataset. We will use first 10,000 images for training and 2,500 images for inference. The training and validation datasets are subsets of respective Wider Face datasets for training and validation with images taken from the beginning of datasets. The small size of the dataset will not affect the experiment result significantly, because we will apply transfer learning by using already pre-trained YOLO weights. We are not going to train MTCNN because this model is created specifically for face detection and already has fully trained weights available.

We are using YOLOv3 and MTCNN model implementations that use TensorFlow and Keras library [14-16]. For comparison of two algorithms, we use the averaged precision as a metric and precision/recall curves [17].

To calculate precision and recall, we must determine what the true positive, false positive and false negative are in our task. For this, we have to use IOU measure. IOU is a measure that evaluates the overlap between two bounding boxes. It requires a ground truth bounding box B_{gt} and a predicted bounding box B_p. The formula for IOU (4) is listed below

$$IOU = \frac{area(B_p \cap B_{gt})}{area(B_p \cup B_{gt})}. \quad (4)$$

In simpler terms, IOU is a fraction between an area of overlap and an area of union (Fig. 3). We can use IOU to determine if detection is a true positive, false positive, or a false negative. If IOU is bigger than a threshold, the detection is correct and therefore is a true positive. If IOU is less than the threshold, the detection is a false positive. The detection is a false negative if no object was detected. Using true positives, false positives and false negatives, we can calculate the precision and the recall [17].



b)

Fig. 3. Intersection Over Union [17, 18]

Precision is the ability of the model to identify only the relevant objects. It is the percentage of correct positive predictions

$$Precision = \frac{TP}{TP+FP} = \frac{TP}{all\ detections}. \quad (5)$$

Recall is the ability of the model to find all the relevant cases (all ground truth bounding boxes). It is the percentage of true positive detected among all relevant ground truths

$$Recall = \frac{TP}{TP+FN} = \frac{TP}{all\ ground\ truths}. \quad (6)$$

To illustrate the performance of the two models we use Precision x Recall curves. It is a good way to evaluate the performance of an object detector as the confidence is changed by plotting a curve. The model can be considered good if precision stays high as recall increases. To compare the two models, we use the Average Precision metrics. We get average precision by calculating the area under the curve (AUC) of the Precision x Recall curve [17].

We calculate metrics for both training and validation data. For YOLO this can demonstrate if the network overfits during training. For MTCNN, it is irrelevant, because we use an already fully trained MTCNN model.

Below we list the experiment outcomes (Tab.1), and the precision x recall curves for each experiment (Fig.4-Fig.7).

TABLE I.
RESULTS OF FEATURE POINTS DETECTING FOR FACE WITH TRANSFORMATION

	YOLO	MTCNN
Training dataset AP	32.55%	41.71%
Validation dataset AP	30.88%	38.35%

Below we list the inference time that was measured during the experiment (Tab.2). Experiment was run on a computer with 6-core CPU, 16 gigabytes of RAM and GTX 1060 GPU with 6 gigabytes of memory.

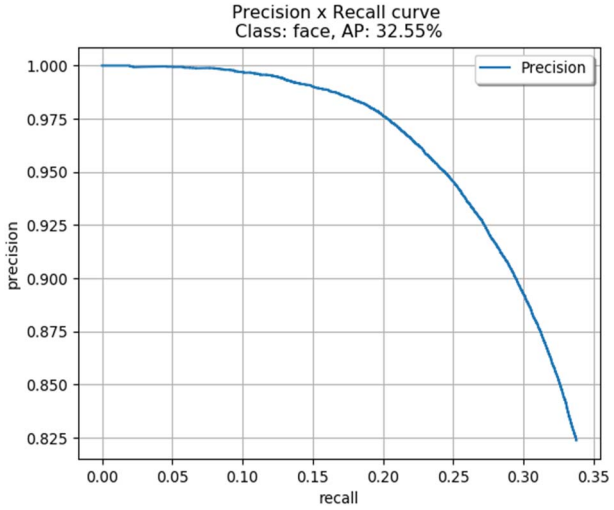


Fig. 4. YOLO training dataset precision x recall curve

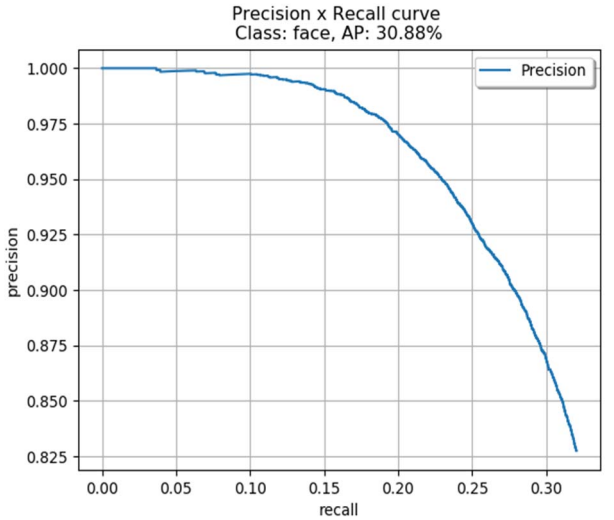


Fig. 5. YOLO validation dataset precision x recall curve

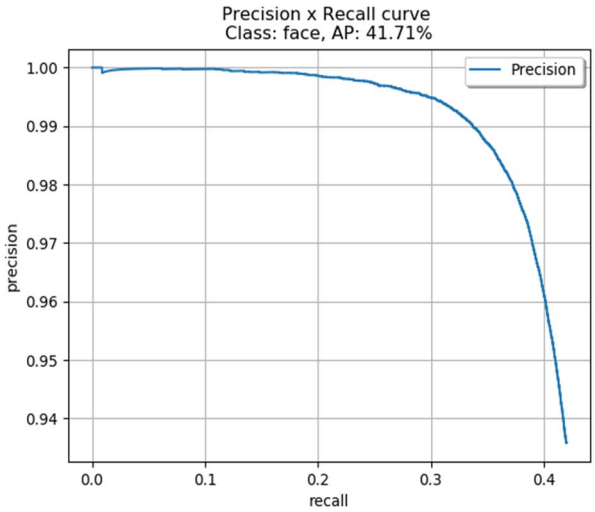


Fig. 6. MTCNN training dataset precision x recall curve

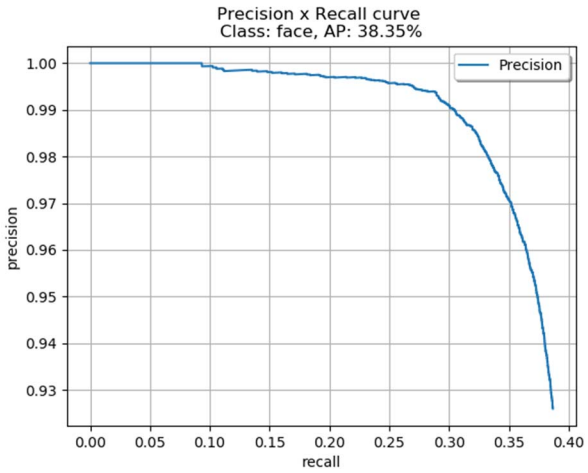


Fig. 7. MTCNN validation dataset precision x recall curve

TABLE II.
RESULTS OF FEATURE POINTS DETECTING FOR FACE WITH TRANSFORMATION

	YOLO	MTCNN
Training dataset (10,000 examples), sec	978.45	2475
Validation dataset (2,500 examples), sec	269.77	732.59
Time per image, sec	0.1029	0.2703

As we can say from the experiment outcomes and precision x recall curves listed above, the MTCNN model performs better and more accurately than the YOLO model. MTCNN model was developed specifically for face detection problems, while YOLO was developed as a general object detection algorithm. MTCNN neural networks are also much less complex than YOLO neural network.

Despite the fact that neural networks in MTCNN models are much less complex than YOLO neural network, in the inference the MTCNN model is about 3 times slower than the YOLO model. This is because MTCNN includes 3 neural networks and uses the Non-Maximum Suppression method.

The final conclusion we can make is that in most cases MTCNN should be used for face detection tasks. YOLO should be used when you have to detect other objects rather than just faces or your dataset is very specific and you have to use transfer learning to adjust the weights. Also, YOLO is preferable when fast inference speed is required. In other cases, MTCNN is a good solution for face detection tasks.

CONCLUSION

In this work, we compared two neural network-based models applied to the task of face detection, YOLO and MTCNN. We did not include another popular neural network-based model Faster R-CNN into the comparison because we lacked computational resources to train it properly. YOLO is a network for general object detection, it can detect any objects, not just faces. MTCNN is a model designed specifically for face detection. YOLOv3 is the latest iterative improvement to the YOLO model. It contains a single but complex neural network that requires a lot of computational resources to train. MTCNN contains 3 much simpler neural networks, but the MTCNN algorithm is complicated and performs a lot of transformations on input data.

The experiment was performed using the Wider Face dataset. YOLO model was trained using pre-trained weights. MTCNN model was used with fully pre-trained weights. Average precision and precision x recall curves were used as comparison metrics. The outcomes of the experiment show that MTCNN performs better on the task of face detection. In most cases, MTCNN should be used for face detection tasks. YOLO should be used if there is a requirement to detect not only faces but other objects as well, fast inference speed is required, or if the dataset is very specific and transfer learning has to be used to adjust the weights. The MTCNN model also detects facial landmarks. This is useful for other advanced facial recognition tasks, such as emotion recognition.

REFERENCES

[1] Asit Kumar Datta, Madhura Datta, Pradipta Kumar Banerjee Face Detection and Recognition. CRC Press, 2015. 352p.
[2] G. Blokdyk Facial Recognition A Complete Guide. 5STARCooks, 2018. 126p.
[3] Smelyakov K., Shupyliuk M., Martovytskyi V., Tovchyrechko D., Ponomarenko O. Efficiency of Image Convolution. 2019 IEEE 8th International Conference on Advanced Optoelectronics and Lasers (CAOL), 6-8 Sept. 2019, Sozopol, Bulgaria. pp. 578-583.
[4] A. Arsenov, I. Ruban, K. Smelyakov, A. Chupryna Evolution of Convolutional Neural Network Architecture in Image Classification Problems. XVIII International Scientific and Practical Conference on Information Technologies and Security (ITS 2018). CEUR Workshop Processing. – Kyiv, Ukraine, November 27, 2018, pp. 35-45.
[5] R-CNN, Fast R-CNN, Faster R-CNN, YOLO: https://towardsdatascience.com/r-cnn-fast-r-cnn-faster-r-cnn-yolo-object-detection-algorithms-36d53571365e.
[6] J. Redmon, S. Divvala, R. Girshick, A. Farhadi You Only Look Once: Unified, Real-Time Object Detection. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 9 May 2016, 10p.
[7] Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, Yu Qiao Joint Face Detection and Alignment using Multi-task Cascaded Convolutional Networks // Signal Processing Let. Vol. 23 (10), 2016. pp. 1499-1503.
[8] WIDER FACE: A Face Detection Benchmark: http://shuoyang1213.me/WIDERFACE.
[9] V. Mukhin, Y. Romanenkov, J. Bilokin, A. Rohovyi, A. Kharazii, V. Kosenko, N. Kosenko, J. Su. The method of variant synthesis of information and communication network structures on the basis of the graph and set-theoretical models. International Journal of Intelligent Systems and Applications (IJISA). 2017. vol. 9 (11). pp. 42-51.
[10] Boiko J., Tolubko V., Barabash O., Eromenko O., Havrylko Ye. Signal processing with frequency and phase shift keying modulation in telecommunications TELKOMNIKA Telecommunication, Computing, Electronics and Control. 2019. Vol. 17 (4), pp. 2025-2038.
[11] O. Yeremenko, T. Lebedenko, T. Vavenko, and M. Semenyaka, Investigation of queue utilization on network routers by the use of dynamic models. Second International Scientific-Practical Conference Problems of Infocommunications Science and Technology (PIC S&T), 13-15 October, 2015, pp. 46-49.
[12] V. Tkachov, V. Tokariev, Y. Dukh,, and V. Volotka, Method of Data Collection in Wireless Sensor Networks Using Flying Ad Hoc Network International Scientific-Practical Conference Problems of Infocommunications. Science and Technology (PIC S&T), 9-12 October, 2018, pp. 197-201.
[13] J. Redmon, A. Farhadi YOLOv3: An Incremental Improvement. ArXiv 2018
[14] TensorFlow: https://www.tensorflow.org.
[15] YOLOv3 implementation: https://github.com/qqwweee/keras-yolo3.
[16] MTCNN implementation: https://github.com/ipazc/mtcnn.
[17] Metrics for object detection https://github.com/rafaelpadilla/Object-Detection-Metrics.
[18] YOLOv3 implementation: https://medium.com/analytics-vidhya/yolo-v3-theory-explained-33100f6d193.

IEEE Third International Conference on Data Stream Mining & Processing
August 21-25, 2020, Lviv, Ukraine

Surface Defects Detection by Clustering and Rotating Image Analysis

Roman Melnyk
Software Department
Lviv Polytechnic National University
Lviv, Ukraine
ramelnyk@polynet.lviv.ua

Ruslan Tusnytskyy
Software Department
Lviv Polytechnic National University
Lviv, Ukraine
ruslan.tushnytskyy@gmail.com

Yurii Havrylko
Software Department
Lviv Polytechnic National University
Lviv, Ukraine
yurii.havrylko@gmail.com

***Abstract* — In this paper the K-means algorithm was used to obtain coordinates of defects by clustering distributed cumulative histogram image. Mean intensity functions of rotating sample and its distributed cumulative histogram image were used to determine coordinates of scratches on the surface image.**

Keywords—image, intensity, defects, clustering, scratch, distributed cumulative histogra, image rotation,variation.

I. INTRODUCTION

There is great amount of published works devoted to various surface defects detection. They consider different approaches as well as different types of defects and images. Published papers contain description of theoretical and industrial systems. In our introduction some typical works are chosen to present existing problems and to underline methods to solve them. In the paper [1] defect detection methods are categorized into seven classes as structural, statistical, spectral, model-based, learning, hybrid and comparison studies. The main criteria are as the accuracy, the computational cost, reliability, rotating/scaling invariant, ability to operate and noise sensitivity. These criteria are taken into account by the majority of researches. The publication [2] presents a complete framework aimed to nondestructive inspection of composite materials: view enhancement, feature extraction, classification and defect detection. These manipulations are also common and the aims of different proposed and planned methodologies to reduce the number of false positives and to detect defects.

A detection method to detect the interesting defective objects for silicon steel based on saliency linear scanning morphology is proposed in the published work [3]. In the proposed method, visual saliency extraction is employed to suppress the clutter background. and a saliency map is obtained for the purpose of highlighting the potential defective objects. The strategy to block background is widely used but tactics to reach it differ between themselves essentially.

In the paper [4] a defect classification system working on the production line is considered. The classifier is constituted of a two-levels hierarchical architecture: a set of adequate features are extracted from the image first and the defect type is identified from them. Statistical analysis allows to reduce the number of features leaving only significant ones. Here we have an approach to decompose the complex problem into two more simpler ones.

In the article [5] an approach for the automatic inspection of defects in randomly textured surfaces is presented. This task arises in fabric, wallpaper, castings, leather and many industrial and investigational materials. It is based on an image reconstruction scheme with the Fourier transform. The difficult defect detection problem in textured images is changed by a simple thresholding problem of segmentation of images without textures.

Artificial intelligence and neural networks are widely used for defects detection. As a rule, they need solid computer calculation but demonstrate good results. In the paper [6] a deep learning-based method for pixel-wise surface defect segmentation is proposed. The algorithm has three steps: a segmentation stage, a detection stage, and a matting stage. In the final matting stage, a filter is utilized to refine the contours of the defect areas to reflect the real defective region. All mention works are characterized by complicated procedures. They demonstrate good experimental results for a variety of real images. Publications show the efficacy of the proposed methods.

In common the majority of the above-mentioned approaches are quite time-consuming to be realized. In our paper we present very simple approaches based on the clustering algorithm, distributed cumulative histogram and transformation of images during a rotation them around the center of gravity.

II. THE CLUSTERING K-MEANS ALGORITHM FOR DEFECTS DETECTION

Firstly, we present the simple formulas of mean intensity of pixels in rows or columns as follows:

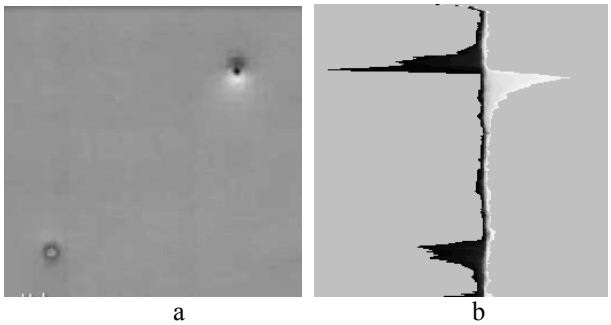
$$I_y(j) = 1/H \sum_{i=1}^{H} I(i,j), \quad 1 \le j \le W, \tag{1}$$

$$I_x(i) = 1/W \sum_{j=1}^{W} I(i,j), \quad 1 \le i \le H. \tag{2}$$

where $I(i,j)$ is pixel intensity in i-row and j-column ($1 \le i \le H$, $1 \le j \le W$).

For an image analysis we choose a model of an image of the distributed cumulated histogram [7]. The last one is more convenient because colors of defects on the DCH image have greater specific weight than colors of defects on the original surface image. It could be visually confirmed by comparison of these images in Fig.1 as well as comparison of mean intensity function in rows (formulas 1,2) of these images by plots given in Fig. 2.

978-1-7281-3214-3/20/$31.00 ©2020 IEEE



a b

Fig. 1. Image with defects (a) and DCH image (b)

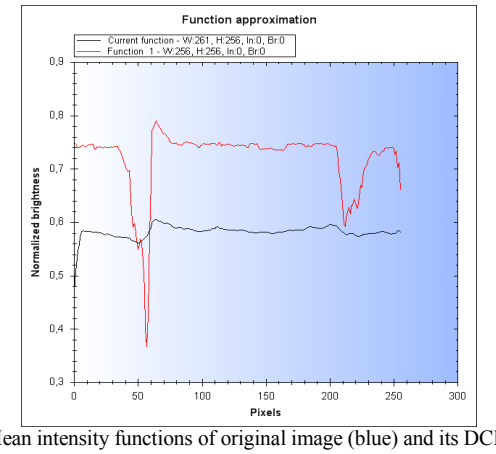


Fig. 2. Mean intensity functions of original image (blue) and its DCH (red)

We cover the DCH image by rectangles (Fig. 3). A width of each of them equals 256 that is a height of the DCH image. A height of the rectangle is being accepted as a control parameter taking value, for example, from the interval of 1-10.

For every rectangle mean intensity value is being calculated by which they as elementary leaves to form clusters are accepted. Clusters are being formed by the K-means algorithm.

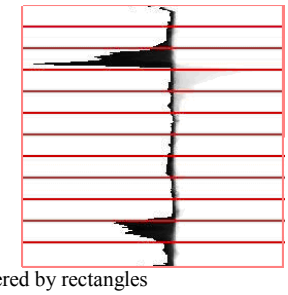
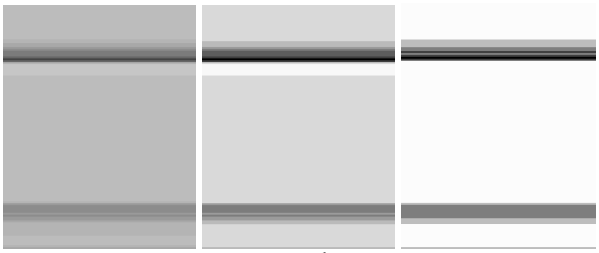


Fig.3. DCH image covered by rectangles

By clustering the DCH image we obtain different results depending from a number of resulting clusters and colors assigned to these clusters. For example, if a final number of clusters is 9 and their colors are native that is corresponding to the input image, we have the image in Fig. 4 a. For the same number of clusters and expanding the color interval to mark them by the following rule

$$I_{\text{int}} = \{\max_{x,y} I(x,y)\} - \{\min_{x,y} I(x,y)\},\ x,y \in Z \quad (3)$$

we obtain the inage given in Fig.4 b. For the same expanded interval and a number of resulting clusters 4 we obtain the image in Fig.4 c.



a b c

Fig. 4. Clustered DCH mages: native colors an 9 clusters (a), expanded interval and 9 clusters (b), expanded interval and 4 clusters (c)

To chose the best result we measure quantitave characteristics of resulting images calculating the mean intensity functin in rows for all of them (Fig.5).

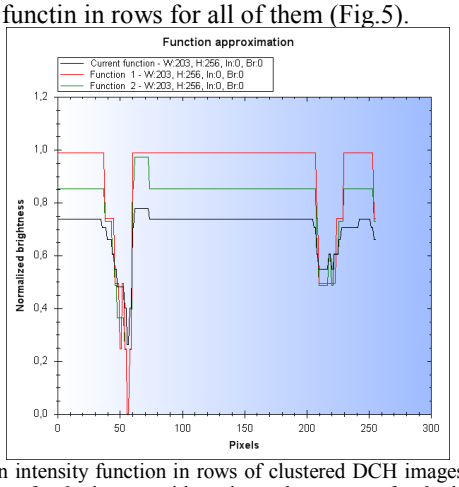


Fig.5. Mean intensity function in rows of clustered DCH images: red for 4 clusters; blue for 9 clusters with native colors; green for 9 clusters with expanded interval

From Fig.5 we see that the best result is for 9 clusters and expanded interval. Dark and light bands are comparatively distinct and remarkable. Their coordinates allow to determine positions of texture irregularities by one coordinate axe.

Positions of defects by the second coordinate axe are being calculated similarly by applying the approach in the appropriate direction of the cut fragment given in Fig.6.

Fig. 6. Cropped fragment and its clustered DCH

III. DEFECTOSCOPE BY ROTATING IMAGE AND DISTRIBUTED CUMULATIVE HISTOGRAM

Here we consider the instrument helping to detect defects of scratching having different intensity and directions.

In Fig. 7 the textured surface has some scratches with colors of poor visibility and intensity differing from those of background [8].

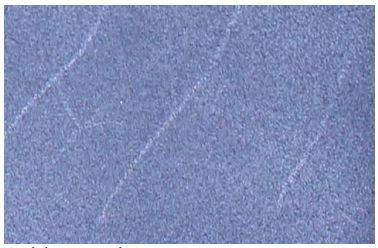


Fig. 7. Texture with scratches

The specific weight of scratch pixels intensity is small comparatively to texture pixels if we measure intensity in rows or columns of the image matrix. The same fact we observe when using distributed cumulative histogram for rows or columns. The picture changes when viewing is being realized from other angles. Pixels of scratches are being accumulated and are being displayed by mentioned above functions. So, to organize defect detection we need observation of the image from different directions. It is performed in so-called visual defectoscope. In this program instead of moving a camera around the image (Fig. 8) the sample image is being rotated inside the circle.

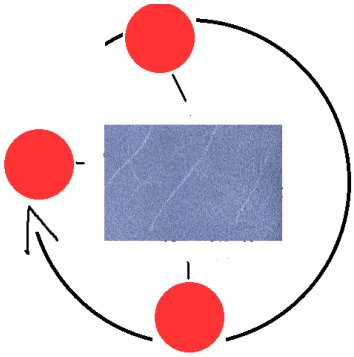
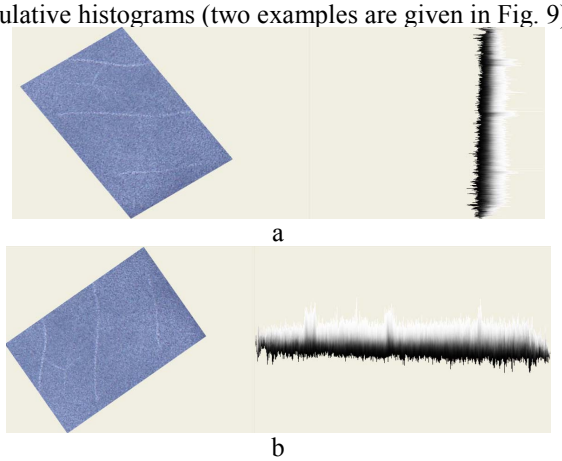


Fig. 8. Scanning the sample image to get cumulative histograms.

Rotation of the original image is carried out by applying combinations of affine transformations and changing the size of the canvas image. The rotation is automatically applied to displaying the original image in the program window [7].

All ideas and formulas are realized in software allowing to rotate the sample image and to build the distributed cumulative histograms (two examples are given in Fig. 9).

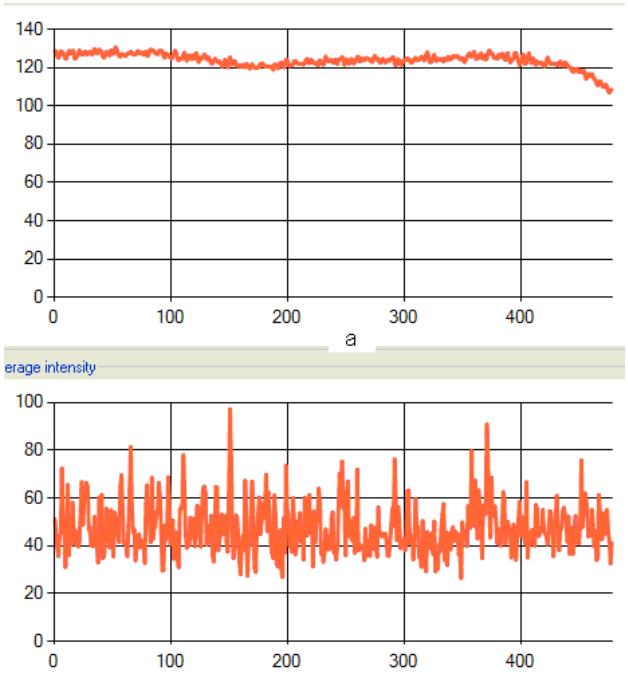


a

b

Fig. 9. Image scanning from OY (a) and OX (b) axes by different angles and the corresponding DCH images

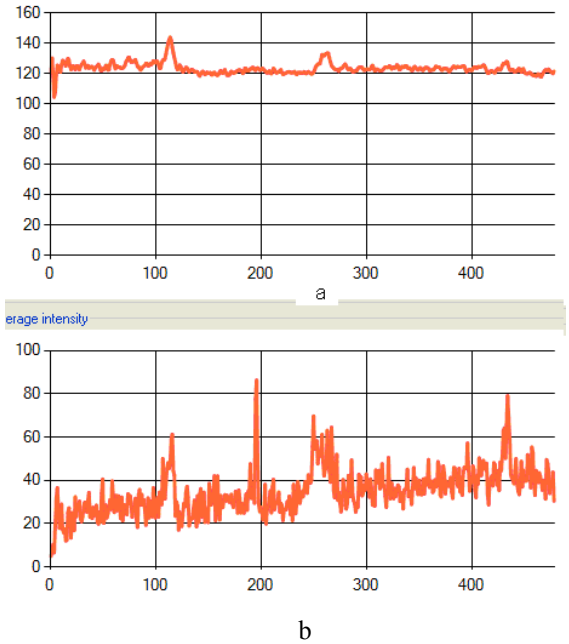
For obtained sample images and their DCH images the mean intensity functions are being calculated using formulas. (1, 2) Their plots for starting position with the angle with zero value are given in Fig.10.



b

Fig.10 Mean intensity function in rows of the sample image(a) and the DCH image (b) for zero angle

Then the sample image is being rotated for all angles from the interval of 0-89 degrees. Simultaneously for samples in these positions the mean intensity functions are being calculated. Their examples for the angle of 35 degrees are given in Fig.11.



b

Fig.11 Mean intensity function in rows of the sample (a) and the DCH image (b) for the angle of 35 degrees

To estimate the position in every angle of rotation we use a variance or a mean standard deviation of calculated functions:

$$E^2(\mathrm{h}) = (1/N)\cdot\sum_{i=1}^{N}((y_i - \bar{y}(h))^2, \qquad (4)$$

where N *is* a number of points of the plot, $\bar{y}(h)$ is an average value of the mean intensity function of the DCH

image, y_i is a point value of the function, $E^2(\mathrm{h})$ is a variance value of the corresponding plot.

The full analysis gives us 90 pairs of plots. For every position we determine what image has the maximal deviation:

$$d(a_{\max}) = \max_a d(a), a = 0,1,2,...,89 \qquad (5)$$

where $d(a)$ is a deviation of mean intensity function depending of the angle a.

Examples for some angles and corresponding functions are given in Table 1. We see a standard deviation of mean intensity functions for samples are smaller than those for the DCH images. They are much smaller in relation to the mean values of corresponding plots. But in one column deviations for zero angle and other angles of samples could differ between themselves in times.

TABLE I. STANDARD DEVIATION OF SAMPLE AND DCH IMAGE BY THE ROTATION ANGLE

	Mean intensity function			
Angle	**Sample image**		**DCH image**	
	deviation	**mean**	**deviation**	**mean**
0	1,67	124,06	10,49	53,65
55	7,77	124,33	14,10	38,06
-35	6,63	123,02	10,70	34,90
			Segmented image	
			12,26	2,679

So, after solving (5) we have the angle indicating the sample position in which scratches have the best visibility. To determine coordinates of scratches we use the DCH image (Fig.12). Cropping the dark part of that image (h=180) we amplify the white colors and obtain the mean intensity function given in Fig.13a by blue color.

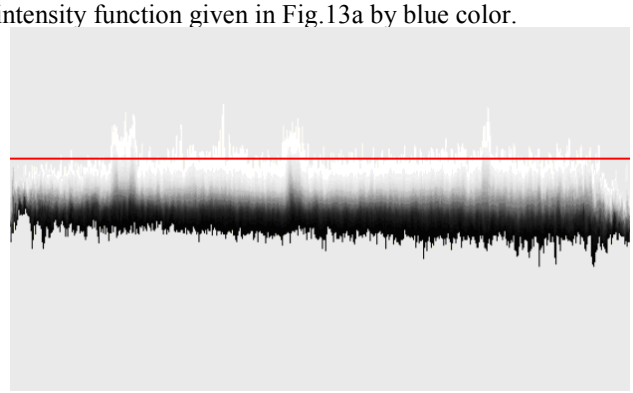


Fig.12 Mean intensity function in rows the DCH image for the angle of 35 degrees

For comparison in Fig.13 the native plot is given by red color. Clustering the cropped part of the DCH image gives us bands in Fig. 14 indicating coordinates of scratches on the rotated sample,

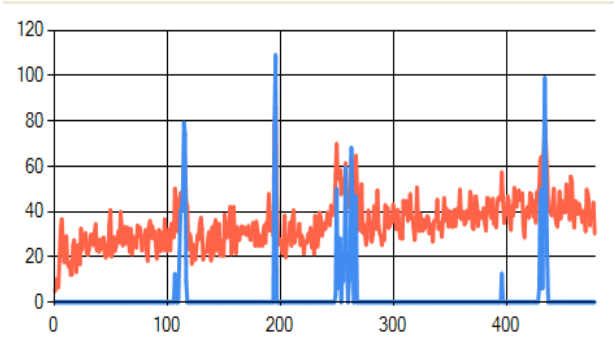


Fig.13 Mean intensity function in rows of the sample (a) and the DCH image (b) for zero angle

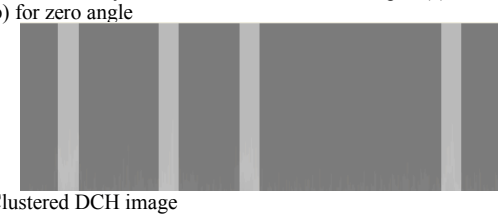


Fig.14 Clustered DCH image

By found coordinates scratches on the investigated surface are marked and given in Fig. 15.

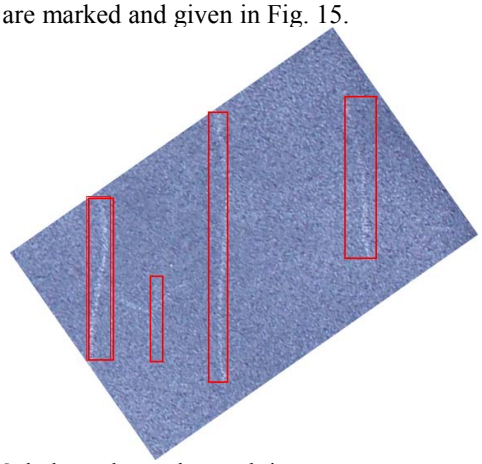


Fig. 15. Marked scratches on the sample images

One more experiment was held with a scratch on rotating image in Fig.16. For it the clustered DCH image is given in Fig.17.

Fig.16 Rotated image of a sample with scratches

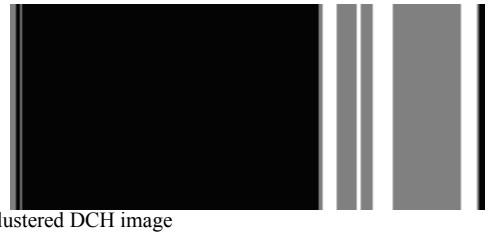


Fig.17 Clustered DCH image

Mean intensity function for the DCH image and the clustered DCH image are given in Fig. 18.

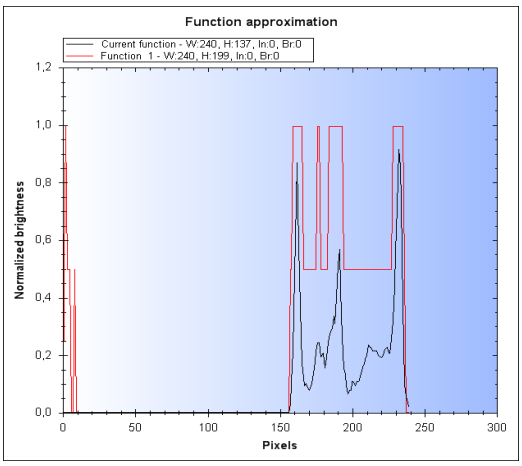


Fig.18 Mean intensity function in rows of the sample (blue) and the clustered DCH image (red) for 20 degree

In common two experiments with rotating must be hold to determine the areas containing scratches or defects of other types.

IV Conclusion

The algorithm is based on distributed cumulative histogram images the K-mean clustering algorithm was proposed and related software developed. It unites area indicating defects into band of equal pixel intensity allowing to determine coordinates of defected areas on the surface. To detect weakly expressed scratches on the surface rotating image and its distributed cumulative histogram to find a maximal variance of meant intensity function were used. It is shown that this algorithm is useful for detecting surface defects.

References

[1] K. Hanbay, M. Talu and Ö.Muhammed, “Fabric defect detection systems and methods—A systematic literature review”, Optik - International Journal for Light and Electron Optics, 127, 2016. 10.1016/j.ijleo.2016.09.110.

[2] R. Maran, D. Palumbo; U. Galietti, E. Stella, Tiziana D'Orazio. Automatic detection of subsurface defects in composite materials using thermography and unsupervised machine learning 2016 IEEE 8th International Conference on Intelligent Systems .516-522. R. Marani, D. Palumbo, U. Galietti, E. Stella and T. D'Orazio, “Automatic detection of subsurface defects in composite materials using thermography and unsupervised machine learning”, 8th International Conference on Intelligent Systems, pp. 516-521, 2016. doi: 10.1109/IS.2016.7737471.

[3] Ke-Chen Song, Shao-Peng Hu, Yun-Hui Yan, Jun Li Surface Defect Detection Method Using Saliency Linear Scanning Morphology for Silicon Steel Strip under Oil Pollution Interference, 2014 ISIJ International , 54(11), 2598-2607. K.-E. Song, H. Shaopeng, Y.-H. Yan and J. Li, “Surface Defect Detection Method Using Saliency Linear Scanning Morphology for Silicon Steel Strip under Oil Pollution Interference”, ISIJ International, 54, 2014, pp. 2598-2607. 10.2355/isijinternational.54.2598.

[4] A. Borghese and M. Fomasi, “Automatic Defect Classification on a Production Line”, Intelligent Industrial Systems, 1, 2015. 10.1007/s40903-015-0018-5.

[5] D.-M. Tsai and T.-Y. Huang, “Automated surface inspection for statistical textures”, Image and Vision Computing, 21, 2003, pp. 307-323. 10.1016/S0262-8856(03)00007-6.

[6] L. Qiu, W. Xiaojun and Y. Zhiyang, “A High-Efficiency Fully Convolutional Networks for Pixel-Wise Surface Defect Detection”, IEEE Access, 2019, p. 1. 10.1109/ACCESS.2019.2894420.

[7] R. Melnyk, Y. Kalychak and R. Kvit, ¨Analysis of cloudiness by segmentation and Monitoring of satellite map images”, International Journal of Computing, vol. 18, iss. 2, 2019, pp. 169–180.

[8] Detecting scratch on image with much noise, [https://stackoverflow.com/questions/33227202/detecting-scratch-on-image-with-much-noise].

IEEE Third International Conference on Data Stream Mining & Processing
August 21-25, 2020, Lviv, Ukraine

Augmented Reality in Web: Results and Challenges

Ruslan Timchenko
It-Jim
Kharkiv, Ukraine
timchenkoruslan97@gmail.com

Oleksiy Grechnyev
It-Jim
Kharkiv, Ukraine
shrike4625@gmail.com

Sergiy Skuratovskyi
It-Jim
Kharkiv, Ukraine
ssnake@it-jim.com

Yurii Chyrka
It-Jim
Kharkiv, Ukraine
yurii.chyrka@it-jim.com

Ievgen Gorovyi
It-Jim
Kharkiv, Ukraine
ceo@it-jim.com

***Abstract*—The paper presents basic concepts of augmented reality applications and challenges in building them in the web. We describe the technical and algorithmic stack required to develop, implement and deploy the augmented reality application. Theoretical concepts behind marker detection and tracking are discussed. Two different pipelines are implemented: server-based with algorithms execution in the cloud and completely front-end solution that runs on a user device. We show advantages and disadvantages of each approach and analyze experimental results as well.**

Keywords—augmented reality, visual tracking, image marker, webAR

I. Introduction

Augmented Reality (AR) is the technology that connects our real life with a digital world. Extending reality is possible by overlying layers with virtual objects over the screen of user devices [1]. Different kind of information can be augmented: text, video, images, audio and both static and dynamic 3d models [2]

Since most of the information comes through the human visual system, AR is becoming more and more widespread. Its strength is already demonstrated in a variety of areas: advertising [3], game industry [1, 4], education [5], medicine [6], entertainment [4, 7-9].

Two important points are taken into consideration during the development of the AR system: technical solutions and hardware platforms used. Several typical computer vision problems are usually solved during the construction of the AR framework. In particular, object detection and recognition (planar markers or arbitrary objects) [2], content-based image retrieval [10] for visual search, simultaneous localization and mapping (SLAM) for 3D object detection [11], markerless tracking and multi-player AR applications [12]. The second point is where to utilize AR technology, i.e. what device to use: mobile phone or tablet, AR glasses, desktop.

There are plenty of different mobile AR systems like ARKit, ARCore, Vuforia, EasyAR, etc. Most of them require installation of the mobile application for AR experience. In the paper, we describe an alternative way to use AR technology: augmented reality in a browser, or web AR [13, 14]. Indeed, the obvious advantage of AR in a mobile browser is instant immersion without need to install any mobile applications.

The paper is structured as follows. The problem of planar marker recognition for AR is discussed in section II. In particular, we consider a typical keypoint-based detection algorithm and analyze various local feature descriptors. In addition, we introduce a hybrid tracking approach which combines sparse optical flow [15] and template-based tracker [2]. The created algorithmic pipeline and its deployment into web AR application are described in section III. We consider two separate architectures: server-based one with algorithm execution in the cloud and pure front-end solution that runs on a user device. We also analyze their strong and weak sides. Finally, experimental results are shown in section IV.

II. Computer Vision Solutions for Marker-Based Augmented Reality

In order to render AR model correctly over the frames from the camera we need to estimate its position. In the case of marker-based AR, the planar marker position in the frame should be known. Hence, we start with the marker detection. Once the marker is found, we track its position in consequent video frames. The marker position in the frame is used to calculate the homography transformation matrix and estimate 6 degrees of freedom (6 DoF) camera position from it.

A. Marker Detection

The estimation of extrinsic camera parameters starts with the detection of the marker position in the scene. By marker we mean not some binary pattern, but a certain image. The fast and robust solution for image marker detection is based on local features. There are plenty techniques for keypoints detection and description. Here we analyzed two of them: ORB [16] and AKAZE [17].

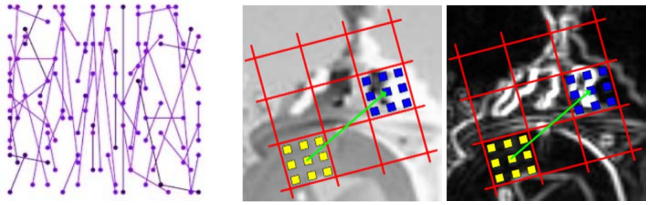
The use of ORB features provides the fastest marker detection procedure comparing to analogues. For the keypoint detection, the oriented FAST [18] is applied in combination with the image pyramid. The local patch around the keypoint is characterized by binary ORB descriptor (Fig. 1, a). It is based on the steered version of BRIEF [19] with binary tests analyzed for correlation to provide more distinctive features.

AKAZE for now is the optimal combination between speed and accuracy. It incorporates fast explicit diffusion, which provides more efficient scale space forming than in SIFT and KAZE. The modified version of LDB descriptor

978-1-7281-3214-3/20/$31.00 ©2020 IEEE

[20] (Fig. 1, b) provides a rotation invariance and preserves computational efficiency without the use of integral images. Scale-dependent sub-sampling (small squares in Fig. 1, b) is used instead of calculating the average value of each area (large squares).

To compare the efficiency of matching ORB and AKAZE features, we used our company business card as a marker (Fig. 2). The keypoint locations were the same in both cases, thus only descriptors and their matching affected the result. The 50 best matches are represented in Fig. 2.



a b

Fig. 1. ORB (a) and M-LDB (b) descriptors.

Fig. 2. Keypoint matching with ORB (upper) and AKAZE (lower).

It is clear that AKAZE provides more reliable matches. Still, most of ORB matches are also good, so random sample consensus (RANSAC) or a similar technique provides a good estimation of the homography matrix in this case, too. And as computation time is crucial for AR in general and WebAR in particular, ORB is our tool of choice.

B. Tracking

Marker detection is a time-consuming operation. Moreover, typical keypoint descriptors handle quite limited view angles [2], and detection fails, which results in a bad AR experience of a user.

To solve both problems, marker tracking algorithms are typically used [2, 15,21]. A common way to track inter frame changes is to use sparse optical flow (OF) methods [2, 23]. This is accomplished by matching adjacent frames instead of straightforward comparison of the reference marker image with the camera frame image, which is not efficient.

A common problem of OF methods in AR applications is the drift of estimated pixel locations. This leads to incorrect camera pose estimation and, as a result, incorrect AR model augmentation. To improve the robustness of our planar tracker, we utilized a template-based tracking called the sum of conditional variances (SCV) [22]. It is an iterative approach, which is used to refine the residual error in homography estimation after OF application.

The proposed fusion of two different trackers leads to a robust tracking.

Firstly, OF is applied to estimate initial homography between adjacent frames. Secondly, SCV refinement is used to receive more accurate warping between the frames (Fig. 3). Such combination allows to compensate OF drift and handles extremal view angles.

Fig. 3. Two-step homography estimation.

C. Camera Pose Estimation

Camera pose estimation is the primary element of AR system that affects the correct rendering of models. Thus, it should be precisely retrieved from the camera frames [2].

Let's consider the common projective geometry for the pinhole camera model [13]. The camera projects points from the 3D world *(x,y,z)* into a 2D pixel in the image plane *(u, v, w)*. Here *w* is a scale parameter for homogeneous coordinates. This transformation can be written as follows:

$$\begin{bmatrix} uw \\ vw \\ w \end{bmatrix} = \begin{bmatrix} f_x & 0 & c_x \\ 0 & f_y & c_y \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} r_{11} & r_{12} & r_{13} & t_x \\ r_{21} & r_{22} & r_{23} & t_y \\ r_{31} & r_{32} & r_{33} & t_z \end{bmatrix} \begin{bmatrix} x \\ y \\ z \\ 1 \end{bmatrix} \quad (1)$$

where c_x, c_y denote the origin of image coordinates, in our application they are equal to zero; f_x, f_y are focal lengths of the camera, responsible for the scaling (zooming). The first matrix is the intrinsic one, which is independent on the scene. It is specific for the particular camera device and can be found once during the camera calibration [14]. The next matrix is the extrinsic one. It contains the camera pose, describing transformation from the world coordinates to the camera coordinate system. The camera pose consists of a translational displacement of the camera (t-vector) and its orientation (r-elements), which represents the transformation (translation and rotation) between the world and the camera coordinate systems.

Let's consider how the marker location in the frame is used to retrieve the camera pose. It is known that the transformation between the planar object locations in two images can be described with the homography matrix [2]:

$$M'(u', v') = H \times M(u, v) = \begin{bmatrix} h_{11} & h_{12} & h_{13} \\ h_{21} & h_{22} & h_{23} \\ h_{31} & h_{32} & 1 \end{bmatrix} M(u, v) \quad (2)$$

Here, M and M' are the original marker and the warped marker on the frame, their pixel coordinates are represented as *(u,v)* and *(u`,v`)* respectively.

The matrix elements are found from matching keypoints in two pictures. Some matches may be wrong, so we used

RANSAC to drop out the outliers and get the right homography.

Homography is a more generic operation compared to the 6 DoF of camera pose (3x1 translation vector and 3x1 rotation vector). So, it is possible to convert homography into a camera pose. The corners of marker are linked with the homography projected ones onto the frame. This set of points is used to estimate rotations and translation vectors basing on Levenberg-Marquardt optimization [24, 25]. The method iteratively minimizes the pixel reprojection error points, represented as the sum of squared distances between the corresponding images.

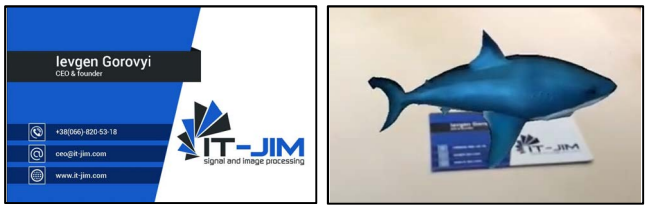


Fig. 4. Example of 3D model augmentation: image marker (left), 3D model over the detected marker (right)

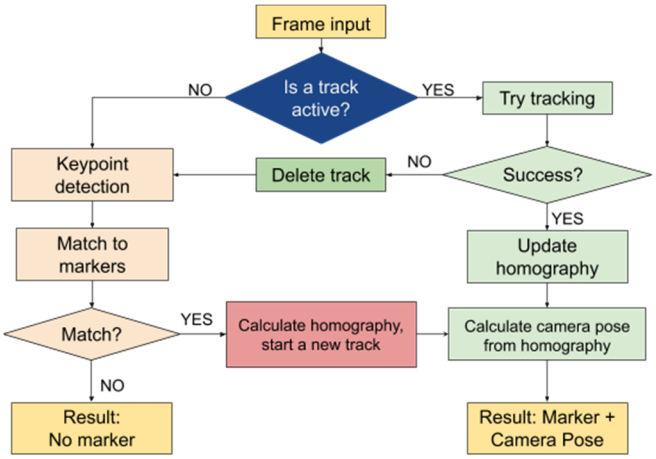


Fig. 5. The main algorithm pipeline.

As a result, 6 DoF camera pose is constantly updated based on the applied planar marker detection and tracking procedures. The estimated camera pose is used to render the AR content. An example of image and 3D model rendering is shown in the Fig. 4.

III. Created Pipeline and Its Implementations

In this section, we consider practical aspects of integration of developed computer vision algorithms and discuss important details of web AR system architecture.

A. High-level Scheme of Camera Frames Processing

The described technical solutions were combined into the following pipeline (Fig. 5).

For a given camera frame, either marker detection or tracking algorithms are used. The homography matrix between marker and camera frame is constantly updated. Finally, the camera pose is estimated in real-time providing the information for AR models rendering in the browser window.

We have created two conceptually different deployments: a front-end-back-end with algorithms running in the cloud and a purely front-end pipeline. The next sub-section describes those two ways of deployment in details.

B. Frontend-Backend Approach

At the front end, we have an outgoing and incoming data streams. Therefore, we divided front end operations into two independent asynchronous threads: camera rendering and model rendering (Fig. 6). Unlike the single sequential thread, when we show video frames after the server response, in this case 3D model is overlaid over the live video stream without any delays and freezes. This approach looks better for a user and provide more opportunities.

The first thread takes a frame from the device camera and immediately displays it on the web page. To use a video stream from the camera, the browser must ensure that the requested web page is safe, thus only HTTPS pages with a SSL/TSL certificate are allowed to access the camera. The captured frame is downsized and converted to JPEG format. At the end, the processed frame is sent to the back end via a secure websocket (wss).

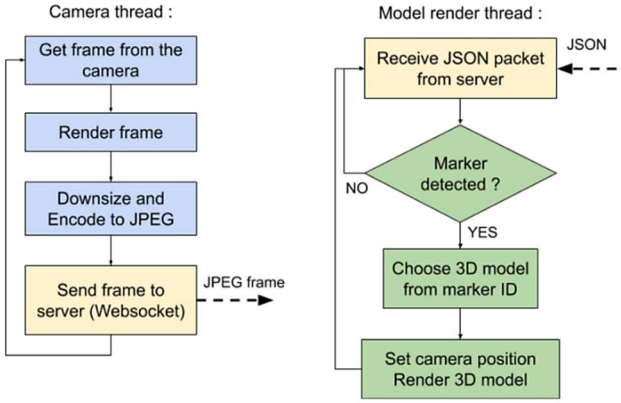


Fig. 6. Frontend architecture. Camera and rendererasynchronical threads.

The second thread, named ‘model render thread’, receives JSON packets from the server. They contain an ID of the identified marker (if there is no marker in the frame, it is equal to -1) and the camera parameters. The latter depend on the render library. In our case, ThreeJS library [26] was used, which creates a camera object by three vectors: ‘position’, ‘lookAt’ and ‘up’.

The calculation of Three JS camera pose has a few important moments. The first is the freedom to choose a render formalism. That means the choice between ‘moving the model’ and ‘moving the camera’. We prefer to move camera rather than the model, because this way is more similar to the real camera moving. Thus, the camera position t_{cam} is calculated as $t_{cam} = -R^{-1}t$, where t is the camera translation for the stationary model described above in the camera pose (1).

The second moment is the usage of left-handed system in contrast with the right-handed one in OpenCV library that was used in C++ pipeline implementation. So, we have to change the sign of z-component of the camera position.

The ‘lookAt’ vector indicates the point the camera is facing at, the ‘up’ vector determines the rotation of the camera view. Both vectors can be found from inverse perspective projection transformation $(u,v) \rightarrow (x,y,z)$. There is an infinite number of solutions, so for simplicity we assume that marker is located in xy0 plane. Considering F: $(u,v) \rightarrow (x,y,0)$, we then project $p_0 = F(0, 0), p_1 = F(0, 1)$. p_0 is the viewport center and

represents the 'lookAt' vector, while $up = p_1 - p_0$ determines the 'up' vector.

The back-end is implemented using Java and Spring Boot. Its architecture is shown in the Fig. 7. It communicates with the front-end via secure web sockets only. For each web session, a separate engine object is created (a wrapper to C++ algorithms), thus the server can process multiple sessions correctly. The engine performs all steps of frame processing, including detection, tracking and calculation of camera pose. At the end, it sends information to the front-end.

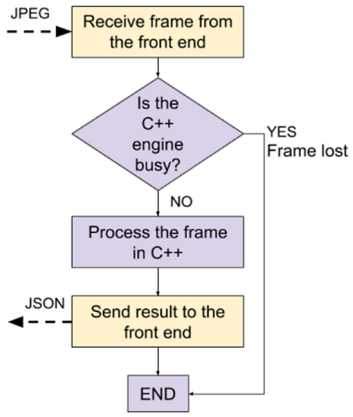


Fig. 7. Server structure.

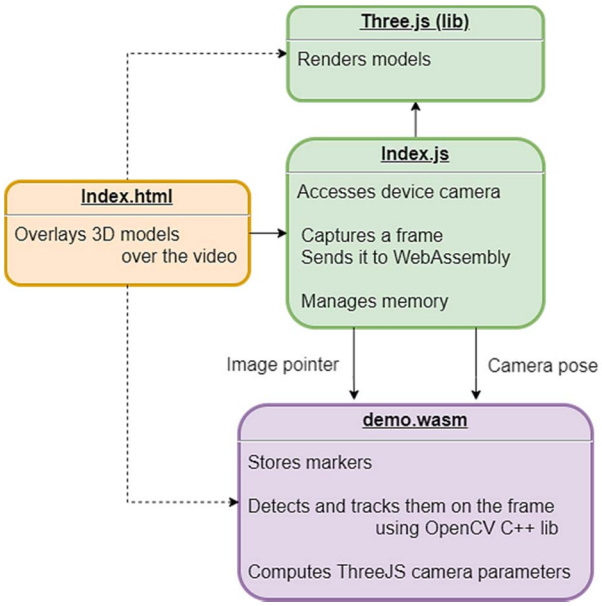


Fig. 8. The architecture of the pure front-end solution.

C. Frontend-only Architecture

To avoid the main problem of the previous architecture, the lag introduced by the network latency, we made a front-end only solution (Fig. 8).

In this way, there is no lag (i.e. transferring an encoded frame to the server and an array of numbers back). On the other hand, all files (including bulky 3D models) have to be downloaded before the web application starts. The required time depends mostly on the user's connection speed and can be quite long.

This application consists of separate JS modules and WebAssembly files [27]. Here the backend server is absolutely replaced by the wasm file, which was compiled from C++ project with the main algorithm pipeline [28].

WebAssembly is an open standard that defines a portable binary-code to run natively in browsers. In order to compile C++ project, we used Emscripten SDK [29]. It is a suitable instrument to call C++ functions from JavaScript side, and often the speed of procedures is higher than of pure JavaScript.

Fortunately, our architecture has not changed much and it keeps the primary logic of the previous architecture unchanged. The only difference is that a user now downloads all files and does all the calculations in the browser.

D. Architecture Summary

A brief summary with the strong and weak sides of each solution are presented in the Table 1. On its basis, one can select any of the solutions for the particular case that is most suitable for the specified conditions.

The frontend-backend solution is the best when we need complicated processing (potentially including some neural networks) or when we want to cover as much devices as possible, regardless of their hardware computational performance. However, it requires a stable connection during an AR session for unceasing data stream. Moreover, it can be very costly in two scenarios: either we want to run our application worldwide (then we need to deploy it over multiple servers over the globe to minimize the network latency between users and the nearest servers)or we want to serve a lot of users simultaneously(then we need to run a powerful backend server).

TABLE I. PROS AND CONS OF ARCHITECTURE TYPES

	Pros	Cons
Frontend-backend	- Provides better performance - Allows to run heavy algorithms - Covers weak devices	- Network latency - Requires a reliable connection - Costly in multiuser and worldwide usage scenarios
Frontend-only	- No network lags - Runs independently from the network - Easier implementation	- Requires a powerful hardware - Downloads heavy models

The frontend-only solution is easier to implement and has a rather weak dependency on the network connection. Once the user gets all the necessary data from the server, the network can be turned off and the web page with AR application will still operate offline in the browser. To enjoy AR experience, the user has to run it on a powerful modern device with a few gigabytes of RAM (it depends on the weight of the models) and mid to high level system-on-chip (SoC).

IV. EXPERIMENTAL RESULTS

This section contains some relevant information about benchmarking of utilized computer vision solutions as well as performance tests of the AR system.

A. Benchmarks for Marker Detectors

To compare different detectors and find their optimal parameters, we have conducted a number of experiments changing detector parameters. We tuned the input image size, number of pyramid levels (octaves) and looked at the dependence on the number of features. As our target is mobile devices, we need to be sure that their computational power is sufficient to use our application in real time and balance between the detection quality and the algorithm speed.

Considering the results, presented in Fig.9 – Fig.11, we have chosen ORB detector with 1000 keypoints and 8 scale levels for 730x410 resolution.

We tested the performance on preliminary recorded videos with markers in various positions. The quality metrics was defined as the ratio of frames with detected markers to the total number of frames in the video sample. The successful detection is registered if the average distance of 20 best matches is below a certain threshold (32.0 in our model).

B. *Performance Tests for Frontend-Backend Architecture*

We have tested the influence of the backend server's computational power on the execution time. The results for two types of AWS instances are shown in Table 2. FPS is the time it takes a frame to complete the entire pipeline with rendering. When a marker is present in the scene, FPS is determined mostly by tracking. If there is no marker, tracking is impossible, and only detection affects speed.

Although FPS is quite high, there is a significant lag between the camera and the render threads. The main problem here is not C++ algorithms, but quality and stability of the internet connection between the browser and the back-end server. The lag mainly stems from the network latency, which varies for different conditions. A delay of a few frames (0.1-0.2s) was very common in our case. In addition, some frames may be lost on the server.

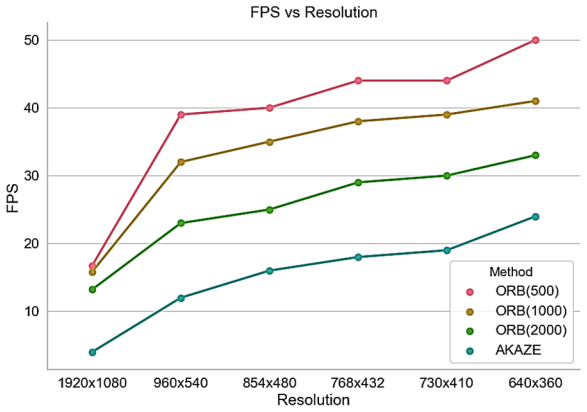


Fig. 9. FPS dependence on the image resolution and a detector type.

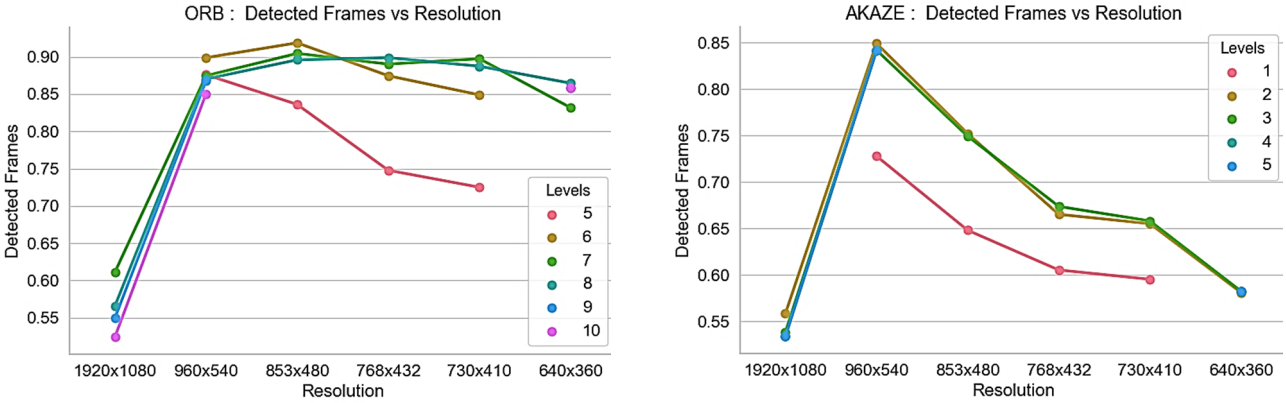


Fig. 10. Accuracy provided by ORB and AKAZE dependion on input image resolution and a number of levels.

TABLE II. TIMING ON DIFFERENT AWS INSTANCES

Functions	T2 small	T2 large
Detection (marker)	120 ms	77 ms
Detection (no marker)	100 ms	46 ms
Tracking	5 ms	4 ms
Decode JPEG	1.2 ms	1.0 ms
Camera pose estimation	0.3 ms	0.2 ms
FPS (marker)	~20	~20
FPS (no marker)	6-7	>10

Unfortunately, network latency can be reduced only by decreasing the distance between the geographical positions of user and server.

C. *Performance Tests for Serverless Architecture*

In the serverless solution, the performance is determined by the device (e.g. cell phone) capabilities and does not depend on any server where the web page is located.

For testing, we used a laptop with Intel Core i5-8250U CPU at 1.60GHz. The results are demonstrated in Table 3. It is about as powerful as modern high-end mobile devices. The performance of the majority of devices is expected to be worse.

TABLE III. TIMING ON THE USER'S LAPTOP

Functions	Time
Detection (1 marker)	150-250 ms
Tracking	15-25 ms
Rendering	10-30 ms
Memory management	1-2 ms
FPS (marker)	~16
FPS (no marker)	3-5

As the resulting timing depends on the current load of the user's CPU, we give the time ranges. WebAssembly reduces the speed of detection and tracking compared to the original C++ code.

V. CONCLUSIONS

We described the important elements required to build a web-based AR application. The applied computer vision solutions were analyzed and discussed. Two different architectures for

the AR system were proposed. The developed AR pipeline is flexible and extendable. We are planning to integrate a new group of computer vision and machine learning algorithms for face tracking, text recognition, arbitrary 2D/3D object recognition in the near future. In addition, additional work will be done for the algorithm optimization, which is crucial for the application running in a mobile browser.

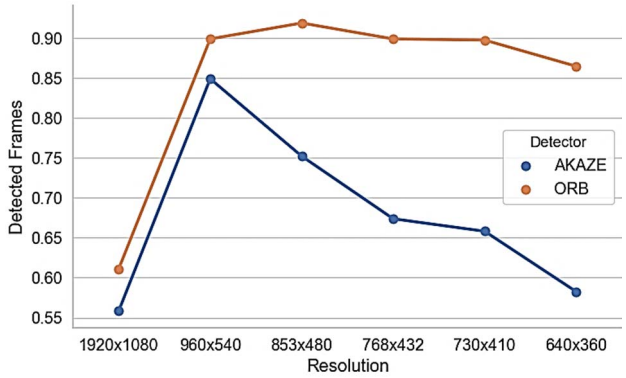


Fig. 11. Accuracy of detectors depending on the Image Resolution. For each resolution was chosen the best properties provided the large number of detected frames.

REFERENCES

[1] V. Vovk et al., "Light-Weight Tracker for Sports Applications" Proceedings of Signal Processing Symposium, 17-19 September, Krakow, Poland, pp.255-259, 2019.

[2] I. Gorovyi et al., "Advanced Image Tracking Approach for Augmented Reality Applications" Proceedings of Signal Processing Symposium, 12-14 September, Jachranka, Poland, pp.266-270, 2017.

[3] Y.-G. Kim and W.-J. Kim, "Implementation of augmented reality system for smartphone advertisements," Int. J. Multimedia Ubiquitous Eng., vol. 9, no. 2, pp. 385–392, 2014.

[4] (Jan. 2017). Augmented/Virtual Reality Report 2017. Accessed: Feb. 17, 2018. [Online]. Available: https://www.digi-capital.com/reports

[5] B. Perry, "Gamifying French language learning: A case study examining a quest-based, augmented reality mobile learning-tool," Procedia-Social Behav. Sci., vol. 174, pp. 2308–2315, 2015

[6] Ha, Ho-Gun & Hong, Jaesung. (2016). Augmented Reality in Medicine. Hanyang Medical Reviews. 36. 242. 10.7599/hmr.2016.36.4.242.

[7] Pokémon GO Report. Accessed: Mar. 5, 2018. [Online]. Available: https://pokemongolive.com/en/post/headsup

[8] Z. Shi, H. Wang, W. Wei, X. Zheng, M. Zhao, and J. Zhao, "A novel individual location recommendation system based on mobile augmented reality," in Proc. IEEE Int. Conf. Identificat., Inf., Knowl. Internet Things (IIKI), Oct. 2015, pp. 215–218.

[9] N. Gavish et al., "Evaluating virtual reality and augmented reality training for industrial maintenance and assembly tasks," Interact. Learn. Environ., vol. 23, no. 6, pp. 778–798, 2015.

[10] M. A. Tahoun, et al., "A robust content-based image retrieval system using multiple features representations," Proceedings. 2005 IEEE Networking, Sensing and Control, 2005., Tucson, AZ, 2005, pp. 116-122.

[11] H. Durrant, T. Bailey. "Simultaneous localization and mapping: Part I", IEEE Robotics & Automation, (2006), pp. 99-108.

[12] Nowacki, Paweł & Woda, Marek. (2020). "Capabilities of ARCore and ARKit Platforms for AR/VR Applications". 10.1007/978-3-030-19501-4_36.

[13] Qiao, Xiuquan & Pei, Ren & Dustdar, Schahram & Liu, Ling & Ma, Huadong & Junliang, Chen. (2019). Web AR: A Promising Future for Mobile Augmented Reality - State of the Art, Challenges, and Insights. Proceedings of the IEEE. 107. 1-16. 10.1109/JPROC.2019.2895105.

[14] A. Charland and B. Leroux, "Mobile application development: Web vs. native," Commun. ACM, vol. 54, no. 5, pp. 49–53, May 2011

[15] Senst, Tobias & Eiselein, Volker & Sikora, Thomas. (2012). Robust Local Optical Flow for Feature Tracking. IEEE Transactions on Circuits and Systems for Video Technology. 22. 10.1109/TCSVT.2012.2202070.

[16] 9. E. Rublee, V. Rabaud, K. Konolige, G. Bradski, Gary, ORB: an efficient alternative to SIFT or SURF. Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2011, pp. 2564-2571. doi 10.1109/ICCV.2011.6126544

[17] P. F. Alcantarilla, J. Nuevo, A. Bartoli, Fast Explicit Diffusion for Accelerated Features in Nonlinear Scale Spaces. 2013, British Machine Vision Conference (BMVC). doi 10.5244/C.27.13.

[18] E. Rosten, T. Drummond. Machine learning for high-speed corner detection. European Conference on Computer Vision (ECCV), 2006, pp. 430-443.

[19] M. Calonder, V. Lepetit, C. Strecha, P. Fua, Brief: Binary robust independent elementary features. European Conference on Computer Vision (ECCV), 2010, pp. 778-792.

[20] X. Yang, K. T. Cheng, LDB: An ultra-fast feature for scalable augmented reality. IEEE and ACM Intl. Sym. on Mixed and Augmented Reality (ISMAR), 2012. doi 10.1109/ISMAR.2012.6402537

[21] Bertrand Delabarre, Eric Marchand. Visual Servoing using the Sum of Conditional Variance. IEEE/RSJ Int. Conf. on Intelligent Robots and Systems, IROS'12, 2012, Vilamoura, Portugal. pp.1689-1694. ffhal-00750602f

[22] Richa, Rogerio & Sznitman, Raphael & Taylor, Russell & Hager, Gregory. (2011). Visual tracking using the sum of conditional variance. IEEE International Conference on Intelligent Robots and Systems. 2953-2958. 10.1109/IROS.2011.6094650.

[23] OpenCV. Open Source Computer Vision Library. 2015.

[24] Levenberg, K.: A method for the solution of certain non-linear problems in least-squares. Quarterly of Applied Mathematics 2 (1944) 164–168

[25] Marquardt, D.: An algorithm for the least-squares estimation for non-linear parameters. SIAM J. Applied Mathematics 11 (1963) 431–441

[26] Dirksen, Jos (2013). Learning Three.js: The JavaScript 3D Library for WebGL. UK: Packt Publishing. ISBN 9781782166283.

[27] A. Haas et al., "Bringing the web up to speed with WebAssembly," in Proc. 38th ACM SIGPLAN Conf. Program. Language Design Implement., 2017, pp. 185–200.

[28] A. Møller, "Technical perspective: WebAssembly: A quiet revolution of the Web," Commun. ACM, vol. 61, no. 12, p. 106, 2018.

[29] Alon Zakai. "Emscripten: an LLVM-to-JavaScript compiler. In Proceedings of the ACM international conference companion on Object oriented programming systems languages and applications companion". Association for Computing Machinery, New York, NY, USA, 301–312. 2011

IEEE Third International Conference on Data Stream Mining & Processing
August 21-25, 2020, Lviv, Ukraine

Comparative Analysis of Classic Computer Vision Methods and Deep Convolutional Neural Networks for Floor Segmentation

Anastasiia Skoryk
It-Jim
Kharkiv, Ukraine
seredkina.a@gmail.com

Yurii Chyrka
It-Jim
Kharkiv, Ukraine
yurii.chyrka@it-jim.com

Ievgen Gorovyi
It-Jim
Kharkiv, Ukraine
ceo@it-jim.com

Oleksiy Grechnyev
It-Jim
Kharkiv, Ukraine
shrike@it-jim.com

Pavlo Vyplavin
It-Jim
Kharkiv, Ukraine
pavlo.vyplavin@it-jim.com

***Abstract*—In the paper, we analyze the problem of automatic room floor segmentation. For this purpose, we consider several classic computer vision algorithms as well as some of the deep convolutional neural network architectures. The segmentation results are illustrated and compared. An idea for combining two groups of methods is proposed. It is demonstrated that a proper fusion provides the best segmentation quality.**

Keywords—indoor image segmentation, superpixels, graph clustering, deep learning, CNN.

I. INTRODUCTION

Image segmentation is one of the key topics in computer vision. Usually, it is interpreted as semantic segmentation, i.e. linking of each pixel in an image with a label from a particular set of classes, for example, “human”, “grass”, “road”, “floor”, “table”, etc. Segmentation appears in a wide range of applications such as scientific image analysis, robotic vision, scene understanding, augmented reality and many more [1, 2].

In the case of segmentation of surfaces with a similar texture or patterns, we can label subsets of pixels that share similar characteristics: intensities, colors, and locations. However, correct separation of different classes may be a challenge due to varying illumination, noise, occlusions, shades, light spots, reflections, and camera perspective changes.

There are many existing methods for image segmentation: from classic ones like simple thresholding [3] or superpixels [4, 5] to quite advanced deep learning-based solutions [6]. In addition, various machine learning methods are often used along with hand-crafted features [7-9].

In this paper, we analyze the problem of automatic room floor segmentation. Such a solution can be used for different purposes like mixed reality (MR) applications, interior design, and entertainment. Our goal is to analyze both classic computer vision methods as well as common deep learning (DL) based convolutional neural network (CNN) architectures. As well we propose a methodology for combination of classic and deep learning based methods in order to get the best overall result.

In Section II, we briefly describe the methods used in our experiments and show some of the intermediate image processing steps. A proposed fusion scheme of classic and DL-based branches outputs is shown in Section III. Finally, datasets and experiment results are described in Section IV.

II. METHODS OVERVIEW

A. Classical Pipeline

Among many different methods for indoor images segmentation [10-12], superpixels are the most widely used technique [4, 5].

According to the definition, a superpixel is a group of a few pixels with common properties **Error! Reference source not found.**. Representing an image as a group of superpixels allows one to get a compact representation and to retrieve the image regions sharing the same properties.

There are different variations of superpixel algorithms [4, 13]. One of the most widely used approach, the simple linear iterative clustering (SLIC), adapts a k-means clustering. The method works by clustering pixels based on their color similarity and proximity in the image plane [4]. The distance between two pixels in the combined five-dimensional LabXY [4] space is defined as follows:

$$\begin{aligned} d_{lab} &= \sqrt{(l_k - l_i)^2 + (a_k - a_i)^2 + (b_k - b_i)^2} \\ d_{xy} &= \sqrt{(x_k - x_i)^2 + (y_k - y_i)^2} \\ D_s &= d_{lab} + \frac{m}{S} d_{xy}, \end{aligned} \tag{1}$$

where S is a grid interval, m is the compactness parameter.

An example of superpixels clustering is given in Fig. 1b. Normally, two important parameters are tuned: the number of superpixels (was set to 300) and the compactness measure (was set to 7). The former corresponds to the maximum amount of superpixels to be extracted from the image, while the latter corresponds to the trade-off between proximity and

978-1-7281-3214-3/20/$31.00 ©2020 IEEE

color-similarity. Compactness controls the shape and smoothness of the superpixels' boundaries: with higher compactness they become smoother and the superpixels become more regular.

The problem is that the straightforward application of superpixels does not provide a perfectly segmented floor. In order to overcome this difficulty, we have created an additional pipeline for image processing. First of all, we group pixels of a color image (Fig. 1a) into superpixels (Fig. 1b). In parallel, we transform RGB image into HSV color space and work with the saturation channel only (Fig. 1d), because it highlights changes between the room surfaces the most. Then, we obtain an edge map of the S-channel image (Fig. 1e). From the combination of the superpixels image and the edge map

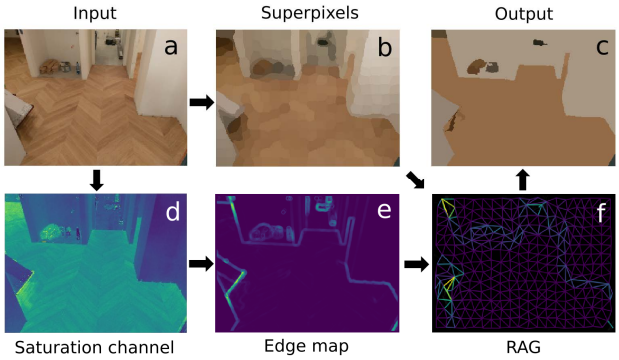


Fig. 1. The main steps of the classical pipeline. a) The input color image; b) the SLIC superpixels; c) the output clusterized image from the merged RAG; d) the saturation color channel; e) the image with highlighted edges; f) the RAG constructed from the superpixels and the edge map.

we construct a region of adjacency graph (RAG), which has greater edge weights on the borders (Fig. 1f). Finally, we cluster the superpixels into groups using graph hierarchical merging algorithm (Fig. 1c). Some details of this pipelines are given below.

RAG is an undirected weighted graph. Its vertices represent image areas (for example, superpixels), while its edges correspond to the connections between the adjacent regions [15]. RAGs give a spatial view of the images and are powerful tools for image processing if neighborhood relationships can be taken into account. In our case, images with emphasized edges (edge maps) are used to present this information. The Sobel gradient magnitude filter [16, 17] and the local binary pattern (LBP) feature map extraction [18] provides the most emphasized edges, so we used these two algorithms to construct the edge maps.

We obtain the output image regions (Fig. 1c) by performing agglomerative hierarchical clustering with mean linkage until a threshold [19, 20]. As the appropriate threshold value highly depends on the image, we estimated it from the distribution of the graph edge weights. The threshold is a value which corresponds to a specific percentile (for example 80% as shown in Fig. 2). In this case, RAGs are associated with their unique threshold while merging.

The graphs before and after the hierarchical merging are visualized in Fig. 3. The boundaries of image regions are also shown. These regions are the return segments, and one of them would be estimated as a floor. To decide which segment is the desired floor, we just take the biggest segment at the bottom of the image.

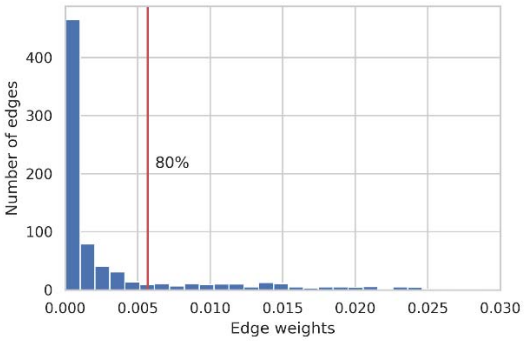


Fig. 2. The distribution of the RAG edge weights. Vertical line shows the percentile value to estimate threshold for hierarchical merging.

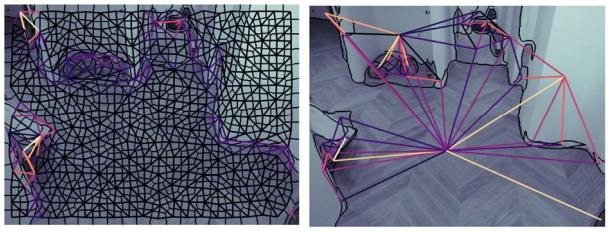


Fig. 3. The RAG before (left) and after (right) hierarchical merging. All nodes with the edge weight less than a threshold are merged together. Border of segments are shown in black.

Since the classical approach is very sensitive to parameter tuning, we have run the classical pipeline several times with different model parameters, resulting in many segmentation masks. The adjusted settings included RGB and HSV images as the SLIC inputs; Sobel filtering and LBP rotation invariant edge extraction as methods for the edge detection; 75%, 80% and 85% percentile values for the threshold estimation.

We add all binary masks together and if more than half of them are positive about a pixel, we label this pixel as true. Otherwise, we label it as false. As a result, we have a single binary segmentation mask.

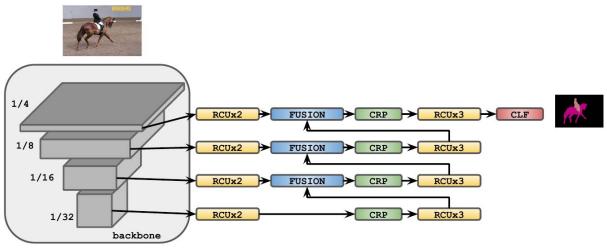
B. Deep Learning Pipeine

There are many different DL architectures available for floor segmentation [6, 21-23]. We used two CNNs: light-weight RefineNet [24] (see Fig. 4a) and FastFCN [25] with a joint pyramid upsampling (JPU) (see Fig. 4b). We used both CNN architectures with minimum changes, only the output layers were transformed to predict just 2 classes: floor and not a floor.

C. *Post-processing: Texture Feature Analysis and Edge Refinement*

Masks predicted either by classical algorithms or by CNN may have Masks predicted either by classical algorithms or by CNN may have complicated boundaries, while the floor shape is usually more or less straight. Moreover, when adding many masks, instead of binary segmentation with two labels, we get determined areas (either a floor or not a floor) where pixels have the same label on each mask, and some undetermined regions where masks have opposite labels. Texture features may be really helpful for analyzing of indoor surfaces and for final classification of undetermined segments to a floor or not a floor.

a)



b)

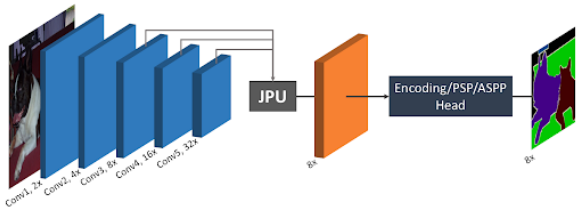


Fig. 4. The CNNs architectures used in the paper. a) The RefineNet architecture for semantic segmentation [23]. b) The FastFCN with a Joint Pyramid Upsampling (JPU) module and a multi-scale/global context module [24].

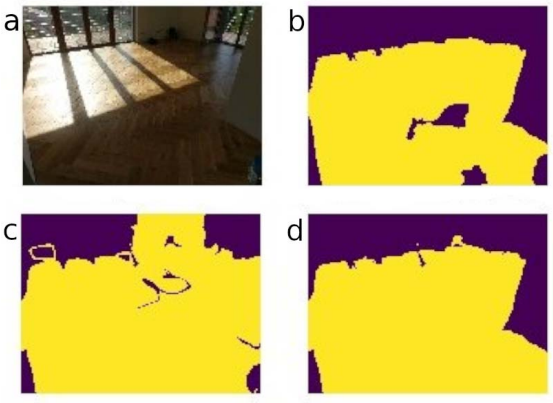


Fig. 5. Post-processing based on the texture feature analysis. a) the input image; b) the classical pipeline output; c) the mask from the deep learning pipeline; d) the mask after post-processing.

The whole image or its separate segments can be represented by features such as shape differences, texture differences, color or light fluctuations. The feature extraction algorithm provides fewer but more meaningful parameters to describe an image or its parts.

A wide range of algorithms for texture features extraction including statistical-based, transform-based, graph-based approaches and many other methods and their heirs are described in the literature [26]. In this study, we use a gray level co-occurrence matrix (GLCM) [27] which determines how often different pairs of pixels appear in an image. The GLCM expresses a matrix with a shape of image bitrates. From this matrix, one can extract features like 'contrast', 'dissimilarity', 'homogeneity', 'ASM', 'energy' and 'correlation' [27].

Extracting GLCM features from different segments of the image makes it possible to calculate the Euclidean distance in multidimensional feature space from the undefined segment to the determined floor (or not a floor) segment (2). The dimensionality n of feature vectors corresponding to the number of features are taken into account.

$$d(x,y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}. \quad (2)$$

The post-processing part combines many segmentation masks obtained by different methods. The main purpose of this stage is the final classification of uncertain areas or blobs. Feature analysis resolves these uncertainties and makes more accurate prediction (see Fig.5).

Finding blobs is done by analyzing contours of the masks. All uncertain blobs are linked to one of two (a floor or not a floor) determined areas by calculating the minimum distance in the feature space (2).

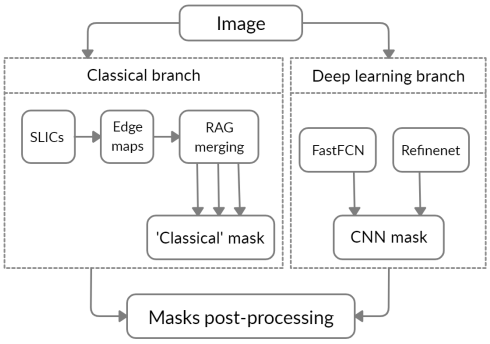


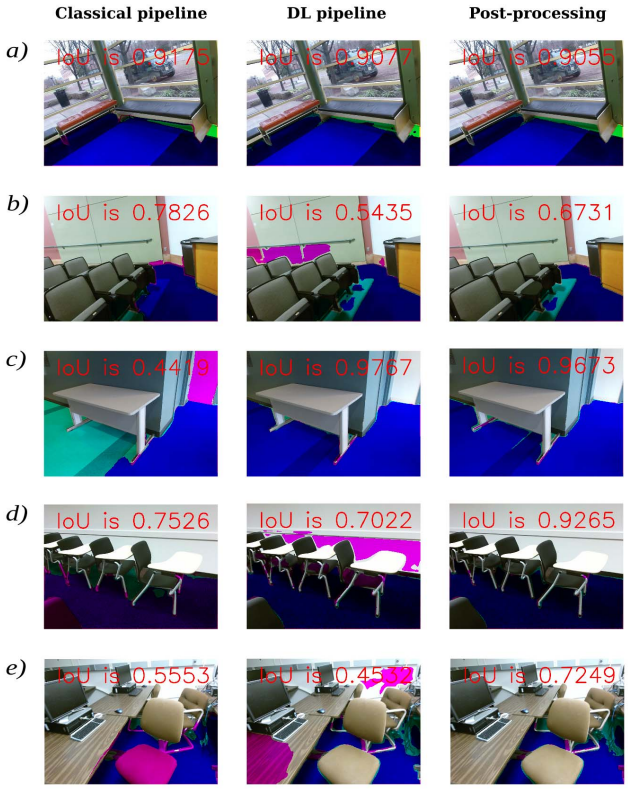
Fig. 6. Full pipeline overview. Masks from 'classical' and DL branches are combined together and post-processed.

D. *Fusion Scheme*

Both classical and DL solutions failed in some cases. In order to additionally refine the quality of segmentation maps, we decided to build a fusion scheme shown in Fig 6.

An RGB image is processed separately by the classical and the DL branches. The classical branch includes SLIC superpixeling, obtaining edge maps, RAG constructing and RAG hierarchical merging. These basic steps repeat with various parameters that provide many segmentation masks, that are summed together. The binary mask obtained by thresholding of the sum is the output mask from the classical branch.

The DL branch consists of two neural networks and independently predicts segmentation masks for the input image. They are also added together. Finally, two outputs from the both branches are combined and the post-processing



using texture feature analysis is implemented. A slight edge refinement is applied in the very end.

III. Experiments and Results

A. Datasets

We worked with several datasets. In order to train CNNs, we used 1449 images from NYUDv2 [28]; 10329 images from the SUN-RGB-D [29-31] and 8880 images from the SUN-RGB-D with NYUD removed. The target dataset was a set of 21 hand-labeled images acquired for evaluation purposes.

B. Results

To evaluate the results of segmentation we used Intersection over Union (IoU) [32]. The best result was achieved with merging of 3 masks (two from the neural networks and one summed mask from the classical pipeline) and applying the post-processing based on texture feature analysis in the end. All intermediate IoU values are shown in the table below.

TABLE I. The Result Evaluation

Mask obtained with:	***IoU***
Classical branch	0.5442
Refinenet	0.7837
FastFCN	0.7893
Deep learning branch	0.7939
Classical + deep learning branches	0.7977
Full pipeline	**0.8013**

Fig. 7. Examples of segmentation masks obtained with classical pipeline, deep learning pipeline and as a result of their combination and post-processing. a) Both classical and deep learning pipelines work well. b) Classical pipeline outperforms the deep learning approach. c) Deep learning pipeline works better than the classical one. d) Both classical and deep learning pipelines work fine and post-processing makes an improvement. e) Both classical and deep learning pipelines work bad, post-processing is used. Color legend in the figure: dark blue is true positive, magentra is false positive, cian is false negative.

Fig. 7 contains some examples of floor segmentation obtained with different setups. As expected, deep learning solution handles more challenging cases better than classical computer vision pipeline. However, for some images developed image analysis procedure provides quite competitive results or even outperforms CNN-based solution. This is explained by the size and quality of the training data, which is crucial for DL-based methods applied for typical computer vision tasks. Finally, the proposed post-processing step based on feature crafting allows refining the quality of segmentation maps.

IV. Conclusions

In this work, we analyzed the problem of room floor segmentation. We have applied both classical computer vision and deep learning techniques for this task. Firstly, we constructed a custom classical pipeline based on superpixels, region adjacency graphs, and graph hierarchical merging. Secondly, we picked two typical CNN architectures and compared their output predictions. Finally, we have built a fusion scheme to combine outputs from two branches and applied post-processing based on textural features analysis. It is clearly seen from the conducted experiments that the proper combination of computer vision methods always gives the best outcomes. In the future, we are planning to additionally improve the segmentation quality and to integrate the developed pipeline into a mobile application.

References

[1] F. Salzenstein and C. Collet, “Fuzzy Markov Random Fields versus Chains for Multispectral Image Segmentation,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 28, no. 11, pp. 1753–1767, 2006.

[2] J. Bruce, T. Balch and M. Veloso, "Fast and inexpensive color image segmentation for interactive robots," Proceedings. 2000 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2000) (Cat. No.00CH37113), Takamatsu, Japan, 2000, pp. 2061-2066 vol.3.

[3] O. J. Tobias and R. Seara, "Image segmentation by histogram thresholding using fuzzy sets," in IEEE Transactions on Image Processing, vol. 11, no. 12, pp. 1457-1465, Dec. 2002.

[4] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua and S. Süsstrunk, "SLIC Superpixels Compared to State-of-the-Art Superpixel Methods," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 34, no. 11, pp. 2274-2282, Nov. 2012.

[5] D. Stutz, A. Hermans, and B. Leibe, “Superpixels: An evaluation of the state-of-the-art,” Computer Vision and Image Understanding, vol. 166, pp. 1–27, 2018.

[6] S. Minaee, Y. Boykov, F. Porikli, A. Plaza, N. Kehtarnavaz, D. Terzopoulos, "Image Segmentation Using Deep Learning: A Survey," [online], Available: arXiv:2001.05566v2.

[7] Y. Artan, "Interactive Image Segmentation Using Machine Learning Techniques," In Canadian Conference on Computer and Robot Vision, St. Johns, NL, 2011, pp. 264-269.

[8] J. Zhang, Z. H. Tang, W. H. Gui, Q. Chen, and J. P. Liu, , “Interactive image segmentation with a regression based ensemble learning paradigm” In *Frontiers of Information Technology & Electronic Engineering*, vol. 18, no. 7, 2017, pp. 1002-1020

[9] A. Madabhushi and G. Lee, “Image analysis and machine learning in digital pathology: Challenges and opportunities,” Medical Image Analysis, vol. 33, pp. 170–175, 2016.

[10] D. C. Lee, M. Hebert, and T. Kanade. “Geometric Reasoning for Single Image Structure Recovery”. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2009.

[11] X.-N. Cui, Y.-G. Kim, and H. Kim. “Floor Segmentation by Computing Plane Normals from Image Motion Fields for Visual Navigation”. In International Journal of Control, Automation, and Systems, 7(5):788-798, 2009.

[12] E. Fazl-Ersi and John K. Tsotsos. “Region Classification for Robust Floor Detection in Indoor Environments ”. In Proceedings of the 6th International Conference on Image Analysis and Recognition, pages 717-726, 2009.

[13] M. V. D. Bergh, X. Boix, G. Roig, and L. V. Gool, “SEEDS: Superpixels Extracted Via Energy-Driven Sampling,” International Journal of Computer Vision, vol. 111, no. 3, pp. 298–314, 2014.

[14] X. Wei, Q. Yang, Y. Gong, N. Ahuja and M. Yang, "Superpixel Hierarchy," in IEEE Transactions on Image Processing, vol. 27, no. 10, pp. 4838-4849, Oct. 2018.

[15] A. Tremeau and P. Colantoni, "Regions adjacency graph applied to color image segmentation," in IEEE Transactions on Image Processing, vol. 9, no. 4, pp. 735-744, April 2000.

[16] M. Nixon and A. S. Aguado, Feature extraction and image processing. Oxford: Academic Press, 2012.

[17] M. Juneja and P. S. Sandhu, “Performance Evaluation of Edge Detection Techniques for Images in Spatial Domain,” International Journal of Computer Theory and Engineering, pp. 614–621, 2009.

[18] T. Ojala, M. Pietikäinen, and T. Maenpaa, “Multiresolution gray-scale and rotation invariant texture classification with local binary patterns,” IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 24, no. 7, pp. 971–987, 2002.

[19] C. K. Reddy and B. Vinzamuri, “A Survey of Partitional and Hierarchical Clustering Algorithms,” Data Clustering, pp. 87–110, Mar. 2018.

[20] J. Jaworek-Korjakowska and P. Kleczek, “Region Adjacency Graph Approach for Acral Melanocytic Lesion Segmentation,” Applied Sciences, vol. 8, no. 9, p. 1430, 2018.

[21] J. Long, E. Shelhamer, and T. Darrell, “Fully convolutional networks for semantic segmentation,” *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[22] H. Noh, S. Hong, and B. Han, “Learning Deconvolution Network for Semantic Segmentation,” *2015 IEEE International Conference on Computer Vision (ICCV)*, 2015.

[23] O. Ronneberger, P. Fischer, and T. Brox, “U-Net: Convolutional Networks for Biomedical Image Segmentation,” *Lecture Notes in Computer Science Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, pp. 234–241, 2015.

[24] V. Nekrasov, C. Shen, I. Reid, "Light-Weight RefineNet for Real-Time Semantic Segmentation," [online], Available: **Error! Reference source not found.**.

[25] H. Wu, J. Zhang, K. Huang, K. Liang, Y. Yu, "FastFCN: Rethinking Dilated Convolution in the Backbone for Semantic Segmentation," [online], Available: arXiv:1903.11816v1 [cs.CV].

[26] A. Humeau-Heurtier, "Texture feature extraction methods: A survey," In. IEEE Access 7, 2019, pp. 8975-9000.

[27] S. Öztürk, and A. Bayram, "Application of feature extraction and classification methods for histopathological image using GLCM, LBP, LBGLCM, GLRLM and SFTA," In Procedia computer science 132, 2018, pp. 40-46.

[28] 28 N. Silberman, D. Hoiem, P. Kohli, and Fergus R, “Indoor Segmentation and Support Inference from RGBD Images,” In: Fitzgibbon A., Lazebnik S., Perona P., Sato Y., Schmid C. (eds) Computer Vision – ECCV 2012. ECCV 2012. Lecture Notes in Computer Science, vol 7576. Springer, Berlin, Heidelberg.

[29] S. Song, P. L. Samuel, and X. Jianxiong, "Sun rgb-d: A rgb-d scene understanding benchmark suite," In Proceedings of the IEEE conference on computer vision and pattern recognition, 2015.

[30] A. Janoch, S. Karayev, Y. Jia, J. T. Barron, M. Fritz, K. Saenko, and T. Darrell, “A category-level 3-d object dataset: Putting the kinect to work,” In ICCV Workshop on Consumer Depth Cameras for Computer Vision, 2011.

[31] J. Xiao, A. Owens, and A. Torralba, “SUN3D: A database of big spaces reconstructed using SfM and object labels,” In ICCV, 2013.

[32] H. A. Alhaija, S. K. Mustikovela, L. Mescheder, A. Geiger, C. Rother, "Augmented Reality Meets Computer Vision : Efficient Data Generation for Urban Driving Scenes," [online], Available: arXiv:1708.01566 [cs.CV].**Error! Reference source not found.**

IEEE Third International Conference on Data Stream Mining & Processing
August 21-25, 2020, Lviv, Ukraine

Automated Measurement of Bone Thickness on SCT Sections and Other Images

Radiy Radutniy
Chief scientific specialist
National Aviation University of Ukraine
Kyiv, Ukraine
radiy@yahoo.com

Alina Nechyporenko
Systems Engineering Department
Kharkiv National University of Radio Electronics
Kharkiv, Ukraine
alinanechiporenko@gmail.com

Victoriia Alekseeva
Department of Histology, Cytology, Embryology of Kharkiv National Medical University
Kharkiv, Ukraine
vik13052130@i.ua

Ganna Titova
Rector of Kharkiv International Medical University
Kharkiv, Ukraine
dr.a.titova@ukr.net

Dmytro Bibik
National Technical University "Kharkiv Polytechnic Institute"
Kharkiv, Ukraine
bibik.dmytro@gmail.com

Vitaliy V. Gargin
Department of Pathological Anatomy
Kharkiv National Medical University
Kharkiv, Ukraine
vitgarg@ukr.net

***Abstract* — Despite the importance of image assessment in the biomedical field (radiation diagnosis, otolaryngology, dentistry, pathology, etc.) there is no clearly defined unified image processing algorithm to date, the result of examination and interpretation of these images often depends on the physician's qualifications and experience. The purpose of our study was implementation of the computer vision system for evaluation of CT images for obtaining impersonal real data. Thus, in the course of the study, a number of medical goals were solved, which included the need to measure bone thickness on radiograms, tomograms, SCTs and other images. This is especially relevant for measuring the minimum thickness, as a thin bone can be easily damaged by a medical instrument during surgery (for example, damage to the upper wall of the maxillary sinus with damage to the orbit). The suggested method could be used for automated measurement of bone thickness of the walls of the paranasal sinuses. The obtained results can be implemented in the work of doctors in a number of specialties (therapists, radiologists, otolaryngologists, dentists, ophthalmologists, neurosurgeons) for study many biomedical images.**

Keywords— SCT, analysis, paranasal sinuses, bone thickness, automated image.

I. INTRODUCTION

Despite the importance of image assessment in the biomedical field (radiation diagnosis, otolaryngology, dentistry, pathology, etc.[1,2,3]), there is no clearly defined unified image processing algorithm to date, the result of examination and interpretation of these images often depends on the physician's qualifications and experience. Single attempts to unify image assessment algorithms did not bring the desired result [4,5].

The use of traditional measurement methods, such as measuring instruments, is impossible because the image can be scaled up or down. In addition, the size of the image is highly dependent on the resolution of the screen or printer. Measuring the thickness of the screen in pixels is not very informative for the doctor, and the conversion of pixels in millimeters requires knowledge of the resolution of the screen (printer), mathematical operations and is associated with the possibility of an error [6,7].

Otolaryngology is one of the medical fields where determination of the correct algorithm for imaging is of paramount importance. Determining the individual anatomical structure of the paranasal sinuses is a major challenge for doctors of various specialties (otolaryngologists, dentists, plastic surgeons, neurosurgeons and ophthalmologists (in terms of the risk of intracranial and intraorbital spread of the inflammatory process) as in many other biomedical research [8,9].

It should be noted that today the number of diseases of the upper respiratory tract, in particular rhinosinusitis - inflammatory processes of the paranasal sinuses, is constantly increasing. There is a tendency towards an increase in the proportion of chronicity of acute rhinosinusitis. Today spiral computed tomography is the "gold standard" for diagnosing rhinosinusitis [10]. This method of research helps to examine, as precisely as possible, in vivo, non-invasively the features of the structure of PNS, to determine the presence or absence of inflammatory or other pathological changes in this area. At the same time, the study of the thickness of the walls of the PNS causes a number of technical problems, in particular the correct choice of examination points, which of the criteria should be preferred - the maximum thickness or minimum one and how to measure correctly and get an accurate and reliable result [11].

Considering the above, the **purpose** of our study was implementation of the computer vision system for evaluation of CT images for obtaining impersonal real data.

II. MATERIALS AND METHODS

Spiral computer tomograms of 5 subjects of different sex aged 25-45 without any signs of ENT abnormalities who underwent SCT examination due to reasons not associated with ENT diseases (suspected stroke which was not confirmed). All the patients agreed to participate in the study. The patients were also examined by an otolaryngologist for final exclusion of ENT abnormalities. SCT findings are described by a radiologist.

This study was performed on a ToshibaAquilion-64 high performance CT scanner with technical characteristics specifically designed to better visualize bodily structures.

978-1-7281-3214-3/20/$31.00 ©2020 IEEE

The wide aperture of the unit (70 cm) and automatic correction of radiation dose depending on the patient's weight and the scanning area in combination with a high speed of reconstruction provide great opportunities for clinical research [12].

We have proposed a method for automated measurement of bone thickness on images. The method is based on the use of machine vision systems [13]. The software analyzes the contour of the image, measures the thickness of the image, denoting bone tissue, setting the size of the image in typical places. If an element of a predetermined size is present in the image, an automatic conversion to millimeters or other units is performed.

TABLE I. DISTRIBUTION OF PATIENTS BY GENDER, AGE

Patient order number	Age	Gender
1.	*40*	*F*
2.	*27*	*F*
3.	*38*	*M*
4.	*25*	*F*
5.	*40*	*F*

III. RESULTS AND DISCUSSION

Typically, DICOM is used to view images. There are many kinds of software for viewing images in this format, for example, Dicom Viewer, Dicom File Viewer, Radiant Dicom Viewer, MictroDicom. Some of them have tools for measuring objects in images, some do not, but in any case, measurements are done manually, which is time-consuming. In addition, all manual operations are associated with "human factor" risks. The physician may do mistakes by incorrectly measuring the size of the object, in particular, the thickness of the bone; incorrectly select the image scale, fail to notice the area where the bone thickness is small. Getting familiar with software also requires spending time on an auxiliary specialty, rather than on improving medical qualification.

Thus, there is a need to automate the process of finding areas of minimum bone thickness and its measurement.

The first part of the task does not present much difficulty. Radiograms and SCT images are usually black and white, with clear contours of soft tissues and bones. Color brightness characterizes densitometric indices of the material. So, according to the Hounsfield scale, soft tissues have +40 units and bones +400 or higher. In the images, soft tissues look like dull areas and the bones are bright white ones. The boundary between the regions is clear.

Threshold gain operations can convert images into fully black and white without halftone. In this case, the dull areas will become black and blend with the background, and bright on the contrary, will become white and will have even clearer outlines.

The following operation is aimed at removing the so-called "noise" - random spots, irregularities, blurriness and inscriptions on the image. It is done by overlaying a filter based on the Gaussian function.

$$g(x) = ae^{\frac{(x-b)^2}{2c^2}} \quad (1)$$

where a, b, c parameters are arbitrary real numbers. As part of this study, they were selected experimentally.

After Gaussian blur, the image becomes convenient to search first, the bone itself, and secondly, areas where the thickness of the bones is minimal.

Finding an area that shows a bone is not difficult. White pixel color means bone, black means background or soft tissue.

Finding areas of minimum thickness does not present difficulties. It is done by sequentially bypassing pixels, constructing segments, accumulating them in an array, and finding the minimum values of the array.

In this study, the search was performed only on an array of horizontal segments. This method is acceptable for thin bones, but can produce inaccurate results on bones of considerable thickness. In the future, the method will be supplemented.

Finding areas of minimum thickness is a bit trickier. It is done by sequentially bypassing pixels, constructing segments, accumulating them in an array, and finding the minimum values of the array.

The values found are displayed on the primary image as a segment and a digital value of the bone thickness in the area.

The thickness is measured in pixels. To convert the thickness into millimeters or other units, a calibration segment of a known length is used. In this study, the segment was 25 mm long and was marked green in the upper left portion of the image. The search for the plotted segment was performed in the same way as the above algorithm, except that instead of white, the software searched for pixels in a certain range, from dark green to light green.

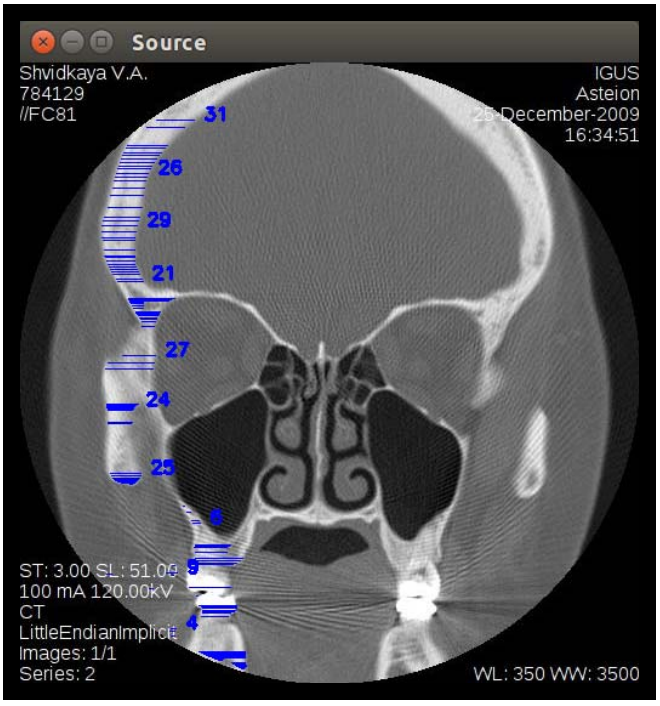


Fig. 1. An exanple of the measured thickness of the wall of skull

The found segment was automatically marked with a successful search tick - a yellow circle. If the segment could

not be found, the message “Cannot recognize measure” was displayed.

Counting the number of pixels per segment helps easily find the scale factor and convert the resulting bone thickness values from pixels to millimeters or other units.

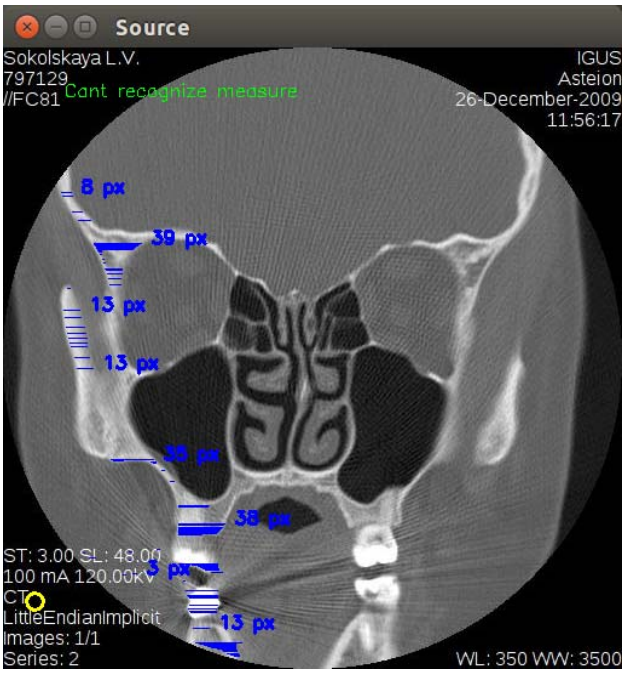


Fig. 2. An exanple of the measured thickness of the wall of skull

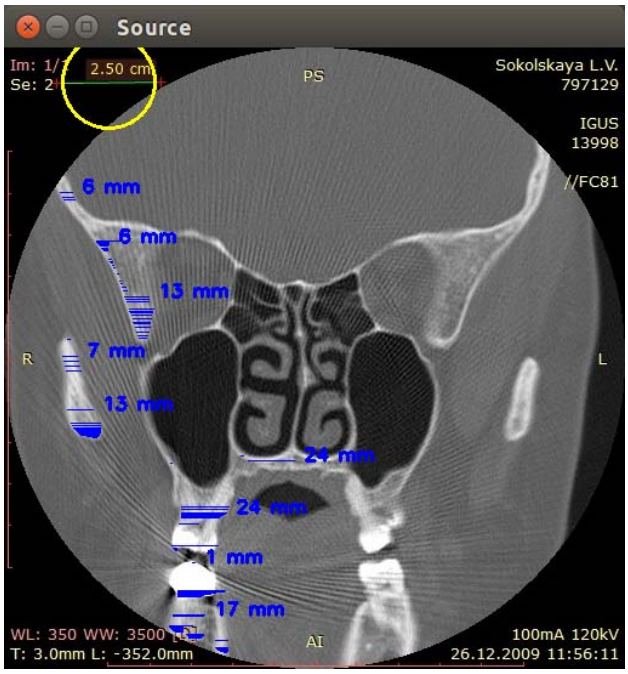


Fig. 3. An example of image with an existing object for calibration.

The technology is being tested. If successful, it is planned to improve technology, in particular, to improve measurement accuracy, as well as to evaluate the densitometric characteristics of bone, in particular, its density.

Fig. 4. The original image has been traslated to HSV format

The image does not have a calibration object, so the thickness is measured in pixels.

The units of measure (pixels) and the warning of the absence of the object to be calibrated are displayed.

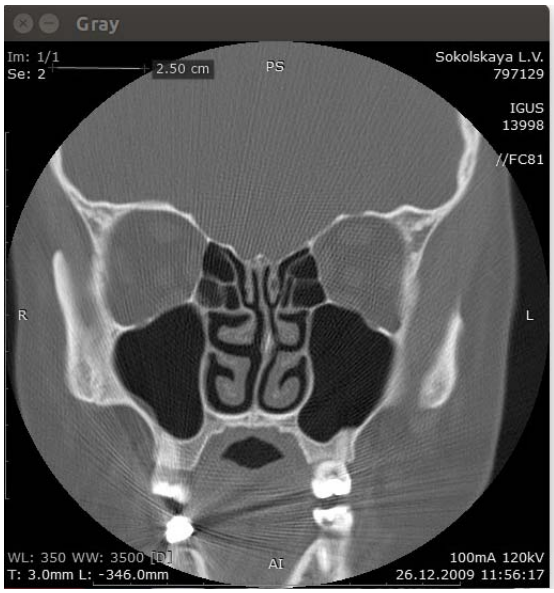


Fig. 5. Color image converted to the grayscale

For the calibration, we used a green line of a known length (25 mm), which is visible in the upper left corner of the image. The line is surrounded by a yellow circle, which indicates the successful automatic recognition of the item for calibration. The thickness of the bones is shown in millimeters. Conversion into millimeter is also fully automatic.

Thus, in the course of the study, a number of medical goals were solved, which included the need to measure bone thickness on radiograms, tomograms, SCTs and other images. This is especially relevant for measuring the minimum thickness, as a thin bone can be easily damaged by a medical instrument during surgery (for example, damage to

the upper wall of the maxillary sinus with damage to the orbit).

Fig. 6. The image was blured using a Gaussian filter

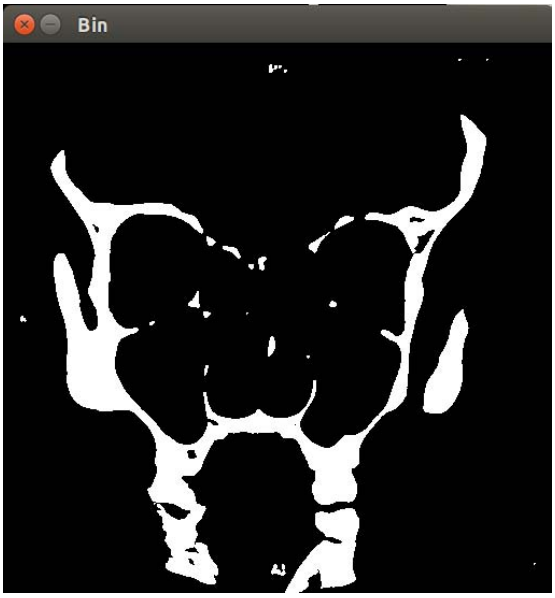


Fig. 7. The image was converted to black and white with a brightness threshold of 130 units. All pixels with less brightness appeared in black, all higher pixels turned to white.

Automated image analysis methods are becoming more and more important to extract and quantify image features in microscopy-based biomedical studies [14,15,8,16] and in anthropometry also [17], research of new material [18]. Work with images is important mathematical problem often [19,20].

It should be noted that this study is not the first one. To date, the Global Osteitis Scale is known, where researchers pay specific attention to the minimum size and, based on the data obtained, classify the sinuses according to the severity of the destructive changes in the walls of the PNS. However, it is not always possible to determine the minimum thickness correctly, and additional measurements will inevitably be associated with spending more time consulting and investigating PNCs [21].

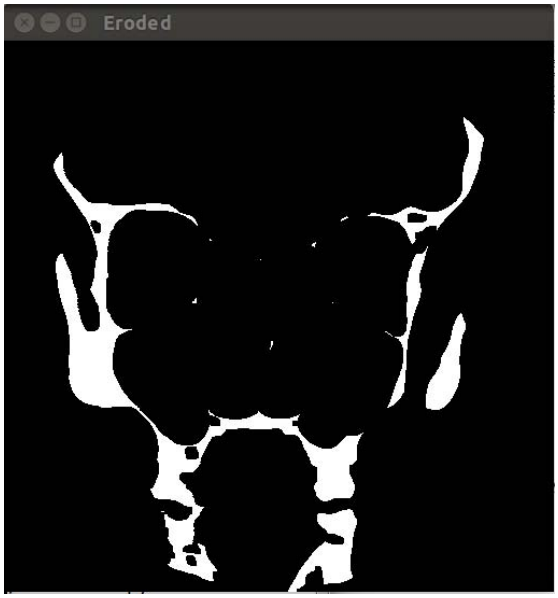


Fig. 8. The image was cleared from noise and noisy components

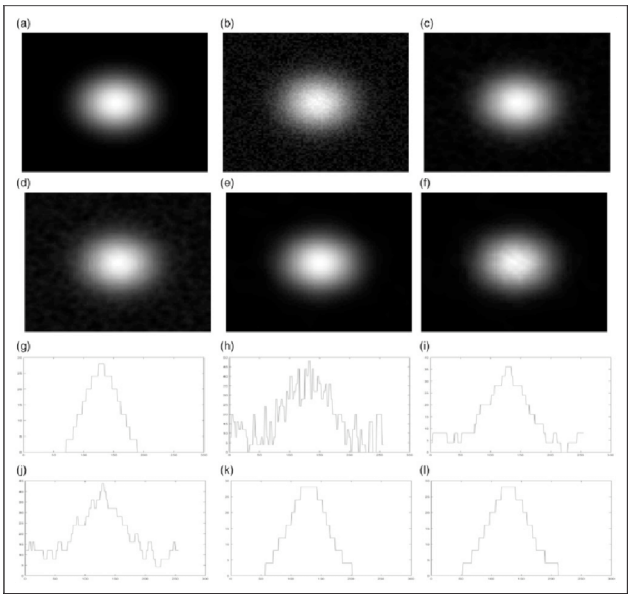


Fig.9 The results of restoring the Gaussian blur function.

Figure 9 showing the results of restoring the Gaussian blur function. Example 9 using the total variation model for different ðu tv Þ, mean curvature u mc ð Þ and the fractional model for different u f.

It is clear from the cross section images (g)-(l) that tv does not achieve very good results for this smooth function, as expected.

Competition between mean curvature and the fractional model is close but the numerical results P u mc ð Þ ¼ 26:32 and P u 1:3 f ¼ 26:69 demonstrate that the fractional model outperforms mc. (a) True image. (b) Received image z. (c) u 10 À3 tv. (d) u 10 À6 tv. (e) u mc. (f) u 1:3 f. (g) True Image. (h) Received Image z. (i) u 10 À3 tv. (j) u 10 À6 tv. (k) u mc. (l) u 1:3 f .

Gaussian blur is widely used for image processing for a variety of purposes. For example, in "A new image deconvolution method with fractional regularization" [21].

It should be noted that this study is not the first one. To date, the Global Osteitis Scale is known, where researchers pay specific attention to the minimum size and, based on the data obtained, classify the sinuses according to the severity of the destructive changes in the walls of the PNS. However, it is not always possible to determine the minimum thickness correctly, and additional measurements will inevitably be associated with spending more time consulting and investigating PNCs [22]. It is the thickness of the bone that determines the features of the course of the inflammatory process in the paranasal sinuses, allowing to calculate precisely the likelihood of the spread of inflammation to adjacent anatomical areas (brain or orbit) with the development of complications. The thickness of the lower wall of the maxillary sinus is of great importance in dentistry during dental implantation, because it helps to properly dose the load during the performed procedure and prevent the development of odontogenic maxillary sinusitis. During otolaryngological surgeries, knowledge of bone thickness will be useful to prevent iatrogenic complications of operations associated with the accidental destruction of the wall of the PNS.

In addition, the method of calculating the uncertainty of the thickness of the bone was used, showing the values of all indicators reliable for the given value. Despite its accuracy, this method has some drawbacks. First of all, it requires a considerable amount of time, which greatly complicates the work of a doctor, especially in the present time, in the conditions of staff load and a large number of patients.

CONCLUSION

The suggested method could be used for automated measurement of bone thickness of the walls of the paranasal sinuses. The obtained results can be implemented in the work of doctors in a number of specialties (therapists, radiologists, otolaryngologists, dentists, ophthalmologists, neurosurgeons) for study many biomedical images.

REFERENCES

[1] D.S. Avetikov, O.P. Bukhanchenko, I.O. Ivanytskyi, V.V. Aipert, D.V. Steblovskyi, "Perspectives for applying the additional study methods for diagnostics optimization of postoperative hypertrophic scars of the head and neck", Wiad Lek., vol. 71(3 pt 1), pp. 470-473, 2018.

[2] N. Hyriavenko, M. Lyndin, K. Sikora, A. Piddubnyi, L. Karpenko, O. Kravtsova, D. Hyriavenko, O. Diachenko, V. Sikora, A. Romaniuk, "Serous Adenocarcinoma of Fallopian Tubes: Histological and Immunohistochemical Aspects," J Pathol Transl Med, vol. 53, pp.236-243, 2019.

[3] V.V. Alekseeva, A.V. Lupyr, N.O. Urevich, R.S. Nazaryan, V.V. Gargin, "Significance of anatomical variations of maxillary sinus and ostiomeatal components complex in surgical treatment of sinusitis," J. Novosti Khirurgii, vol. 27, pp. 168–176, Mar.-Apr. 2019. (in Russian).

[4] M. Brinkmann, D. Lutkemeyer, F. Gudermann, J. Lehmann, "New technologies for automated cell counting based on optical image analysis ;The Cellscreen'," Cytotechnology, vol. 38(1-3), pp. 119-127, Jan 2002.

[5] S. Wienert, D. Heim, M. Kotani, B. Lindequist, A. Stenzinger, M. Ishii, P. Hufnagl, M. Beil, M. Dietel, C. Denkert, F. Klauschen. "CognitionMaster: an object-based image analysis framework," Diagn Pathol, vol. 27; pp. 8-34, Feb 2013.

[6] S. Krivenko, D. Demchenko, I. Dyogtev, V. Lukin, "A Two-Step Approach to Providing a Desired Quality of Lossy Compressed Images", Advances in Intelligent Systems and Computing 1113 AISC, pp. 482-491, 2020.

[7] S. Krivenko, S. Abramov, V. Lukin, B. Vozel, K. Chehdi, "Lossy DCT-based compression of remote sensing images with providing a desired visual quality," Proc. SPIE 11155, Image and Signal Processing for Remote Sensing XXV, 2019.

[8] V. Romanenko, L. Podrigalo, S. Iermakov, "Functional state of martial arts athletes during implementation process of controlled activity - Comparative analysis", Physical Activity Review, vol. 6, pp. 87-93, 2018.

[9] A. Romaniuk, M. Lyndin, V. Sikora, Y. Lyndina, K. Panasovska, "Histological and immunohistochemical features of medullary breast cancer", Folia medica cracoviensia, vol. 2, pp. 41-48, 2015.

[10] V.V. Gargin,V.V. Alekseeva, A.V. Lupyr, N.O. Urevich, R.S. Nazaryan, V.M. Cheverda, "Correlation between the bone density of the maxillary sinus and body mass index in women during the menopause", J. Problemi Endokrinnoi Patologii, vol 2 (68), pp. 20-26, 2019. (in Russian).

[11] A.S. Nechyporenko, S.S. Kryvenko, V. Alekseeva, A. Lupyr, N. Yurevych, R.S. Nazaryan, V.V. Gargin, "Uncertainty of measurement results for anatomical structures of Paranasal sinuses," in Proc. 8th Mediterr. Conf. on Embedded Computing, Budva, Montenegro, pp. 570-574, 2019

[12] A. Nechyporenko, V. Reshetnik, V. Alekseeva, N. Yurevych, R. Nazaryan, V. Gargin, "Implementation and analysis of uncertainty of measurement results for anatomical images" 2020 IEEE 40th International Conference on Electronics and Nanotechnology, ELNANO, 2020.

[13] V. Gargin, R. Radutny, G. Titova, D. Bibik, A. Kirichenko, O. Bazhenov, "Application of the computer vision system for evaluation of pathomorphological images 2020 IEEE 40th International Conference on Electronics and Nanotechnology", ELNANO, 2020.

[14] S. Krivenko., A. Pulavskyi. "Accuracy improvement of noninvasive determination of glucose concentration in human blood", Experience of Designing and Application of CAD Systems in Microelectronics (CADSM), pp. 181-183, 2013.

[15] Sean M. Johnson, Adam L. Honeybrook, Vaibhav H. Ramprasad, Ralph Abi Hachem, David W. Jang, "Radiodensity of the ostiomeatal complex in recurrent acute rhinosinusitis", Otolaryngol Head Neck Surg, vol. 157 (5),pp. 887-890, Nov 2017.

[16] S.A. Krivenko, A.A. Pulavskyi, S.S. Krivenko, L.S. Kryvenko, "Many-To-Many Linear-Feedback Shift Model for Training of Artificial Neural Network in Dentistry", 2019 IEEE 39th International Conference on Electronics and Nanotechnology, ELNANO, Proceedings 8783543, pp. 429-434, 2019.

[17] .Y. Kuzmina, G.I. Hubina-Vakulik, G.J. Burton, "Placental morphometry and Doppler flow velocimetry in cases of chronic human fetal hypoxia", European Journal of Obstetrics and Gynecology and Reproductive Biology, vol.120, pp. 139-145, June 2005.

[18] M. Pogorielov, A. Kravtsova, G. Reilly, V. Deineka, G. Tetteh, O. Kalinkevich, O. Pogorielova, R. Moskalenko, G. Tkach, "Experimental evaluation of new chitin–chitosan graft for duraplasty", Journal of Materials Science: Materials in Medicine, vol 28(2), pp. 34, 2017.

[19] D. Bilyk, N Feldheim, "The Two-Dimensional Small Ball Inequality and Binary Nets", Journal of Fourier Analysis and Applications, vol 23(4), pp. 817-833, 2017.

[20] D. Bilyk, L. Markhasin, "BMO and exponential Orlicz space estimates of the discrepancy function in arbitrary dimension" Journal d'Analyse Mathematique, vol 135(1), pp. 249-269, 2018.

[21] B. Williams, J. Zhang, Ke Chen. "A new image deconvolution method with fractional regularisation", Journal of Algorithms & Computational Technology, vol. 10(4), pp. 265-276, 2016.

[22] B.B. Hoyer, C.H. Ramlau-Hansen, M. Vrijheid, D. Valvi, H.S. Pedersen, V. Zviezdai, B.A. Jonsson, C.H. Lindh, J.P. Bonde, G. Toft, "Anthropometry in 5- to 9-Year-Old Greenlandic and ukrainian children in relation to prenatal exposure to perfluorinated alkyl substances", Environ Health Perspect., vol.123(8), pp. 841-6, Aug. 2015.

IEEE Third International Conference on Data Stream Mining & Processing
August 21-25, 2020, Lviv, Ukraine

Geometric Properties of Adversarial Images

Bogdan Ivanyuk-Skulskiy
National University of Kyiv-Mohyla Academy
Kyiv, Ukraine
ivanyuk.skulskiy@ukma.edu.ua

Galyna Kriukova
National University of Kyiv-Mohyla Academy
Kyiv, Ukraine
kriukovagv@ukma.edu.ua

Andrii Dmytryshyn
School of Science and Technology
Örebro University
Örebro, Sweden
andrii.dmytryshyn@oru.se

***Abstract*—Machine learning models are now widely used in a variety of tasks. However, they are vulnerable to adversarial perturbations. These are slight, intentionally worst-case, modifications to input that change the model's prediction with high confidence, without causing a human eye to spot a difference from real samples. The detection of adversarial samples is an open problem. In this work, we explore a novel method towards adversarial image detection with linear algebra approach. This method is built on a comparison of distances to the centroids for a given point and its neighbors. The method of adversarial examples detection is explained theoretically, and the numerical experiments are done to illustrate the approach.**

***Index Terms*—adversarial learning, autoencoder, artificial neural network**

I. Introduction

In recent years, machine learning and, in particular, deep learning (DL) models have improved their performance in various tasks, e.g., image classification, speech recognition, natural language processing. However, even the state-of-the-art models are vulnerable to so called adversarial perturbations. These perturbations aren't visible for a human eye but, applied to a correctly classified sample, lead to misclassification of the sample [1]–[5]. Obviously, such an issue may cause serious consequences in the applications where safety and security are priority, for example, autonomous driving and medical image processing.

There have been recent attempts to explain this phenomenon, see e.g., [1], but a consistent theory is still missing.

In this paper, we propose a new approach to adversarial image detection based on linear algebra methods. Our approach relies on the assumption, that an adversarial perturbation pushes a sample away from a manifold, where the correctly classified samples are concentrated. This allows us to use distributions of certain distances for detecting adversarial samples.

II. Background

We provide some basic background on machine learning, Generative Adversarial Networks (GAN), and autoencoders.

A. Deep Learning Classifiers

A deep learning classifier can be expressed as a mapping $F(x;\theta_c) : \mathbb{R}^{\mathrm{D}} \to \mathbb{R}^{\mathrm{C}}$, where $\mathbb{R}^{\mathrm{D}}$ is the input space, $\mathbb{R}^{\mathrm{C}}$ is the space of all the classes, and θ_c is a vector of trainable parameters. In this work we use neural networks with sigmoid output layer. For a given input x a predicted label is either 0 or 1, and is denoted by $\hat{y} = \arg\max_{i\in[0;1]} F(x;\theta_c)_i$, where $F(x;\theta_c) = \mathrm{sigmoid}(\theta_c \cdot x + b_c)$ is a vector output from neural network bounded by 0 and 1. Common training objective in this setting is to minimize the binary cross-entropy (BCE) loss, defined as:

$$\mathcal{L}_{\mathrm{BCE}}(x,y) = -y\log F(x;\theta_c) - (1-y)\log(1-F(x;\theta_c))$$

for a single input pair (x, y).

B. Adversarial Networks

Originally adversarial neural networks were represented as a multilayer perceptrons [6] that were composed from two parts: Generator network $G(z;\theta_g)$ which is a differentiable function with parameters θ_g that learns the mapping from input noise variable $p_{adv}(z)$ to data space, and a discriminator (i.e., a deep learning classifier) network $D(x;\theta_d)$ with parameters θ_d that represents the probability of input x coming from real distribution rather than adversarial sample distribution. In general we are aiming to maximize correct classification of training and generated samples for discriminator network and simultaneously minimize $\log(1 - D(G(z;\theta_g);\theta_d))$:

$$\min_G \max_D V(G,D) = \mathbb{E}_{x\sim p_{data}(x)}[\log D(x;\theta_d)] + \mathbb{E}_{z\sim p_{adv}(z)}[\log(1 - D(G(z;\theta_g);\theta_d))]$$

given $V(G, D)$ as a value function of a two-player minimax game.

In 2016 Radford, Metz, and Chintala [7] introduced an extension of GAN described above, known as Deep Convolutional Generative Adversarial Network (DCGAN). DCGAN consists of a discriminator network, made of stridden convolutional layers, batch normalization layers, and LeakyReLU activations; as well as a generator network, comprised of convolutional-transpose layers, batch normalization layers, and ReLU activations.

C. Autoencoders

An autoencoder is a type of neural network that aims to copy its input to its output. The network is composed of two parts: an encoder function $h = \mathrm{Enc}(x;\theta_{enc})$ and a decoder function that performs reconstruction $r = \mathrm{Dec}(h;\theta_{dec})$. Autoencoders are designed to copy an input approximately, meaning the model is forced to prioritize which aspects of the input should be copied, it often learns useful properties of data [8].

Galyna Kriukova is grateful to the Charity Foundation "Believe in Yourself" for financing her sabbatical.

Fig. 1. Sample images from real MNIST (1st row) and CelebA (3rd row) datasets and adversarial images for these datasets (2nd and 4th rows respectively) generated by DCGAN for 10 epochs.

$$\mathcal{L}(x, \mathrm{Dec}(\mathrm{Enc}(x;\theta_{enc});\theta_{dec})) \\ = \min_x ||x - \mathrm{Dec}(\mathrm{Enc}(x;\theta_{enc});\theta_{dec})||^2$$

given $\mathcal{L}$ a loss function penalizing $\mathrm{Dec}(\mathrm{Enc}(x;\theta_{enc});\theta_{dec})$ being dissimilar from input x.

III. Methods

Many learning algorithms exploit the idea that data concentrates around a low dimensional manifold. In the following method, we use geometrical characteristics of such a manifold for detecting adversarial examples.

We make the following assumptions on our data:

(a) Data points belong to a smooth manifold;
(b) For every point on this manifold (data point or not), the centroid of its k nearest neighbors (KNN) is approximately at the same distance from this point;
(c) For a point that does not belong to the manifold the set of its KNN coincides with the set of KNN of its projection on the manifold.

We suggest a method for detecting adversarial examples by comparing the distances to the centroids for a given point and its KNN. This method is base on the following theoretical justification.

Let x be a point outside of a given manifold, and M_x be the centroid of its k nearest neighbors. For the Euclidean distance we have

$$(M_x x)^2 = (x\mathrm{Pr}_x)^2 + (\mathrm{Pr}_x M_x)^2,$$

where Pr_x is the projection of x on the manifold.

From our assumption (c) we have that if M_x is a centroid for KNN of x then M_x is also a centroid for KNN of Pr_x. Combining this with the assumption (b) we have that $\mathrm{Pr}_x M_x = M'_{x'} x'$, for any points x' and its KNN's centroid $M'_{x'}$. Therefore $(M_x x)^2 = (x\mathrm{Pr}_x)^2 + (M'_{x'} x')^2$.

Now since X is outside of our manifold we have $x\mathrm{Pr}_x > 0$ and thus $M_x x > M'_{x'} x'$. Moreover,

$$x\mathrm{Pr}_x = \sqrt{(M_x x)^2 - (M'_{x'} x')^2}.$$

Therefore we may use the following statement as a criterion if a given example is adversarial or not:

For every point outside of a given (differential) manifold and large enough integer k, the distance to the centroid of its k nearest neighbors is significantly larger than the distance from a point on this manifold to the centroid of its k nearest neighbors.

Obtaining similar theoretical results with relaxed conditions (a)–(c) is a part of our future work.

The above theory results in the following method for detecting adversarial examples. For a given image x we find the set of its nearest neighbors. Then for this set, we find the centroid point M_x and the distance $M_x x$. We do the same for each neighbor-image from selected set of neighbour images, resulting in the distance distribution of neighbor images to their neighbor-sets centroids. These distances suggest us whether the input image x is adversarial or not. This procedure for our experiments is described in more details in the following section, see also Algorithm 1.

Adversarial images are generated using DCGAN introduced in Section II-B. The dimensionality of the images is reduced using autoencoders, see Section II-C for the definition.

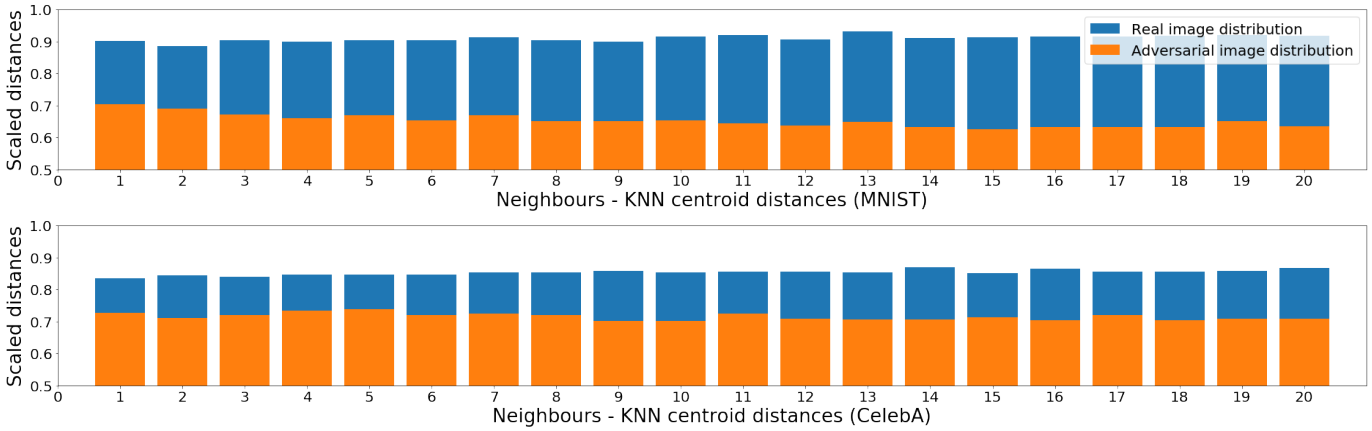


Fig. 2. Distance distribution to centroid for KNN (K = 20) of input images x_i scaled by the distance $M_{x_i}x_i$ (where M_{x_i} is a centroid point of KNN of x_i) averaged after 1000 experiments. The experiments are performed for the both real and adversarial samples

IV. Experiments

In this section we present the results of our experiments. We start by describing how the experiments were performed.

For the experiments we took **MNIST** (60000 images) [9] and **CelebA** (200000 images) [10] datasets. The pixel values of images were scaled to be in the range of $[-1.0, 1.0]$. Given this data, we have generated 10000 adversarial images both for MNIST and CelebA data using DCGAN introduced in Section II-B. Few samples of the generated adversarial images for both MNIST and CelebA datasets are presented in Fig. 1.

In order to reduce the dimensionality and keep the key features of the images, we use autoencoders trained on real images. To be exact, in the first experiment, we reduce dimensions of an MNIST sample image from $1 \times 28 \times 28$ to $8 \times 4 \times 4$ and then flatten it to obtain a 128-dimensional vector. In the second experiment, we reduce each input CelebA image from $3 \times 218 \times 178$ to $8 \times 4 \times 4$ and flatten it like in the previous experiment. After obtaining 128-dimensional vectors, we proceed as described in the following paragraph, all these steps are also formalized in Algorithm 1.

We start by finding neighbouring images of a given image x. For this purpose, we have tried to use both the Euclidean distance and cosine similarity, resulting in almost the same arrays of neighbors. We denote this array of neighbours by S_x, and sort its entries by the distance to x, starting from the nearest image. For S_x we calculate the coordinates of the centroid point M_x and the corresponding $M_x x$ distance. Our next step is to perform the same actions for each entry of the array S_x, resulting in the distance distribution of neighbor-images to the centroid of their neighbor-arrays. These distances give us an evidence whether the input image x is adversarial or not.

In practice, for each of our datasets, we have randomly selected 1000 samples x_i of both real and adversarial images, and perform Algorithm 1 for each of them. Therefore, for each x_i, we obtain the distance $M_{x_i}x_i$ and the distribution array M_{distr_i}. We divide each value from the distribution array M_{distr_i} by the distance $M_{x_i}x_i$. Each of the elements of the resulting scaled array can be interpreted as a fraction

Algorithm 1 comparing the distances to the centroids for a given point x and its KNN

Input: sample image x
Output: $M_x x$, distance distribution
Initialisation : M_{distr} empty array
$S_x \leftarrow$ search for neighbours of x
$M_x \leftarrow$ centroid point of obtained set S_x
$M_x x \leftarrow$ distance between x and M_x
for s_i from set S_x **do**
  $S_i \leftarrow$ search for neighbours of s_i
  $M_{s_i} \leftarrow$ centroid point of obtained set S_i
  $M_{s_i} s_i \leftarrow$ distance between M_{s_i} and s_i
  $M_{distr}[i] \leftarrow M_{s_i} s_i$
end for
return $M_x x$, M_{distr}

of the distance $M_{x_i}x_i$ but we also call them scaled distances. This scaling allows us to compare the results for different x_i, in particular, for adversarial versus real images. Our results are presented in Fig. 2. The two plots in Fig. 2 show that in all the cases the fractions of the corresponding $M_{x_i}x_i$ for an adversarial sample x_i (orange bar plots) are significantly smaller than the fractions of $M_{x_j}x_j$ for a real sample x_j (blue bar plots). For the MNIST dataset, the difference is about 20% - 25% and for CelebA data around, it is about 15%. Slightly different percentages for MNIST and CelebA data can be explained by the nature of images as follows. MNIST images are greyscaled and thus they have much lower variance in pixel values comparing to images taken from CelebA data.

In summary, our experiments show that for both MNIST and CelebA images, real samples are concentrated around the same manifold while the adversarial samples are pushed further away from this manifold. These results may be seen as a justification of our method and as a supporting argument for the assumptions (a)–(c) made on the data, see Methods III.

V. Related Work

Investigation of adversarial perturbations is one of the most active areas of deep learning research. A fundamental understanding of both the adversarial attacks and defenses is crucial and has already resulted in a large number of publications.

Here we list just a few, most relevant results on the detection of adversarial examples. In [11] and [12] the authors show that adversarial examples are not drawn from the same distribution as the original data, and can thus be detected using statistical tests. The dimensional properties of adversarial regions are characterized in [13] using so-called Local Intrinsic Dimensionality. In [14] the authors propose to use the knowledge extracted from a deep neural neetwork to improve its resilience to adversarial samples.

Unfortunately, the defense strategies are typically not successful if an attacker is allowed to modify the attack algorithm using the information about the defense mechanism of a particular network, see e.g., [15] and [16]. This results in a large number of open problems and a need for further research.

VI. Conclusion

In this paper, we present a method of detecting adversarial examples. To be exact, we compare the distance from a given example to a centroid of its nearest neighbors with the distribution of the distances from these neighbors to the centroids of their neighbors. For an adversarial example, such a distance is much larger than the corresponding distances for its neighbors while for a real example it is not.

The presented method is a step towards a better understanding of adversarial examples using the geometrical properties of data. We hope that such geometrical methods can make a significant contribution to our ability of detecting adversarial examples. Moreover, we expect the methods based on geometrical properties to be effective in the cases where the architecture-based-methods fail. Thus there is a good potential that combining both these types of methods will create a powerful tool for detection of adversarial examples.

Another argument in favor of the methods based on the geometry of the data is that adversarial examples seem to be generalizable across different models, see e.g., [4] and [17]. Therefore methods based on the geometry of the data may be more universal but also more dependent on the available training data.

In the manuscript we consider common Euclidian linear vector space for explanations, whereas the approach may be modified for kernel methods as well.

As a part of our future research we plan to test our method on the sets of adversarial examples that are generated in different ways [18] as well as on natural adversarial examples [19].

Acknowledgment

The work on this paper was initiated when Andrii Dmytryshyn visited National University of Kyiv-Mohyla Academy. Andrii Dmytryshyn thanks Galyna Kriukova for the invitation and kind hospitality during his visit.

References

[1] I. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," *ICLR*, vol. abs/1412.6572, 2015.

[2] A. Nguyen, J. Yosinski, and J. Clune, "Deep neural networks are easily fooled: High confidence predictions for unrecognizable images," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2015.

[3] N. Papernot, P. McDaniel, S. Jha, M. Fredrikson, Z. B. Celik, and A. Swami, "The limitations of deep learning in adversarial settings," in *2016 IEEE European Symposium on Security and Privacy (EuroS P)*, March 2016, pp. 372–387.

[4] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. J. Goodfellow, and R. Fergus, "Intriguing properties of neural networks," *CoRR*, vol. abs/1312.6199, 2013. [Online]. Available: http://arxiv.org/abs/1312.6199

[5] Y. Vorobeychik and M. Kantarcioglu, "Adversarial machine learning," *Synthesis Lectures on Artificial Intelligence and Machine Learning*, vol. 12, no. 3, pp. 1–169, 2018. [Online]. Available: https://doi.org/10.2200/S00861ED1V01Y201806AIM039

[6] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2*, ser. NIPS'14. Cambridge, MA, USA: MIT Press, 2014, p. 2672–2680.

[7] A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," *arXiv:1511.06434*, 2016.

[8] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016.

[9] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[10] Z. Liu, P. Luo, X. Wang, and X. Tang, "Deep learning face attributes in the wild," in *Proceedings of International Conference on Computer Vision (ICCV)*, December 2015.

[11] K. Grosse, P. Manoharan, N. Papernot, M. Backes, and P. McDaniel, "On the (statistical) detection of adversarial examples," *arXiv preprint arXiv:1702.06280*, 2017.

[12] X. Li and F. Li, "Adversarial examples detection in deep networks with convolutional filter statistics," *2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 5775–5783, 2017.

[13] X. Ma, B. Li, Y. Wang, S. M. Erfani, S. Wijewickrema, G. Schoenebeck, D. Song, M. E. Houle, and J. Bailey, "Characterizing adversarial subspaces using local intrinsic dimensionality," *Proceedings of the 6th International Conference on Learning Representations (ICLR), Vancouver, BC, Canada*, 2018.

[14] N. Papernot, P. McDaniel, X. Wu, S. Jha, and A. Swami, "Distillation as a defense to adversarial perturbations against deep neural networks," *Proceedings of the 37th IEEE Symposium on Security and Privacy, San Jose, CA.*, 2016.

[15] A. Athalye, N. Carlini, and D. Wagner, "Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples," *arXiv preprint arXiv:1802.00420*, 2018.

[16] N. Carlini and D. Wagner, "Adversarial examples are not easily detected: Bypassing ten detection methods," in *Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security*. ACM, 2017, pp. 3–14.

[17] N. Papernot, P. McDaniel, and I. Goodfellow, "Transferability in machine learning: from phenomena to black-box attacks using adversarial samples," *arXiv preprint arXiv:1605.07277*, 2016.

[18] N. Papernot, F. Faghri, N. Carlini, I. Goodfellow, R. Feinman, A. Kurakin, C. Xie, Y. Sharma, T. Brown, A. Roy, A. Matyasko, V. Behzadan, K. Hambardzumyan, Z. Zhang, Y.-L. Juang, Z. Li, R. Sheatsley, A. Garg, J. Uesato, W. Gierke, Y. Dong, D. Berthelot, P. Hendricks, J. Rauber, and R. Long, "Technical report on the cleverhans v2.1.0 adversarial examples library," *arXiv preprint arXiv:1610.00768*, 2018.

[19] D. Hendrycks, K. Zhao, S. Basart, J. Steinhardt, and D. Song, "Natural adversarial examples," *arXiv preprint arXiv:1907.07174*, 2019.

IEEE Third International Conference on Data Stream Mining & Processing
August 21-25, 2020, Lviv, Ukraine

Acquiring Custom OCR System with Minimal Manual Annotation

Jan Hula*, David Mojžíšek, David Adamczyk, Radek Čech
Institute for Research and Applications of Fuzzy Modeling
Ostrava, Czechia
*jan.hula@osu.cz

***Abstract*—We describe a development of a custom OCR system, which is designed specifically for a linguistic analysis of texts printed during the early modern period. This analysis requires precise detection of individual graphemes, and we, therefore, could not apply standard approaches that transcribe whole lines in an end-to-end fashion. We also describe our use of synthetically generated images, which allow us to avoid manual annotation of a large training set.**

***Index Terms*—OCR, Synthetic Data, Historical Texts, Neural Networks**

I. Overview

Optical Character Recognition (OCR) is nowadays considered a solved problem with very little space for innovation [1]. Still, some applications have very specific requirements for which custom solutions are needed. Here, we focus on a problem arising in a linguistic analysis of printed texts from the early modern period (1500-1750). The solution to this problem must include a precise detection of graphemes because the analysis will deal with relative sizes of printed graphemes and spaces between them as described in the next section. Because we have not found an open-source system which would fulfill our requirements, we decided to design the system by ourselves. It follows a two-stage pipeline. The first stage is based on a state-of-the-art model for object detection to detect individual graphemes and in the second stage, individual graphemes are recognized using bidirectional LSTM [2] which recognizes every grapheme in a context in which it appears. We solve these two stages separately. After training the grapheme detection model, we use it to detect thousands of instances of generic graphemes which we cut-out and cluster using a simple K-means algorithm. We then label these clusters and use them to pre-label training examples for the recognition model.

As annotating bounding boxes for individual graphemes can be a tedious task, we wanted to avoid as much of manual annotation as possible. For this purpose, we synthesize artificial examples of printed pages which help us to bootstrap a labeling process with a model pre-trained on these examples. Our methodology follows a simple principle where we train a weaker method requiring few training examples to pre-label training data for a stronger method. In this contribution, we describe the whole pipeline of our solution. We believe that it contains ideas which can generalize to similar problems.

II. Motivation and Problem Statement

The approach presented in this paper is motivated by needs that have emerged in linguistics focused on the development of the Czech orthographic system. Specifically, Voit [3] introduced new explanation of the usage of orthographic variations (such as ”uo” $\sim$ ”ů” / ”ú”, ”ie” $\sim$ ”ij” $\sim$ ”j”) in 16th century. He claims that the usage of either singlegraphic or digraphic grapheme was caused by pragmatic factors related to typesetting praxis. According to him, a typesetter was forced to fulfill the following requirements: 1) to align the right edge of the text, 2) to avoid splitting of words at the end of the line. In other words, an option to use either longer or shorter realizations of the grapheme was a tool for dilation or compression of text in the line.

This explanation offers setting up several empirically testable hypotheses. For instance, ”the higher the number of types in the line, the higher probability of occurrence of digraphic grapheme” or ”the higher the number of types in the line, the lower the number of spaces in the line”. Testing of hypotheses of this kind must be performed on a large sample of original texts. Further, proper testing needs careful operalization, which is not a trivial task in this case. For instance, a width of both the grapheme and space must be determined unambiguously. To our knowledge, up to now, only in [4], the problem had been analyzed empirically. However, they used the sample consisting of only four texts which were transcribed and annotated manually. The lack of adequately processed documents is the main obstacle for a thorough analysis of this phenomenon, and this could be solved by automatization of the annotation process.

In other words, we want to prepare an annotated dataset (together with the annotation tool) for further linguistic analysis. This analysis should confirm or disprove the hypothesis that changes in written Czech language were driven by technological limitations and needs in typesetting practice.

III. Description of the data

In this section, we briefly describe our data and their specifics. We are working with scanned printed documents mainly from the second half of the 17th century. They were printed in different Czech cities (Praha, Litomyšl, Olomouc, etc.). Not all of the documents are clearly readable - few pages are torn, or the text is faded out. Also, the typeset is very specific and different from contemporary documents. Even Czech

978-1-7281-3214-3/20/$31.00 ©2020 European Union

native speakers are not able to fluently read such texts if they are not trained. In figures 1,2,3 we present a few examples that point out some problematic areas in our dataset. Our dataset contains hundreds of scanned documents without annotations. We also dispose of dozens of transcriptions containing similar language, which we leverage when training the classification model, as described in section VIII.

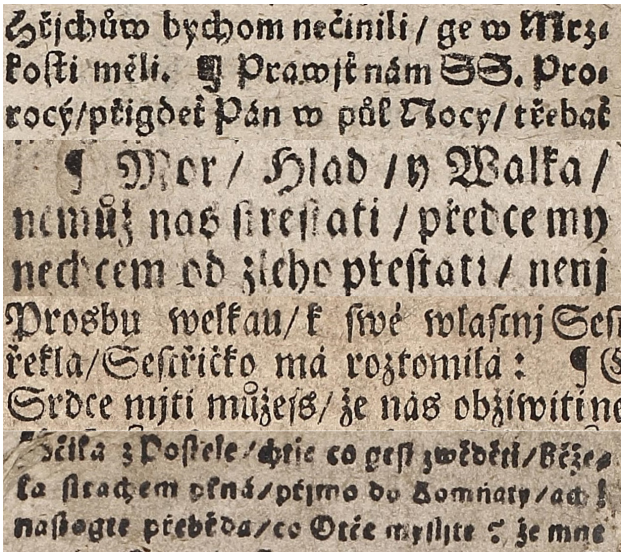


Fig. 1: Altough these text are from approximately the same period, they contain different typesets.

Fig. 2: Character difficulties. a) Ligatures, i.e. two or more characters printed as one. b) Some glyphs are hard to read and distinguish when the context is unknown.

IV. Outline of Our Pipeline

As most of OCR systems in use [5]–[8], we developed a pipelined system in which we solve individual steps independently of others. On a high level, we rotate each document so that lines are horizontally oriented, then we detect all graphemes in a given document, and finally, we classify each grapheme in the context of the text it appears in. The schematic representation of this process is depicted in figure 4. Most of modern OCR systems do not work by detecting individual graphemes but transcribe whole lines in an end-to-end fashion. This approach has the benefit that it does not require annotated

Fig. 3: Difficult pages. a) Some documents are damaged (for example folded or torn) or the qaulity of scan is low. b) Page containing an image and faded graphemes of different size.

bounding boxes. In our case, obtaining these bounding boxes is necessary, so we decided to split the pipeline to grapheme detection stage and grapheme classification stage. We could have also tried to classify and detect the graphemes in parallel, as it is usually done in object detection. By doing so, we would miss the opportunity to incorporate the structure of the text into the classification. The detection network would classify the grapheme as a patch in the image, instead of a grapheme in the sequence of graphemes. The second minor benefit of separating these two stages is the fact that we do not need to annotate the class of every bounding box, as will be described in section VIII. To obtain a final version of our system, we rely heavily on synthetic data and other tricks which allow us to avoid as much manual annotation as possible. Here follows a description of all the steps we execute during the construction of our system:

Alignment part

1) Train a neural network *NN-rot* (VGG-11) to regress angles of rotated images.
2) Use *NN-rot* to straighten all images in the dataset.

Detection part

3) Annotate bounding boxes around graphemes in 3 random pages.
4) Use the annotated graphemes to generate a large training set of synthetic images.
5) Pre-train a neural network *NN-det* to detect graphemes in synthetic images.
6) Fine-tune *NN-det* on the 3 annotated real images.
7) Pre-label 6 new pages with *NN-det* and correct wrong detections to enlarge the annotated set.
8) Repeat steps 4-7 two more times to increase the variability of the dataset and to obtain an accurate detection model.

Classification part

9) Detect few thousands of graphemes in still unlabeled images using *NN-det* trained in previous steps.
10) Cut out the detected graphemes and train an autoencoder to obtain low-dimensional representations for every grapheme.

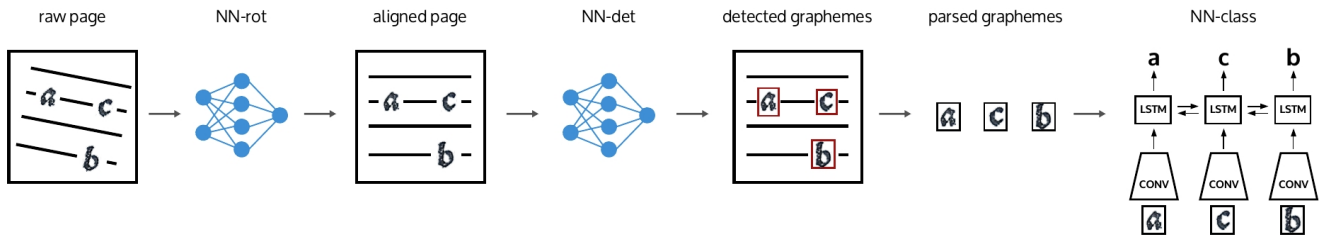


Fig. 4: A schematic representation of the pipeline used for every new image. The system first alignes the page according to a predicted angle from *NN-rot*; next it detects all graphems on the page with *NN-det*; these graphemes are parsed into a sequence which is finally classified by *NN-class*

11) Cluster the graphemes using K-means based on the representation from the autoencoder.
12) Create a labeled dataset of graphemes by manually labeling the clusters and deleting incorrectly assigned graphemes.
13) Use available transcriptions of texts from the same period containing the same language to generate training sequences of cut-out graphemes.
14) Train a convolutional bi-LSTM *NN-class* to classify each grapheme in the context of other graphemes using training examples from step 13.

By following these steps, we avoid manual labeling of hundreds of pages. Their details will be described in the subsequent section. At the end, we use *NN-rot*, *NN-det*, and *NN-class* for every new image.

V. Page Alignment

The scanned documents in our dataset were not always aligned, and therefore, individual lines were not aligned horizontally. Page alignment is a part of all OCR systems because when the letters are aligned, the subsequent recognition model does not need to learn rotation invariance. We decided to train a neural network to predict the angles from cropped patches of the image. Fortunately, obtaining a labeled dataset for this task does not require a lot of manual effort. We manually aligned 30 pages, and these aligned images are then rotated by a random angle from an interval -15 to +15 degrees, quantized to increments of 0.5. Subsequently, crops are taken from these rotated images, and the rotation angle is saved as their target label. We train a convolutional network with ResNet-18 backbone [9] to predict these angles, and for every new image, we average predictions from crops taken from it. We achieve nearly perfect accuracy ($\sim$98%) with this approach.

VI. Synthetic Data and Domain Knowledge

One of the biggest drawbacks of data-driven approaches for problem-solving is that they often need a lot of labeled examples to be trained on. To avoid this problem and save a lot of manual annotation, we decided to use synthetic data, which we generate using our understanding of the problem domain. Synthetic data has been recently used in all kinds of domains of Computer Vision and Machine Learning in general [10]–[13], and their use in OCR for modern print marks one of their first successful use-case [14]. The main pitfall of using synthetic datasets to train systems for real data is that the training distribution may be very different from the testing distribution because, in some domains, it may not be trivial to synthesize realistic examples (e.g., synthesizing realistic images of human faces). OCR for modern print was a successful use-case mainly because it is possible to create highly realistic synthetic examples using known fonts and simple image distortions and noise. Creating realistic examples of old prints is more involved because the variability of the appearance of graphemes is larger due to all kinds of problems arising during the printing process (e.g., leakage of ink) and degradation of documents after long periods. Examples of such problems can be seen in figures 2 and 3.

After a visual inspection of many real documents, we model the realism and variability of the page as closely as possible. For this, we use cut-out examples of various graphemes [1] with background removed and documents containing blank pages. We generate each page by sampling random words from which we create whole lines and place them to an empty page. On each grapheme, we apply a random set of augmentations simulating fading of the ink, elastic distortion, and various kinds of noises and scratches. To gain the variability in the background, we also apply similar augmentations on the few blank pages we had at our disposal. Also, many documents contained random drops of ink, which could be possibly mistaken for a grapheme, and therefore we add such drops to the background at random positions. An example of a such generated page can be seen in figure 5. For our experiments, we created 300 synthetic pages together with ground truth bounding boxes for every grapheme.

VII. Glyph Detection

As described in section IV, we first detect all graphemes as one generic class, then we parse the segmented graphemes to a sequence of these graphemes, and finally, we classify each grapheme using a sequence-based model. For the detection of generic graphemes, we use a state-of-the-art model for object

[1] Synthetic data described in this section are used only to detect generic graphemes, i.e., to detect a grapheme without classifying it into a class.

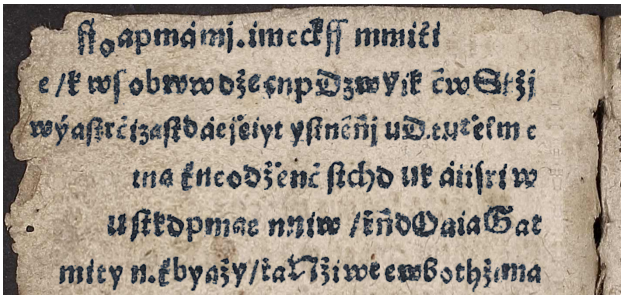


Fig. 5: An example of a crop from artificially generated page. We used different grapheme and background augmentations to add variability.

detection called RetinaNet [15], which we slightly modify to suit our task. Object detection models are usually categorized into one-stage and two-stage methods. As the names suggest, one-stage methods classify and detect objects in one stage, whereas two-stage methods first propose a candidate bounding boxes, which are then refined and classified in the second stage. One-stage methods are usually faster but less precise due to the imbalance of positive and negative bounding box proposals. RetinaNet solves this short-coming of one-stage methods by using special loss function called Focal Loss [15], which takes this imbalance into account. Therefore RetinaNet is a fast and accurate architecture for object detection.

Most of the time, models for object detection use pre-trained backbones (feature extractors) trained on big classification datasets such as ImageNet [16]. These backbones extract already useful features that can be leveraged by subsequent layers in the network. The pre-trained backbone is most useful when the pre-training domain is similar to the target domain. In our case, the images of scanned printed documents are very different from photographs capturing random objects, and therefore we do not use a pre-trained backbone but train it from scratch on our synthetic dataset.

As most of the one-stage detectors, RetinaNet uses anchors when detecting individual objects. During detection, the image is divided into a grid of rectangular cells, and inside each cell, multiple anchors of predefined sizes and proportions are used to detect possible objects. The final bounding box is being regressed from each such anchor and classified as a particular class or as a background. Setting up the sizes and ratios correctly is an important step. It, for example, does not make sense to have anchors large in size if we know that there won't be large objects within any image. We, therefore, take special care to set up these sizes and ratios so that they cover the sizes of graphemes in our dataset. For a more complete overview of current methods in object detection see [17].

Also, we wanted to retain the resolution of images as high as possible, and so we cut the whole page into overlapping patches of size 256x256 px and process each patch separately. After the system processes all patches from an image, we merge all boundary boxes in a post-processing stage. We train the detection model on 300 synthetically generated pages, and

TABLE I: Comparison of average precision (AP) and Focal loss between the same model (RetinaNet) trained with and without syntetic images.

Training Data	AP	Focal Loss
With synthetic data	**0.3110**	**1.143**
Without synthetic data	0.03767	2.181

then we fine-tune it on three real and manually labeled pages. This model already produces quite precise bounding boxes, so we use it to pre-label six more pages in which we manually correct wrong predictions. We then enlarge the training set of synthetic and real images using these newly labeled pages. We repeat this process two times and thus acquire a model of satisfying accuracy. In table I, we compare average precision and Focal loss between the same model trained with and without synthetic images created from three annotated pages. As these metrics are not easily interpretable, we show a visual example in figure 6.

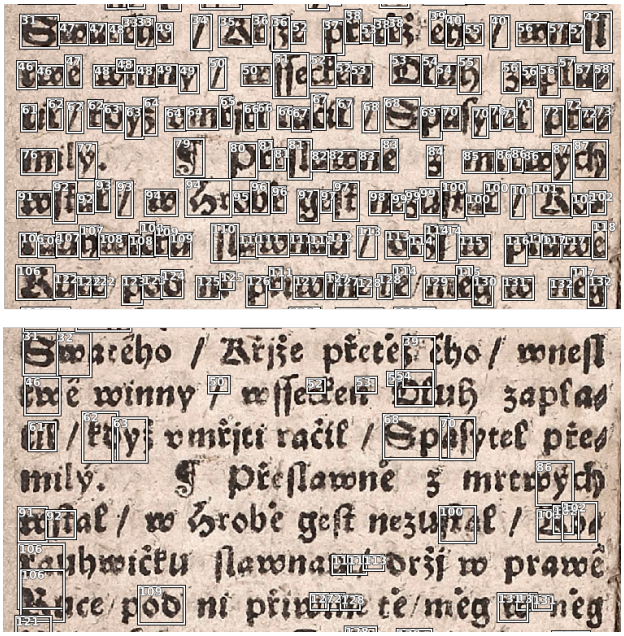


Fig. 6: A visual example of a quality of detection after the first round using only 3 annotated images. The top image shows results from a model trained on synthetic images and fine-tuned on the 3 annotated images. The bottom image shows results from a model trained only on the 3 annotated images (with standard augmentations, i.e. resize, lightness, etc.).

VIII. Glyph Classification

In the last section, we described the model for grapheme detection. We use this model to detect generic graphemes, which will be parsed into a sequence where the order of graphemes is the same as the order in which we would read

the text. These sequences are then classified with a sequence-based model, which we describe in this section. We use a sequence-based model because, in some instances, the identity of a grapheme may not be recognizable without a context. The sequence-based model can leverage statistical regularities within sequences of letters found within the use of a language.

First, we need to create a dataset for classification. Again, we want to avoid manual annotation as much as possible. We came up with two ways how to achieve it. Our classification model is convolutional bidirectional LSTM which takes a sequence of cut-out graphemes and produces a sequence of labels, one label for each grapheme. Therefore we need to obtain labeled training examples of such sequences. Our idea was to use transcriptions of texts which contain a language being used in our documents. Given these transcripts, we can create many different training sequences by sampling graphemes according to the letters in the text. This has an advantage that using one sentence, we can generate many different training examples by sampling different examples for the same letter each time. In order to do this, we need to have many examples of each letter. To avoid manual annotation of separate graphemes, we use the detection model to cut-out thousands of generic graphemes, which we then cluster and label only the clusters. This lowers the amount of work we need to do by order of magnitude.

In order to cluster the graphemes, we first train an autoencoder with ResNet-18 in the encoder to obtain a low-dimensional representation of every grapheme. Umap [18] visualization of this low-dimensional space can be seen in 7. We then cluster all graphemes using a simple K-means algorithm. We set K to 2*C where C is the number of classes because when K is close to C, the algorithm mixes too many examples of different classes together in the same cluster. We manually check all clusters and remove incorrectly assigned examples. Thus, we acquire thousands of labeled examples with very little manual effort. From these, we construct a dataset of labeled sequences.

Our classification model first extracts low-dimensional (512) representation of each grapheme by processing it with a backbone from ResNet-18, and the sequence of these low-dimensional representations then goes as an input into bidirectional LSTM (with 2 layers, both of which have the output dimension equal to 512). Finally, the sequence of representations in the hidden layer of the LSTM is then processed by a linear layer with a softmax to predict the class of the grapheme at every position of the sequence.

In table II, we show a comparison between our model, which takes the context of each grapheme into account and a model with the same backbone that classifies each grapheme separately.

IX. Related Work

The OCR problem has been studied for many years [1]. Especially for modern prints, there exist software solutions with nearly perfect accuracy. Most of these systems use a pipelined approach such that in one step, individual lines are segmented out, and subsequently, they are transcribed into sequences of letters by a neural network that processes the whole line as a sequence of vertical strips of pixels. The number of such strips in a line does not correspond to the number of letters (labels) in that line, and therefore special loss function called CTC loss [19] is used to account for the alignment of these strips and labels. The same loss function is frequently being used in speech recognition, where the same problem arises. Examples of systems based on this approach include Tesseract [7], Calamari [6] build on top of TensorFlow and OCRopus [5] build on top of PyTorch. We started our development with OCRopus but soon realized that we need a custom solution. We have also tried a system called OCR4ALL [8], which was developed specifically for historical OCR. It is targeted mostly on people with no coding skills, so the emphasis is being given mainly on user-friendliness and not much on customizability. In most of the use-cases, it is not needed to segment out individual graphemes, and so there is no need to innovate over these solutions. As mentioned in the motivation, our use case was quite specific, and therefore we needed to develop our own solution.

The usefulness of synthetic datasets for training machine learning models was realized by many [10]–[12]. Synthetic

Fig. 7: Umap visualization of low-dimensional representation of cut-out graphemes obtained from a trained autoencoder.

TABLE II: Comparison between a model which classifies each grapheme separately and a model which takes the ohter graphemes in the same context into account. Both models share the same backbone (ResNet-18).

Method	Validation Accuracy
With bi-LSTM	**0.9337**
Without bi-LSTM	0.498

datasets are especially useful in robotics [13] and other domains where training on real data would be too expensive or infeasible. The main problem arising with the use of such datasets is that a machine learning model may overfit to specifics of synthetic examples and may not transfer well to real examples. As a possible solution to this problem, a technique called Domain Randomization became popular in recent years [20]–[22]. The idea behind Domain Randomization is that if we do not want to overfit to particular properties of data, such as for example color of objects when training object classifier, we should randomize that property as much as possible so that the model can not learn a statistical correlation between this property and some other variables of interest. We took inspiration in this idea when we were creating our synthetic dataset and randomized some properties of the generated images such as the size of graphemes, their sharpness, and other distortions.

X. Conclusion

In this contribution, we described creation of a custom OCR system explicitly designed to help linguistic analysis of printed texts from the early modern period. In contrast to mainstream OCR systems, we do not transcribe whole lines in an end-to-end fashion, but we first segment out individual graphemes, which are then classified using a sequence-based model. We also showed the usefulness of synthetically generated images and a bootstrapping process for annotation, which reduced the amount of manual work we needed to do by order of magnitude. In the future, we aim to design an intuitive user interface for our system which will be released together with the final version of the source code. We believe that our work, driven by the practical needs of linguists, is a valuable contribution to the interdisciplinary research between computer science and humanities.

Acknowledgment

The work was supported from ERDF/ESF "Centre for the development of Artificial Intelligence Methods for the Automotive Industry of the region" (No. CZ.02.1.01/0.0/0.0/17_049/0008414).

References

[1] D. Doermann, K. Tombre *et al.*, *Handbook of document image processing and recognition*. Springer, 2014.

[2] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[3] P. Voitem and J. Linka, "Udržet pravỳ okraj stránkové sazby (od literární historie k samostudiu)," *Česká literatura*, vol. 59, no. 2, pp. 242–260, 2011.

[4] R. Čech and J. Mačutek, "Orthography system of broadside ballads from 17. and 18. century. quantitative approach." Transformations of Czech 'kramářské písně' (Broadside Ballads) – media, traditions, contexts. Masaryk University in Brno, 2019.

[5] T. M. Breuel, "The ocropus open source ocr system," in *Document Recognition and Retrieval XV*, vol. 6815. International Society for Optics and Photonics, 2008, p. 68150F.

[6] C. Wick, C. Reul, and F. Puppe, "Calamari-a high-performance tensorflow-based deep learning package for optical character recognition," *arXiv preprint arXiv:1807.02004*, 2018.

[7] R. Smith, "An overview of the tesseract ocr engine," in *Ninth International Conference on Document Analysis and Recognition (ICDAR 2007)*, vol. 2. IEEE, 2007, pp. 629–633.

[8] C. Reul, D. Christ, A. Hartelt, N. Balbach, M. Wehner, U. Springmann, C. Wick, C. Grundig, A. Büttner, and F. Puppe, "Ocr4all—an open-source tool providing a (semi-) automatic ocr workflow for historical printings," *Applied Sciences*, vol. 9, no. 22, p. 4853, 2019.

[9] S. Wu, S. Zhong, and Y. Liu, "Deep residual learning for image steganalysis," *Multimedia tools and applications*, vol. 77, no. 9, pp. 10 437–10 453, 2018.

[10] J. Hula, I. Perfilieva, and A. A. M. Muzaheed, "Towards visual training set generation framework," in *International Work-Conference on Artificial Neural Networks*. Springer, 2017, pp. 747–758.

[11] S. Sankaranarayanan, Y. Balaji, A. Jain, S. Nam Lim, and R. Chellappa, "Learning from synthetic data: Addressing domain shift for semantic segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 3752–3761.

[12] S. R. Richter, V. Vineet, S. Roth, and V. Koltun, "Playing for data: Ground truth from computer games," in *European conference on computer vision*. Springer, 2016, pp. 102–118.

[13] P. Martinez-Gonzalez, S. Oprea, A. Garcia-Garcia, A. Jover-Alvarez, S. Orts-Escolano, and J. Garcia-Rodriguez, "Unrealrox: an extremely photorealistic virtual reality environment for robotics simulations and synthetic data generation," *Virtual Reality*, pp. 1–18, 2019.

[14] T. K. Ho and H. S. Baird, "Evaluation of ocr accuracy using synthetic data," in *Proceedings of the 4th Annual Symposium on Document Analysis and Information Retrieval*. Citeseer, 1995.

[15] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2980–2988.

[16] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 2009, pp. 248–255.

[17] L. Jiao, F. Zhang, F. Liu, S. Yang, L. Li, Z. Feng, and R. Qu, "A survey of deep learning-based object detection," *IEEE Access*, vol. 7, pp. 128 837–128 868, 2019.

[18] L. McInnes, J. Healy, and J. Melville, "Umap: Uniform manifold approximation and projection for dimension reduction," *arXiv preprint arXiv:1802.03426*, 2018.

[19] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber, "Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks," in *Proceedings of the 23rd international conference on Machine learning*, 2006, pp. 369–376.

[20] J. Tobin, R. Fong, A. Ray, J. Schneider, W. Zaremba, and P. Abbeel, "Domain randomization for transferring deep neural networks from simulation to the real world," in *2017 IEEE/RSJ international conference on intelligent robots and systems (IROS)*. IEEE, 2017, pp. 23–30.

[21] X. B. Peng, M. Andrychowicz, W. Zaremba, and P. Abbeel, "Sim-to-real transfer of robotic control with dynamics randomization," in *2018 IEEE international conference on robotics and automation (ICRA)*. IEEE, 2018, pp. 1–8.

[22] K. Bousmalis, A. Irpan, P. Wohlhart, Y. Bai, M. Kelcey, M. Kalakrishnan, L. Downs, J. Ibarz, P. Pastor, K. Konolige *et al.*, "Using simulation and domain adaptation to improve efficiency of deep robotic grasping," in *2018 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2018, pp. 4243–4250.

IEEE Third International Conference on Data Stream Mining & Processing
August 21-25, 2020, Lviv, Ukraine

Unsupervised Object-aware Learning from Videos

Jan Hula
Institute for Research and Applications of Fuzzy Modeling
Ostrava, Czechia
jan.hula@osu.cz

***Abstract*—We consider a novel unsupervised learning setup in which training examples are grouped into small bundles that preserve an identity of an object. Such setup may practically arise when we are able to detect moving objects in videos without being able to classify their identity. Our approach is based on a construction of a similarity graph of bundles from which we are able to recover the identities of objects by applying a community detection algorithm. Finally, we train Siamese Neural Network to discriminate examples from different components and show that thus acquired representations produce well-separated clusters. Part of our contribution is also a unique dataset we assembled in order to test the presented idea.**

***Index Terms*—Unsupervised Learning, Clustering, Community Detection, Computer Vision**

I. Introduction

We present a novel unsupervised learning setup in which certain examples are grouped into bundles respecting the class of these examples. This idea was inspired by how toddlers learn by exploring unknown objects[1]. Our approach developed for this setup is called Unsupervised Object-aware Learning and starts by extracting parts of videos belonging to an object which is being tracked. We then compare these extracted parts to each other and construct a similarity graph in which we discover components using a standard community detection algorithm. Finally, we train Siamese Neural Network [1] to discriminate between these components. We empirically show that the discovered components correspond to the classes present within a dataset and that the representations obtained from Siamese Neural Network create well-separated clusters within a Umap visualization [2]. We test our approach on two datasets. One easier, created by modifying the MNIST dataset and second, more realistic, which we create specifically to test our approach.

Here are our main contributions:

- We introduce a novel unsupervised learning setup that exploits the temporal coherency of videos.
- For testing our approach, we assemble a dataset which we release online.
- We show a way to filter out correlated frames extracted from the same video.
- We provide experiments showing the promising direction of our approach.

The rest of the paper is structured as follows: Section 2 describes the motivation for this work; section 3 contains a detailed description of individual steps in our approach; in section 4 we describe a dataset we created in order to test our approach; experiments are presented in section 5; section 6 contains related work and sections 7 and 8 contain discussion and conclusion.

[1]https://www.youtube.com/watch?v=8vNxjwt2AqY

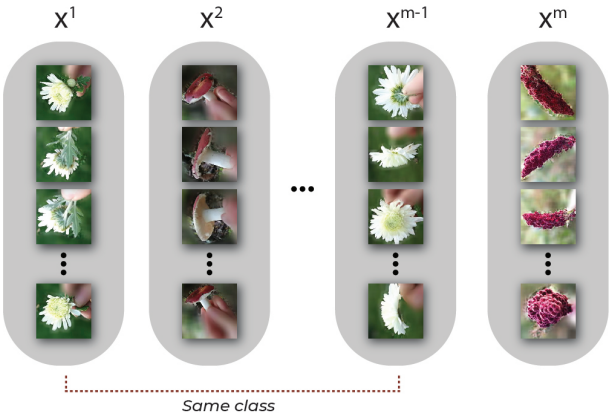


Fig. 1. A schematic structure of a dataset we assume in our approach. Images are divided into image bundles where in each bundle there are only images of the same class. Many bundles may correspond to the same class as depicted by the dotted arrow.

II. Motivation and Problem Statement

Discovery of unsupervised learning algorithms that would allow us to avoid the large effort of manual labeling of training examples is a highly desirable goal. Nonetheless, it is not known whether a completely unsupervised approach which would work with i.i.d. datasets is practically realizable. We may hope for its existence on the basis of evidence from nature where unsupervised learning is believed to be abundant [3]. One crucial difference between machine learning algorithms and animals is that the later do not learn from i.i.d. datasets but from temporally coherent streams of sensory-motor information. It is also known that many animals dispose of innate inductive biases which allow them to learn much more efficiently [4]. One of the innate abilities of most animals is the ability to track moving objects [5]. We may speculate that this ability allows them to focus their learning on relevant parts of the sensory stream. It may also enable them to learn invariant properties of objects because their innate learning algorithm may exploit the fact that objects will not change their identity when they move or are viewed from a different angle. This observation serves as the main inspiration for our work.

A. Problem Statement

We assume a dataset of images X which is organized into small bundles X^i where i is the number indexing the bundle and not the class. Each bundle is restricted to contain only images of the same object which may be viewed from different angles or in different poses. We assume that images in a bundle will be created by tracking an object within a video stream. There may be multiple bundles corresponding to the same class of object (e.g. observations of different instances of the same type of flower) but we assume that this information is not disclosed to us. Our aim is to provide an unsupervised learning algorithm that will recover the classes contained within the dataset and produce representations which cluster according to these classes. A schematic illustration of the dataset is depicted in Figure 1.

III. Unsupervised Object-aware Learning

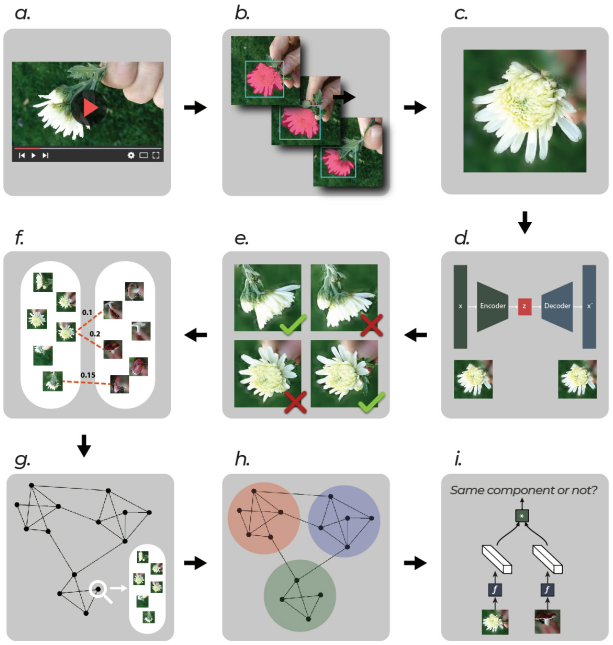


Fig. 2. Schematic depiction of the whole pipeline. Individual figures correspond to steps described in this section.

In this section, we describe the process of extracting representations of images starting with videos capturing a set of objects. This means that aside from unsupervised learning on a dataset with image bundles, we also describe how we prepare such dataset from raw videos. The whole process is depicted in Figure 2 and works as follows:

1) We start with a set of n videos capturing m different objects (one object per video, $n \gg m$).
2) In every video, we track the main object together with an approximate mask of the object.
3) In each frame, we blur out the masked background and extract a set of frames from each video by cutting out the bounding box of the object.
4) We train Variational Auto-encoder on all extracted frames to acquire low-dimensional representation for every image.
5) As we do not want too many correlated frames, frames extracted from a given video are filtered out according to the similarity of their low-dimensional representations. We call the resulting set images for one video an image bundle.
6) We compute the average similarity between every pair of tracked observations.
7) Using this similarity, we construct a similarity graph of image bundles by connecting every image bundle to its 5 most similar image bundles.
8) Community detection algorithm is applied on the resulting graph to find highly interconnected components in it.
9) Finally, we train a Siamese Neural Network to discriminate between components discovered by the community detection algorithm.

A. Tracking objects within videos

For our algorithm to be completely automatic and purely unsupervised it would need to be able to detect general objects within a video. Unfortunately, class-agnostic object recognition – recognizing that a certain set of pixels constitute an object – is not a well-studied task in Computer Vision and according to our knowledge, there are no reliable algorithms that we could use. Therefore we approach this problem with a visual object tracking algorithm which requires an initialization with a bounding box. Concretely, we use a model called SiamMask [6], a neural network which simultaneously performs both visual object tracking and object segmentation. Therefore, after this step, we obtain an approximate mask around the object in every frame and a bounding box we can use to cut out the object.

B. Background blurring

Our final representations should be ideally invariant with respect to a change of background. Also, when we measure the similarity of images in steps 5 and 6, we do not want this similarity to capture similarity of backgrounds. Standard practice in Computer Vision is to augment the dataset with transformations with respect to which we want to be invariant [7]. This would in our case mean to sample different backgrounds for the masked object from the previous step. Instead of doing this, we decided to blur out the background with the intention to remove any irrelevant signal. Because the mask produced from SiamMask is not tight around the object, we blur also the border of the mask to remove unintended edges between blurred and unblurred parts. Samples of blurred and cut-out images can be seen in Figure 3.

C. Obtaining low-dimensional representation

We want to measure the similarity of pairs of images in order to be able to filter out correlated frames from each video and also to construct the similarity graph of image bundles. For

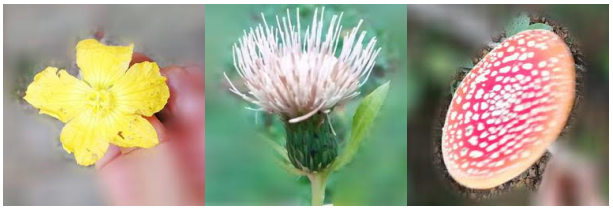


Fig. 3. Examples of cut-out images with blurred background. The approximate mask for the background is produced with SiamMask architecture.

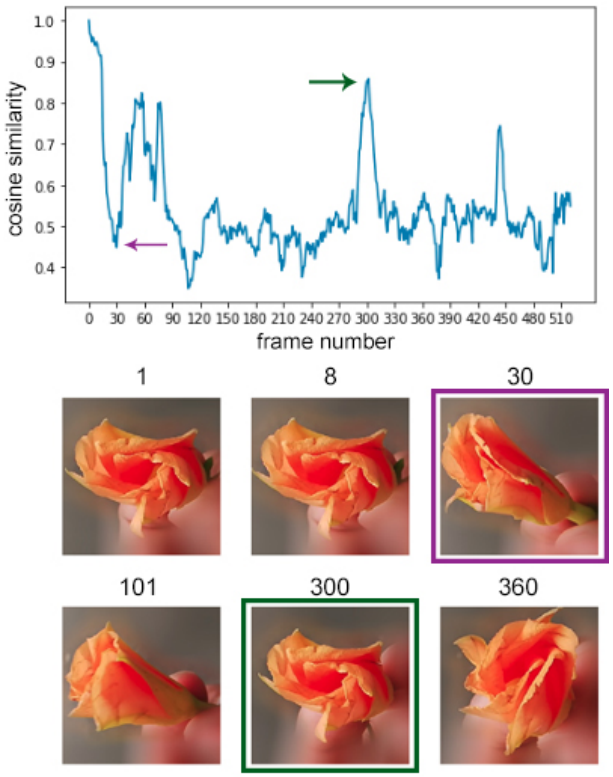


Fig. 4. ***Top***: a graph of cosine similarity between first and other frames in the video. ***Bottom***: Picked out frames from the video with their corresponding frame numbers. The two highlighted frames correspond to two arrows within a graph.

this purpose, we train a convolutional version of Variational Auto-encoder [8] and use the representation produced by the encoder. As we partially blur out the background, most of the capacity of the model can be used to capture the object.

D. Filtering out correlated frames

If we would simply use all the frames from each video, the distribution of the dataset may end up skewed because some parts of a video may be more static then others and these parts would produce many correlated frames. We, therefore, need to filter out correlated frames extracted from a given video. We first tested whether the subjective visual similarity of images can be captured by cosine similarity between their low-dimensional representations. As can be seen in Figure 4, it captures the visual similarity well enough for our purposes[2].

To extract n uncorrelated frames for every video, we run k-means clustering (with $k == n$) on the low-dimensional representations from Variational Auto-encoder and take the most similar frame to every centroid of the resulting clusters. We did not try anything more complicated as this simple heuristic already produced uncorrelated images[3]. The results from this phase constitute what we call image bundles.

E. Similarity graph and community detection

We would like to cluster all images corresponding to the same class together. We approach it by constructing a graph where nodes correspond to different image bundles and edges arise by connecting every image bundle to its 5 most similar image bundles. The similarity between two image bundles is computed by averaging the similarity of l most similar pairs of images (one image from the first bundle, second image from the second bundle). We do not average across all pairs as this average may be shifted by outliers. The intuition behind this step is that two image bundles corresponding to the same class may contain few very similar images (e.g. few frames may capture the object from the same angle and in the same position) and these will link the corresponding bundles together.

After constructing the similarity graph, we run Louvain method for community detection [9]. This is a standard community detection algorithm whose goal is to find components within a graph that are highly interconnected and only sparsely connected to the rest of the graph. In our experiments – described in section V – the discovered clusters corresponded to classes of objects within a dataset with only a few misclassified cases.

F. Training a Siamese Network to discriminate between components

The previous step can be seen as an assignment of labels to a training set. Finally, we train a Siamese Neural Network with a triplet loss [10] to discriminate images belonging to different components of the graph. The training works by sampling batches of image triples – where two images belong to the same component and one belongs to a different component – and optimizing the following loss function:

$$\sum_{i}^{N} \left[\|f(x_i^a) - f(x_i^p)\|_2^2 - \|f(x_i^a) - f(x_i^n)\|_2^2 + \alpha \right]_+ . \quad (1)$$

x_i^a is an anchor image for which we sample one positive example x_i^p (belonging to the same component) and one negative example x_i^n (belonging to a different component). f is a function which produces low-dimensional representations of images; in our case convolutional neural network. The hyperparameter α is a margin between a similarity between

[2]We also tried the SSIM algorithm but it did not capture the visual similarity as well as the cosine similarity between the representation from Variational Auto-encoder.

[3]Even simpler heuristic would be to take every n-th frame but our intuition was that k-means may produce more representative examples.

negative and positive pairs. The loss function forces this network to produce representations in which the distance for the positive pair is small and distance the negative pair is high. We show visualizations of resulting representation in section V.

IV. Description of the Dataset

In order to test our idea, we needed a dataset that inherently contains the above-mentioned notion of image bundles. As we could not find any such dataset, we first tested the idea by artificially splitting MNIST dataset into bundles of 50 images each containing only digits of the same class. After obtaining positive results – mentioned in the next section – we decided to assemble our own dataset where the notion of image bundles would appear naturally. The resulting dataset

Fig. 5. Samples from the Organic Objects dataset. Images are already cropped and have blurred out background

contains 18 classes of organic objects, some of which are depicted in Figure 5. We have chosen organic objects because they naturally produce large variability between instances. For every class, we shoot 10 different instances of that class (e.g. 10 different roses if a rose is the class) capturing the object from different viewpoints. These 10 instances correspond to 10 images bundles.

In every video, we mark the bounding box around the object of interest and track it together with the approximate mask using SiamMask architecture. As described in the previous section, we blur out the background and cut out the bounding box in every frame and finally filter out correlated frames. All images have a resolution of 256x256 pixels. When running the k-means clustering to find the representative frames, we set k to 20 so we end up with 20 images per one image bundle. The final dataset can be downloaded at the following address: https://github.com/Jan21/Organic-objects-dataset/

V. Experiments

We test our idea on a modified version of MNIST dataset and our own dataset capturing organic objects.

A. Experiments on modified MNIST

In order to create image bundles for MNIST, we split the training set into bundles of 50 images each containing only instances of the same class. This will produce around 120 image bundles per class which enables more robust community detection.

To extract low-dimensional representations, we use Variational Auto-encoder with 3 linear layers in the encoder (with dimensions $784 \to 500$, $500 \to 500$, $500 \to 20$). The body of Siamese Network architecture contains 2 convolutional layers (with feature-map counts 32 and 64) interleaved with max pooling layers and 2 fully-connected layers (with dimensions $1024 \to 256$, $256 \to 20$). In Figure 6 we show discovered components from community detection and in Figure 7 we show Umap visualization of representations obtained from Siamese Network.

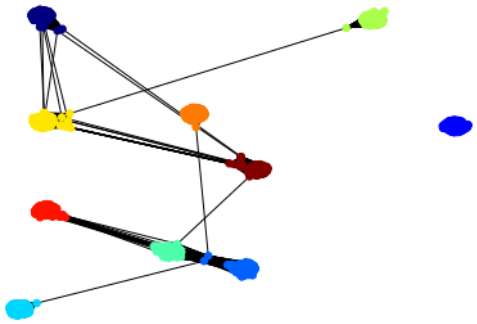


Fig. 6. Visualization of detected communities in the modified MNIST dataset using the Louvain method. Nodes are colored according to the component (community) they are assigned to. The method discovered 10 components that belong to 10 different digits even though some image bundles are incorrectly connected to image bundles from a different class. Also, nodes from the same component are spatially clustered because the visualization technique pulls connected nodes together.

Fig. 7. Umap visualization for the modified MNIST dataset. The visualized representations were obtained by training a Siamese Neural Network to discriminate between components of the similarity graph. Colors represent ground-truth classes.

B. Experiments on Organic Objects dataset

In comparison with the modified MNIST dataset, this dataset contains image bundles with fewer images (20) and there are also fewer image bundles per class (10). Therefore the construction of the similarity graph and the following

community detection on it may be noisier. To obtain the low-dimensional representations which will be used for filtering out correlated frames and construction of the similarity graph, we resize all images to resolution 64x64 pixels and train a convolutional version of Variational Auto-encoder with 5 convolutional layers (with feature-map counts: 16, 32, 64, 128, 256 and stride 2) and 1 fully-connected layer (with dimensions $1024 \rightarrow 96$).[4] After running community detection on the

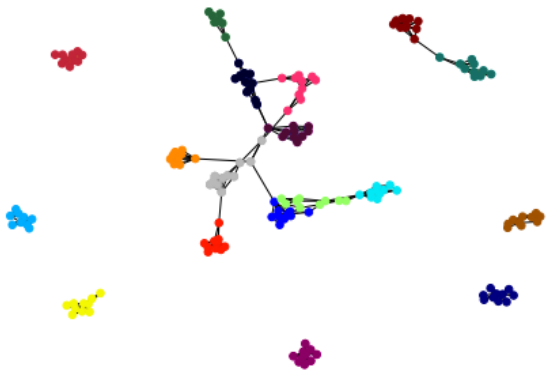


Fig. 8. Visualization of detected communities in the Organic Objects dataset with the Louvain method. Nodes are colored according to the component (community) they are assigned to. The method discovered 18 components which belong to 18 different classes.

similarity graph, we inspect how many image bundles were assigned to a wrong component. We found out that only 5 of 162 image bundles in the training set were assigned to a wrong component. In Figure 8, we show the result of community detection on the constructed similarity graph. In Figure 9 we show Umap visualization of representations obtained from Siamese Network colored by corresponding classes. These representations do not cluster that well as in the case of modified MNIST dataset which should be expected because digits are symbols designed to be well distinguishable. The source code for all experiments will be released together with this contribution and accessible from the same URL as the dataset.

VI. Related Work

Our research is marginally related to current self-supervised approaches using large quantities of unlabeled data such as [11], [12] and also the once which try to exploit some kind of coherency in the data [13], [14], [15]. Also, our approach can be seen as a version of clustering with constraings. This task was heavily studied in the past for example by [16], [17], [18].

VII. Discussion and Future Work

The main limitation of the presented approach is the fact that we have to initialize bounding boxes to track objects within

[4]We also tried to extract representations by using VGG16 which was pre-trained on ImageNet. These representations better discriminated very similar objects (e.g. two types of red flower).

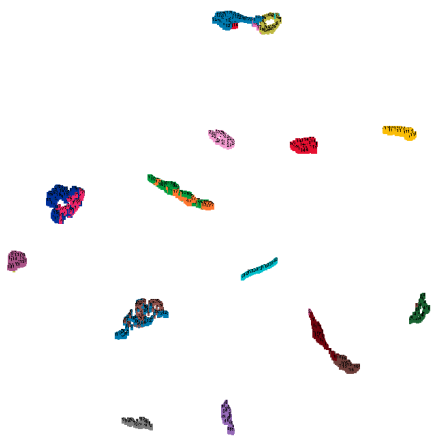


Fig. 9. Umap visualization for the Organic Objects dataset. The visualized representations were obtained by training a Siamese Neural Network to discriminate between components of the similarity graph. Colors represent ground-truth classes. In this case, the clusters are not that well-separated as in the case of the modified MNIST dataset.

videos. Doing this automatically will be the main focus of our future work. This functionality will allow us to automatically extract image bundles from a large number of videos [19].

Another limitation is the fact that the dataset we constructed may be too "curated". We assume that our approach would be used in scenarios where there is no or very limited curation of data like for example the automatic extraction of image bundles from a large number of videos. Such image bundles may be very noisy and disproportionally distributed (e.g. there may be a lot of bundles for some class and very few for others). It may be the case that in such scenarios the construction of the similarity graph and the subsequent community detection will not produce sensible results and so we will need to develop more sophisticated solutions which will consider noise and class imbalance. Also, the subsequent training of Siamese Network may be improved by different sampling strategies of positive and negative examples. Currently, the sampling works by uniformly picking positive examples from the same component as the anchor example and negative example from a different component. In future, we would like to modify the sampling so it takes the similarity of image bundles into account.

Lastly, we would like to point out that our method may be useful for the creation of datasets for supervised learning. By extracting image bundles with decorrelated frames from a large number of videos and finding out components of the resulting similarity graph, we may quickly assemble a lot of training examples that would require only one label per component (probably with quick visual inspection).

VIII. Conclusion

We presented a novel setup for unsupervised machine learning where certain examples are grouped into bundles respecting the class of these examples. These bundles are semi-automatically created by exploiting the fact that frames

extracted from videos are temporally coherent and that objects do not change their class from moment to moment. We also point out the importance of class-agnostic object detection (detecting an object without classifying it) which is a task that is rather neglected by the community of Computer Vision researchers. Our approach called Unsupervised Object-aware Learning shows a promising direction for this novel setup. As a part of our contribution, we also release a dataset designed to naturally contain the notion of image bundles which is required by our approach.

ACKNOWLEDGMENT

The work was supported from ERDF/ESF "Centre for the development of Artificial Intelligence Methods for the Automotive Industry of the region" (No. CZ.02.1.01/0.0/0.0/17_049/0008414). The author would like to thank to Anand Pratap Singh and Hana Zamecnikova for their help with the creation of the dataset.

REFERENCES

[1] R. Hadsell, S. Chopra, and Y. LeCun, "Dimensionality reduction by learning an invariant mapping," in *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, vol. 2. IEEE, 2006, pp. 1735–1742.

[2] L. McInnes, J. Healy, and J. Melville, "Umap: Uniform manifold approximation and projection for dimension reduction," *arXiv preprint arXiv:1802.03426*, 2018.

[3] C. J. Mitchell, "Human and animal perceptual learning: Some common and some unique features," *Learning & behavior*, vol. 37, no. 2, pp. 154–160, 2009.

[4] A. M. Zador, "A critique of pure learning and what artificial neural networks can learn from animal brains," *Nature communications*, vol. 10, no. 1, pp. 1–7, 2019.

[5] E. S. Spelke, "Core knowledge." *American psychologist*, vol. 55, no. 11, p. 1233, 2000.

[6] Q. Wang, L. Zhang, L. Bertinetto, W. Hu, and P. H. Torr, "Fast online object tracking and segmentation: A unifying approach," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 1328–1338.

[7] J. Tobin, R. Fong, A. Ray, J. Schneider, W. Zaremba, and P. Abbeel, "Domain randomization for transferring deep neural networks from simulation to the real world," in *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2017, pp. 23–30.

[8] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," *arXiv preprint arXiv:1312.6114*, 2013.

[9] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre, "Fast unfolding of communities in large networks," *Journal of statistical mechanics: theory and experiment*, vol. 2008, no. 10, p. P10008, 2008.

[10] F. Schroff, D. Kalenichenko, and J. Philbin, "Facenet: A unified embedding for face recognition and clustering," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 815–823.

[11] Q. Xie, E. Hovy, M.-T. Luong, and Q. V. Le, "Self-training with noisy student improves imagenet classification," *arXiv preprint arXiv:1911.04252*, 2019.

[12] H. Bagherinezhad, M. Horton, M. Rastegari, and A. Farhadi, "Label refinery: Improving imagenet classification through label progression," *ArXiv*, vol. abs/1805.02641, 2018.

[13] P. Bhattacharjee and S. Das, "Temporal coherency based criteria for predicting video frames using deep multi-stage generative adversarial networks," in *Advances in Neural Information Processing Systems*, 2017, pp. 4268–4277.

[14] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2223–2232.

[15] Y. Xie, E. Franz, M. Chu, and N. Thuerey, "tempogan: A temporally coherent, volumetric gan for super-resolution fluid flow," *ACM Transactions on Graphics (TOG)*, vol. 37, no. 4, p. 95, 2018.

[16] I. Davidson and S. Ravi, "Clustering with constraints: Feasibility issues and the k-means algorithm," in *Proceedings of the 2005 SIAM international conference on data mining*. SIAM, 2005, pp. 138–149.

[17] S. Ravi and I. Davidson, "Agglomerative hierarchical clustering with constraints: Theoretical and empirical results," in *European Conference on Principles of Data Mining and Knowledge Discovery*. Springer, 2005, pp. 59–70.

[18] S. Basu, M. Bilenko, A. Banerjee, and R. J. Mooney, "Probabilistic semi-supervised clustering with constraints," *Semi-supervised learning*, pp. 71–98, 2006.

[19] S. Abu-El-Haija, N. Kothari, J. Lee, P. Natsev, G. Toderici, B. Varadarajan, and S. Vijayanarasimhan, "Youtube-8m: A large-scale video classification benchmark," *arXiv preprint arXiv:1609.08675*, 2016.

IEEE Third International Conference on Data Stream Mining & Processing
August 21-25, 2020, Lviv, Ukraine

3D Fire Front Reconstruction in UAV-Based Forest-Fire Monitoring System

Vladimir Sherstjuk
Department of Software Tools and Technologies
Kherson National Technical University
Kherson, Ukraine
vgsherstyuk@gmail.com

Maryna Zharikova
Department of Software Tools and Technologies
Kherson National Technical University
Kherson, Ukraine
marina.jarikova@gmail.com

Irina Dorovskaja
Department of Technologies
European University
Kryvyi Rih, Ukraine
irina.dora07@gmail.com

***Abstract*—This work presents a new method of 3D reconstruction of the forest-fire front based on uncertain observations captured by remote sensing from UAVs within the forest-fire monitoring system. The use of multiple cameras simultaneously to capture the scene and recognize its geometry including depth is proposed. Multi-directional observation allows perceiving and representing a volumetric nature of the fire front as well as the dynamics of the fire process. The novelty of the proposed approach lies in the use of soft rough set to represent forest fire model within the discretized hierarchical model of the terrain and the use of 3D convolutional neural network to classify voxels within the reconstructed scene. The developed method provides sufficient performance and good visual representation to fulfill the requirements of fire response decision makers.**

Keywords— forest fire monitoring; fire front; remote sensing; process reconstruction; neural network; voxel; soft rough set.

I. INTRODUCTION

Due to global climate change, industrialization, urbanization, population growth, and the other outstanding features of new century, intensive forest fires have become a whole-planetary problem. They grow year by year, so forest fire response operations become challenging and expensive. Traditionally, these operations continue to be based on visual observations and decision-maker's estimations. However, smoke and flame substantially distort observations, while high temperature does not allow to approach the fire closely. That is why, inaccurate and incomplete observations cannot be a reliable basis for planning response operations. Although a forest fire can be considered as a poorly modeled and unpredictable process, it is well known that its fast development requires always a high responsibility of decision-maker with acute lack of time. It is also known that the efficiency of response operations depends mainly on the availability and usability of real-time forest fire monitoring tools. Therefore, today a lot of attention is paid to unmanned vehicles, remote sensing, image processing, and a range of other modern tools and technologies that can be synergistically used to forest fire monitoring. Since unmanned aerial vehicles (UAV) can fly closely enough to the fire carrying optical and infrared cameras as remoted sensors, they are the most suitable tools for the real-time forest fire monitoring [1].

However, obtaining a real and reliable picture of the ongoing fire spreading process is not an easy task due to effects of wind, smoke, and fire. There are many other factors such as multiplicity of fire spots, segmentation of the fire front, spatial and temporal variability of weather, fuels, and topography conditions, which make also their contribution. All of this complicates obtaining credible grounds for decision-making on planning and executing fire response operations. Decision-maker always requires a clear picture of the ongoing processes in order to understand a direction and a rate of fire spreading. At the same time, existing implementations of real-time forest fire monitoring systems offer a result in the form of a flat two-dimensional image of a burning area on a map [2]. Thus, the topic of our interest is a study of the ways of 3D reconstruction of the forest fire process during its real-time monitoring, which can provide decision-maker with a clear model of the fire front spreading to make in-time decisions on the fire response.

II. LITERATURE ANALYSIS

The considered problem is very close to three-dimensional (3D) object reconstruction that is a computer vision task having several applications including robotics, object tracking, etc. The object reconstruction is always reduced to taking a set of scans of the object's surface with a certain sensor or scanner, which must be located in different points due to both a limited field of view and occlusions of the objects [3]. A 3D model of the object can be generated by a physical sensing its surface from several views. Since objects have various shapes and sizes, and sensor locations have certain observation constraints, it is necessary to choose different views carefully. Usually, the set of views around the object can be chosen manually by operator [4] (in process or predefined), or automatically by a view planning algorithm finding proper locations for sensors [5].

Today, there are many well studied online and offline object reconstruction algorithms [6]. Online Simultaneous Localization and Mapping (SLAM) algorithms are based on an incremental appearance-based loop closure detector and usage of range sensors such as laser or stereo cameras. They are quite robust, have revisiting capability and work in real time. However, in general they provide insufficient opportunities for the detailed object reconstruction due to sufficiently sparce observations, although the shapes, features and colors can be recognized as good as possible [7]. The detailed review of SLAM odometry-, depth sensor-based, and other methods including their advantages and performance are presented in [8]. Offline algorithms such as Multi-View Stereo (MVS) are aimed at finding pairwise stereo correspondences to estimate dense and accurate

978-1-7281-3214-3/20/$31.00 ©2020 IEEE

reconstructions [6]. Sparse feature matching and patch growing method with photometric and visibility constraints is proposed in [9]. In [10], a Next-Best View (NBV) approach is proposed to select feasible stereo pairs views. This approach is expanded in [11] to use multi-views. However, the computational complexity of such methods is too high to use them in the real-time. A volumetric method based on hierarchical octree reconstruction was also proposed in [12]. Such methods are divided into real 3D methods, which describe the modeled object by the 3D point clouds returned by sensors, and 2.5D methods, which describe only measured height (depth) for each cell within a certain 2D grid. Point clouds can not distinguish free and unknown areas, so they are mainly suitable in static environments [13]. Besides that, they are sensitive overmuch to sensor noise. Height (depth) measurements represented as elevation map are mainly used to model the outdoor environment having a single surface [14]. Thus, they only discretize the environment vertically but do not provide its volumetric representation [15]. To overcome this problem, a hybrid approach is proposed in [16], where each cell in a 2D grid stores a list of vertically ordered voxels. Another important area of improvement is concerned with the use of Next-Best View algorithms to enhance the efficiency of the preformed reconstruction. NBV methods can improve quality and coverage using a minimum amount of data [17].

It should be noted that forest fire is not an object but a dynamic process, within which the presence of a flame is a certain "eigenfeature". From the image recognition point of view, such eigenfeature is significantly variable with respect to its shape and color. There are also effects of smoke, flares, and flickers, which significantly complicate the recognition and distort the picture of the process. Moreover, forest fires are always spreading, and a rate of fire spreading is also variable under the influence of wind, type of vegetation, and other factors. Despite the overall progress in the field of object reconstruction, the issues of 3D process reconstruction based on multiple observations still remain open and have a little reflection in the literature. We can reduce the problem of the fire process reconstruction to the problem of the reconstruction of the fire front. However, this process has such decisive dynamics that all above-considered methods are poorly applicable to solve the reconstruction problem. The solution of the process reconstruction problem must be three-dimensional (volumetric), must take into account the inaccuracy and uncertainty of the observations and provide adequate representation of the dynamics of the real process. The images captured by the cameras mounted on the UAVs can provide information necessary for such reconstruction. However, the geometry of the fire front does not allow observing the process completely from one viewpoint, even in the case of the continuous movement of the camera. Since the viewpoint position cannot get to the fire front closer than a certain safe distance, a simultaneous observation from several different points is required to solve the reconstruction problem. Since several issues related to the forest fire front reconstruction still remain insufficiently investigated, this needs further research.

III. Problem Statement

A forest-fire monitoring system must provide real-time information to the decision maker during response operation. The most important information regarding the forest fire process is information about the dynamics of the fire front. Understanding this dynamic enables decision-maker to assess fire intensity based on the rate of fire front spreading and the estimated length/height of flame. Thus, reconstruction of the fire front should mainly be aimed at the determining its geometric parameters within the spatial and temporal scale. Actually, forest fire monitoring systems involve UAVs equipped with optical and infrared cameras to remote sensing that allows identifying the features of flame and smoke on the base of the processing of the captured images [18]. Although both optical and infrared sensors are sensitive to noise and interferences so the observations are often ambiguous, imprecise, and inconsistent, nonetheless, they can be used for obtaining the feasible estimates of the fire front parameters. The main questions will be how accurate the reconstruction of the fire front is and how much the degree of its credibility is. Fortunately, in most cases, decision-maker does not require precise estimates of fire front parameters, but he needs to know how the fire front evolves over time and understand the course of the process and its drivers. Clearly, a visual representation of the fire front dynamics can be a valuable information for making adequate decisions.

Therefore, the aim of this work is to develop a method of 3D reconstruction of the fire front based on uncertain observations captured by remote sensing from UAVs within the forest fire monitoring system. We propose to use multiple cameras simultaneously to capture the scene and recognize its geometry. We assume that multi-directional views of the process can be used to estimate volumetric nature of the fire front as well as the fire processes.

IV. Remote Sensing and Image Processing in Forest-Fire Monitoring System

A. Observable attributes of the forest fiore

The main attributes of the forest fire are heat, smoke, and flame, including such its manifestations as light, flicker, and motion [19]. All of them are observable by sensors.

Flame emits their own visible light and can be considered as "eigenfeature" of the forest fire. Usually, flame has also such visible properties as flickers, flares, movement, and transparency [19]. Depending on the temperature, the flame color can vary from dark red to light yellow and even up to white at the developed stage of the fire [20]. Despite of the variability of its color, flames are usually distinctive to recognition on images with respect to the background. Another visual feature of flame is its shape. However, it can essentially vary depending on fuel consuming and composition, wind variation, etc. Thus, the flame shape recognition is more challenging. Obviously, the image recognition can be complicated by interferences (i.e., sunlight) and distortions (shadow, smoke, flares, and flickers) that affect images. Moreover, smoke and precipitation must be considered as noise.

Smoke is always a reliable visual attribute for the detection of forest fires but during the fire monitoring it is rather an occlusion that masks eigenfeature and obscures the visibility of flame. Smoke has also dynamic properties changing fire's shape, size and color as well as rising in plumes. Usually, smoke can be recognized through low values of chrominance in captured images [19].

Heat is usually transported from the fire by convection, conduction, and radiation but only radiated heat can be remotely sensed and measured by infrared cameras.

Unfortunately, such measurements do not have a depth effect, so they cannot be directly used to reconstruct a fire process. However, images captured by infrared camera can be used to unmask a flame covered by smoke.

B. Sensors

Modern electro-optical cameras have a high enough resolution and wide field of view, but the quality of their images depends heavily on lighting conditions. Thus, they are sensitive to darkness (nighttime), smoke and precipitations such as rain, haze, mist, fog, etc. Although the image processing algorithms are mainly based on color properties, they should also analyze flickering of pixels over time as well as intermittency and irregular oscillations of the edges of the flame regions since the latter usually vary in height, size, and brightness. Smoke can be identified by low values of chrominance and variations in color and density.

The broadband thermal infrared camera measures energy release within the combustion reaction but it has a limited dynamic range. Images captured by infrared (IR) camera are usually affected by such interferences as saturation, reflected sunlight, energy radiated from non-fire sources, etc. In daytime, its images have too small contrast. However, IR camera can be used overnight due to good flame-to-background contrast and under the smoke conditions because smoke is quite transparent on the corresponding wavelength.

Thus, both optical and infrared cameras have their drawbacks and can provide imprecise, uncertain, or ambiguous information in captured images because of interferences and distortions. Turbulence and vibrations of UAV distort captured images additionally. Therefore, we cannot build a clear and accurate model of the fire front based on remote sensing information, but we can reconstruct a model of the fire front approximately.

C. Remote Sensing

Monitoring UAVs need to have hovering capabilities providing a respectively long hovering time. Usually, multi-rotary-wing UAVs equipped with a gimbal carrying both infrared and high-resolution electrooptical camera with pan and tilt units are used for monitoring purposes. In [2], the monitoring UAVs are equipped with the 16-Mp optical camera (5376x3024 pixels) providing scalability of images, the thermal infrared camera, weather sensors, GPS receiver, and inertial measurement unit for self-localization and navigation purpose.

Monitoring UAVs capture images flying towards the fire and hovering around it. Obviously, such UAVs should perform onboard only stabilization, geo-localization, and geo-rectification of images (Fig. 1). Their sensors capture and transmit images to the ground control station. The digital elevation model represents cartographic dataset describing terrain surface, so the captured images are transformed to a stream of geo-mapped frames where each frame is complemented with the coordinates of the upper left corner of the frame and scale value. This makes it possible to merge images taken from different positions of observation.

D. Image processing

It is assumed that that fire spreads in three-dimensional space C above the terrain discretized by a grid $D=\{d_{xyz}\}$ of isometric cubic cells d_{xyz} with the size being $\delta\times\delta\times\delta$.

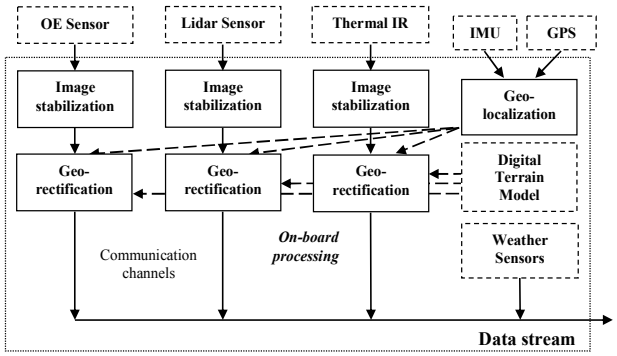


Fig. 1. Onboard image processing

There are three channels to process image from each UAV at the ground center: one for image frames captured by infrared camera and two separate channels providing flame and smoke recognition based on the image frames captured by optical cameras (Fig. 2). In the infrared channel, image pixels represent a heat radiation by colors ranged from black to white, so the image analysis is performed in three stages: image mapping, image averaging, and gray color evaluation. Burning areas are always represented as white areas within the image, while non-burning areas are black. Clearly, a plenty of pixels is greyed due to uncertainty of observations. As the result of the image analysis, each cell $d_{ijk}\in D$ is associated with a certain "degree of grayness" μ_{ijk} ranged in the interval $[0..1]$ and based on the average brightness B_{ijk} of this cell.

In the optical channels, the process of image analysis is performed in such sequential stages as mapping, transformation (only in optical camera channels), averaging, filtering, generalization, and conversion of images. This analysis is quite similar in smoke and flame channels and performed in HSI color space, but the differences are in the intervals of processed hue and saturation values. In the flame channel, color can vary from dark red to light yellow with the high saturation while in the smoke channel, color can be in gray-and-silver interval with low saturation but high intensity. As the result of image processing, each cell d_{ijk} is associated

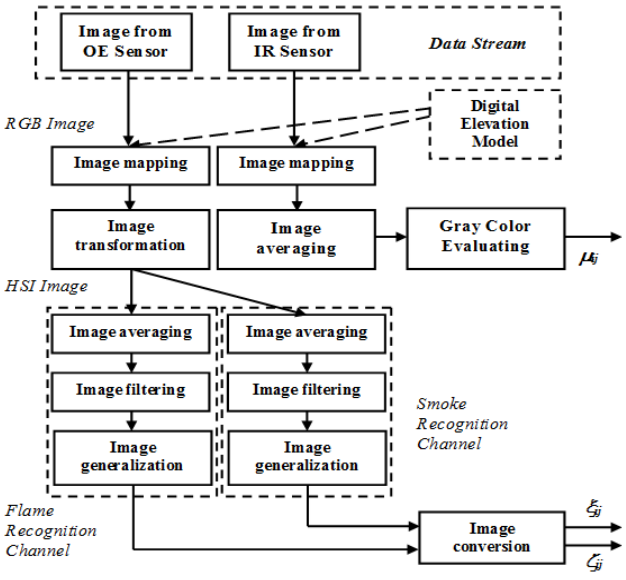


Fig. 2. The image analysis at the ground center

with certain degrees of flame (ξ_{ijk}) and smoke (ζ_{ijk}) recognized within this cell and normalized in the range [0, 1].

Finally, the values ξ_{ijk} , ζ_{ijk} , and μ_{ijk} are processed into merged value η_{ijk} with respect to each cell d_{ijk} . The final stage of the image analysis consists of image filtering, cleaning and merging. Such multi-stage processing algorithm allows reducing interferences, distortions, and noise during remote sensing. However, the implementation of this method in [2] implies averaging images captured from different points of view (different UAVs) and projecting the resulting grayscale image frame based on values η_{ijk} for the cells $d_{ijk} \in D$ into a two-dimensional plane $d_{ij} \in D'$ lying "at the ground" to recognize the flat forest fire front.

V. Method of 3D Fire Front Reconstruction

To obtain the volumetric fire front representation, we must reconstruct it from uncertain observations based on the results of the image analysis described in the previous section. However, such reconstruction is quite challenging because of fire spreading. Fig. 3 shows the forest fire observed by UAVs' sensors from two distinct viewpoints. Obviously, a part of the fire process is masked with smoke, which spreads out mainly in the wind direction. This makes it impossible to directly observe the combustion process covered by smoke. Therefore, the viewpoints for observation are often selected at the opposite sides (windward A vs leeward B in Fig. 3) of the fire front. Nevertheless, a significant part of the combustion process remains hidden due to presence of hidden areas and occlusions in the observed scene such as shown in Fig. 3. Furthermore, fire fronts can cover a sufficiently deep area that is poorly visible from two opposite viewpoints. Often, they are segmented and even broken up into several parts.

The fire front spreads in such a way that the fuel burns out inside of the fire perimeter remaining burned-out areas, but due to the environmental effects, new areas of vegetation are covered by burning outside the fire perimeter. Accordingly, monitoring UAVs should move avoiding unsafe areas and choosing positions that can maximize gaining of observable information overcoming uncertainty.

A. Scene geometry

The forest fire scene is represented in Fig. 4. A configuration of viewpoints can be defined as a set of poses (Pose A, B, C in Fig. 4). The pose is a description representing the UAV location and the sensor orientation

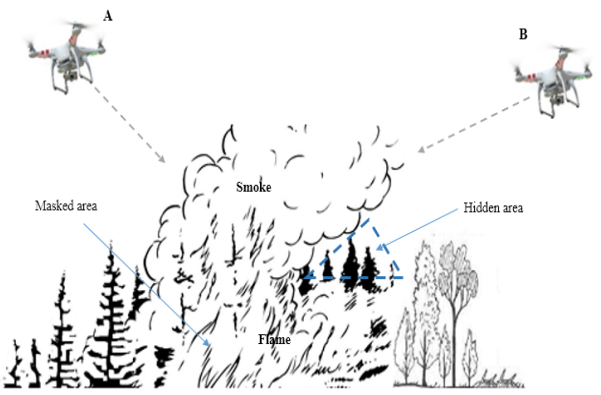


Fig. 3. Multi-view fire front monitoring

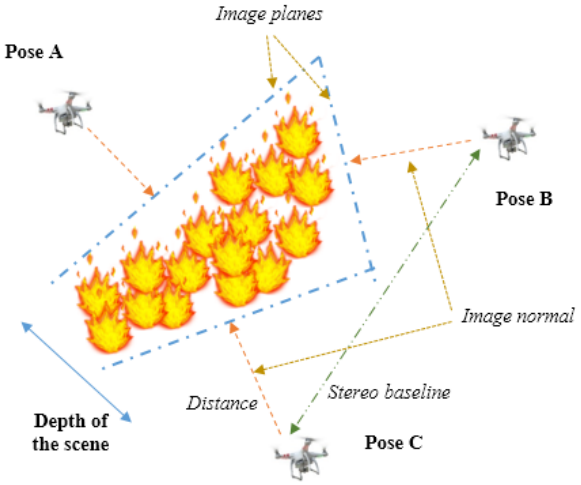


Fig. 4. The reconstructed scene

within three-dimensional space. Usually, the sensor orientation is chosen in accordance with the normal to the image plane. Each pair of UAVs can be involved in a stereo-pair view; therefore, they must be located in accordance with stereo baseline determined through the crossing of their image normals at the certain point at the maximal depth of the scene. Thus, images from UAV cameras can be used to scene depth perception both on its own (through stereo-view) and with the help of additional neural network processing. Determining a set of optimal poses is a separate independent research task, which will not be considered in this paper.

B. Spatial Representation

The grid of cells D used in the image processing at the ground center is not suitable for the information presentation during the process reconstruction because of the spreading of the fire front. Thus, the initial grid needs to be extended with new spaces dynamically along with fire spreading process while burned-out areas lose their value. Consequently, we must initially build a space such big as the maximal bounds of the considered forest fire, but they are not known. Therefore, we cannot use the grid D due to performance requirements and memory limitations.

We create a new 3D structure to reconstruct fire front based on a grid of voxels, which are considered as certain cubic volumes of equal size. Although voxels are a bit like cells that discretize a space in a grid D , nonetheless, they are organized in a completely different way. Each voxel is considered as a node in a certain tree-like structure called octree. Each node of the octree can be recursively divided into 8 sub-nodes, each of which is also a voxel but has a smaller size. Such recursive process can be both top-down and bottom-up. Initially, we can start from the voxel that covers the overall forest fire area and subdivide it recursively down to the minimum voxel size reachable with respect to the sensor resolution. After the fire front achieves a bound of the discretized area, we can just move a step up and create a voxel containing the original voxel as a sub-node. As the result, we describe a hierarchical data structure based on octrees that can be cutted at any level of hierarchy. Such hierarchical data structure known as OctoMap can be used to representation of the reconstruction scene, where voxels constitute a certain 3D vector where each unit is a voxel. We use a revised version of an open source library OctoMap [21].

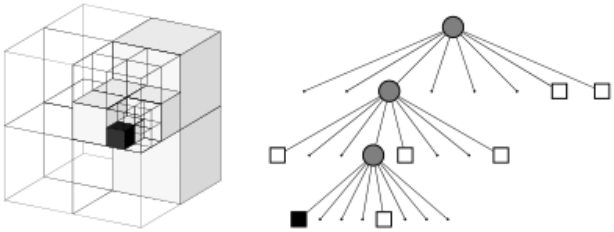


Fig. 5. OctoMap data structure [21]

C. Scene Representation

Within the scene, each voxel belongs to a specific class w_i, which describes the type of its inner content within the considered reconstruction scene:

- Unmarked ($i=0$). It is a type of voxels that have not been seen by sensors. Initially, the reconstruction scene is represented by a 3D vector filled by unmarked voxels.
- Empty ($i=1$). It is a type of voxels that lie between the sensor position and sensed "surface" of the fire process represented by flame and smoke. Such voxels usually represent a "free of fire" space.
- Flame ($i=2$). These voxels represent a remotely sensored "surface" of the fire. Such voxels correspond to a burning kernel of the fire process.
- Smoke ($i=3$). This type of voxels represents the areas within the scene shrouded in smoke, which prevents sensoring the inner voxels. Such voxels can mask "surface" of the fire.
- Uncertain ($i=4$). It is a type of voxels that can not be precisely assessed as "burning", i.e. voxels that are possibly involved in the combustion process due to the uncertain observation but there is no certainty about it.
- Burnt ($i=5$). Such voxels represent the areas that are already burnt and, therefore, cannot be involved in the combustion processes.
- Fuel ($i=6$). This type of voxels corresponds to the vegetation areas that do not participate in combustion processes but can be ignited due to readiness of the fuel. It should be noted that voxels, which do not contain flammable vegetation, belong to the "empty" class instead of the "fuel" class.
- Occluded ($i=7$). These voxels represent the areas occluded by other voxels that prevent perception of the fire front depth.

D. Classification of Voxels

A 3D convolutional neural network (3D-CNN) can be used to classify voxels within the reconstructed scene. It is a special kind of CNNs using the 3D volumes in the kernels instead of 2D maps. We propose to use VoxNet [22] architecture (Fig. 6). The first convolutional layer has 3 channels with 128 features, a kernel of size 5×5×5, stride 2×2×2, and max pooling operation of stride 2×2×2; the second one has 32 features, a kernel of size 4×4×4, stride 2×2×2, and max pooling operation of stride 2×2×2; and the third one - 8 features, a kernel of size 3×3×3, stride 2×2×2, and max pooling operation of stride 2×2×2. Three fully connected layers have 1200, 400, and 50 parameters as output.

E. Soft Rough Fire Front Model

The model of the fire front can be represented using soft sets of cells [23]. Considering a set of voxel classes $W=\{w_0,...,w_7\}$ and a 3D vector of voxels V, we can define a *soft set of voxels* over V as a pair (Υ,W) where Υ is a mapping of W into the set of all subsets of the set V. This set can be defined as $\Upsilon(t)=\{(w_i,\Upsilon(w_i,t)): w_i\in 2^W, \Upsilon(w_i,t)\in 2^V\}_{i=0}^{7}$, where each $\Upsilon(w_i,t)$ is an w_i-element of the soft set (a set of voxels of a class $w_i\in W$ at the reconstruction moment $t\in T$). Using the voxels of a class w_2, we can define a lower approximation containing the cells, which definitely belong to the w_2-element of the soft set $\underline{\Upsilon}(t)=\Upsilon(w_2,t)$, while using the voxels of classes w_3 and w_4, we can define an upper approximation containing the cells, which possibly belong to the w_3 - and w_4 - elements of the soft set, $\overline{\Upsilon}(t)=\Upsilon(w_3,t)\cup\Upsilon(w_4,t)$. As the result, both lower and upper approximations constitute a soft rough set of voxels $\hat{\Upsilon}(t)=\{\underline{\Upsilon}(t),\overline{\Upsilon}(t)\}$ that represent a 3D fire front model at the reconstruction time. Other voxels belong to the negative area of the rough set $NEG(\hat{\Upsilon}(t))=V-\overline{\Upsilon}(t)$. The boundary area of the fire process is a subset of the set of voxels, which belong to the upper approximation, but don't belong to the lower approximation, $BND(\hat{\Upsilon}(t))=\overline{\Upsilon}(t)-\underline{\Upsilon}(t)$.

VI. Implementation

The proposed method of 3D reconstruction of the forest fire front has been implemented using Visual C++, OctoMap framework, ConvNet and Fast Artificial Neural Network (FANN) libraries. The software prototype has been tested on PC based on the Pentium i5-7400 3-3,5 GHz processor and 16 GB RAM, the reconstruction rate has been evaluated during the simulation. The initial weights of neural networks have been randomly generated. The simulation results show that the method can achieve an accuracy of reconstruction up to 98% (Fig. 7), which shows the rate of correctly classified voxels (true positive) from the prepared test set. Thus, the method is effective for the forest fire front reconstruction, it has

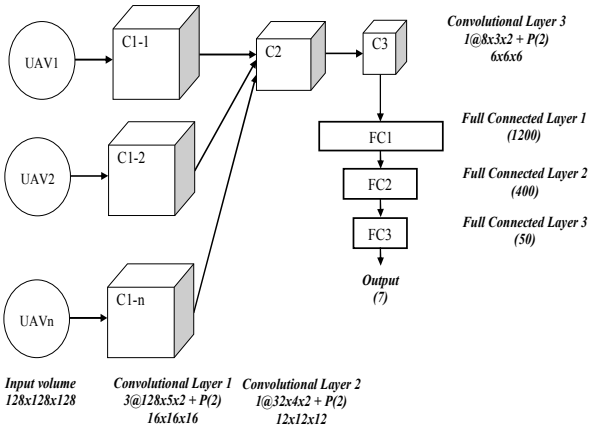


Fig. 6. 3D-CNN Architecture

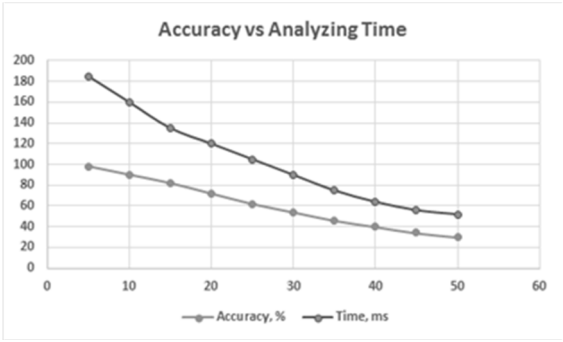


Fig. 7. Simulation results

relatively the same performance as the method of the fire front recognition proposed in [24]. At the same time, reconstructed 3D model of the fire front gives to decision-maker a significantly better visual representation of the forest fire dynamics (Fig. 8), so the credibility of fire response decision making is respectively increased, which makes the proposed method of practical use.

VII. Conclusion

The proposed in the paper method of 3D reconstruction of the forest fire front is based on uncertain observations captured by remote sensing from UAVs within the forest fire monitoring system. We suggest to use multiple cameras simultaneously to capture the scene and recognize its geometry including depth. Multi-directional observation allows us to perceive and represent a volumetric nature of the fire front as well as the dynamics of the fire processes. The novelty of our approach lies in the use of soft rough set to represent forest fire model within the discretized hierarchical model of the terrain as well as in the use of 3D convolutional neural network to classify voxels within the reconstructed scene. The developed method provides sufficient performance and good visual representation to fulfill the requirements of fire response operations.

References

[1] 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.

[2] V. Sherstjuk, M. Zharikova, and I. Sokol, “Forest Fire-Fighting Monitoring System Based on UAV team and Remote Sensing,” In Proc. of 2018 IEEE 38th Int. Conf. on Electronics and Nanotechnology. IEEE, Kyiv, 2018, pp. 663–668.

[3] J. I. Vasquez-Gomez, L. E. Sucar, R. Murrieta-Cid. and E. Lopez-Damian, “Volumetric Next-Best-View Planning for 3D Object Reconstruction with Positioning Error,” Int. J. of Advanced Robotic Syst., vol. 11, no. 10, pp. 1–13, 2014.

[4] V. Lippiello, F. Ruggiero, B. Siciliano, and L. Villani, “Visual grasp planning for unknown objects using a multifingered robotic hand,” in IEEE Trans. on Mechatronics, vol. 18, no. 3, pp. 1050–1059, 2013.

[5] W. R. Scott, G. Roth, and J.-F. Rivest, “View planning for automated three-dimensional object reconstruction and inspection,” ACM Comput. Surv., vol. 35, no. 1, 2003, pp. 64–96.

[6] O. Mendez, S. Hadfield, N. Pugeault and R. Bowden, "Taking the Scenic Route to 3D: Optimising Reconstruction from Moving Cameras," 2017 IEEE Int. Conf. on Computer Vision (ICCV), Venice, 2017, pp. 4687–4695.

[7] R. Mur-Artal, J. M. M. Montiel and J. D. Tardós, "ORB-SLAM: A Versatile and Accurate Monocular SLAM System," in IEEE Transactions on Robotics, vol. 31, no. 5, pp. 1147–1163, Oct. 2015.

[8] J. Delmerico, S. Isler, R. Sabzevari, D. Scaramuzza, “A comparison of volumetric information gain metrics for active 3d object reconstruction,” Autonomous Robots, vol. 42, pp. 197–208, 2018.

[9] Y. Furukawa and J. Ponce, "Accurate, Dense, and Robust Multiview Stereopsis," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 32, no. 8, pp. 1362–1376, 2010.

[10] M. Jancosek, A. Shekhovtsov and T. Pajdla, "Scalable multi-view stereo," 2009 IEEE 12th Int. Conf. on Computer Vision Workshops, ICCV Workshops, Kyoto, 2009, pp. 1526–1533.

[11] S. Galliani, K. Lasinger and K. Schindler, "Massively Parallel Multiview Stereopsis by Surface Normal Diffusion," 2015 IEEE Int. Conf. on Computer Vision (ICCV), Santiago, 2015, pp. 873-881.

[12] S. Isler, R. Sabzevari, J. Delmerico and D. Scaramuzza, "An information gain formulation for active volumetric 3D reconstruction," 2016 IEEE Int. Conf. on Robotics and Automation (ICRA), Stockholm, 2016, pp. 3477–3484.

[13] A. Nüchter, K. Lingemann, J.Hertzberg, and H. Surmann, “6D SLAM – 3D mapping outdoor environments,” J. Field Robot., vol. 24, no. 8–9, pp. 699–722, 2007.

[14] R. Triebel, P. Pfaff and W. Burgard, "Multi-Level Surface Maps for Outdoor Terrain Mapping and Loop Closing," 2006 IEEE/RSJ International Conference on Intelligent Robots and Systems, Beijing, 2006, pp. 2276-2282.

[15] J. S. Gutmann, M. Fukuchi, M. Fujita, “3D perception and environment map generation for humanoid robot navigation. Int. J. Rob. Res. Vol. 27, no. 10, pp. 1117–1134, 2008.

[16] J. Ryde, H. Hu, “3D mapping with multi-resolution occupied voxel lists,“ Autonomous Robots, vol. 28, no. 2, pp. 169–185, 2010.

[17] E. Marchand and F. Chaumette, "Active vision for complete scene reconstruction and exploration," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 21, no. 1, pp. 65–72, 1999.

[18] V. Sherstjuk, M. Zharikova and I. Sokol, "Forest Fire Monitoring System Based on UAV Team, Remote Sensing, and Image Processing," 2018 IEEE Second International Conference on Data Stream Mining & Processing (DSMP), Lviv, 2018, pp. 590–594.

[19] R. Allison, J. Johnston, G. Craig, and S. Jennings, “Airborne Optical and Thermal Remote Sensing for Wildfire Detection and Monitoring,” Sensors, vol. 16, no. 8, 1310, 2016.

[20] Y. Qiang, B. Pei, and J. Zhao, “Forest Fire Image Intelligent Recognition based on the Neural Network,” Journal of Multimedia, vol. 9, no. 3, pp. 449–455, 2014.

[21] A. Hornung, K.M. Wurm, M. Bennewitz, C. Stachniss, W. Burgard, “Octomap: An efficient probabilistic 3d mapping framework based on octrees,” Autonomous Robots, vol. 34, no. 3, pp. 189–206, 2013.

[22] D. Maturana and S. Scherer, "VoxNet: A 3D Convolutional Neural Network for real-time object recognition," 2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Hamburg, 2015, pp. 922–928.

[23] D. Molodtsov, “Soft Set Theory – first results,” Comput. and Math. with Appl., vol. 37, pp. 19–31, 1999.

[24] V. Sherstjuk and M. Zharikova, "Fire-Front Recognition in UAV-Based Forest-Fire Monitoring System Using Fuzzy Rough Soft Sets," 2019 IEEE 2nd Ukraine Conference on Electrical and Computer Engineering (UKRCON), Lviv, Ukraine, 2019, pp. 1091–1096.

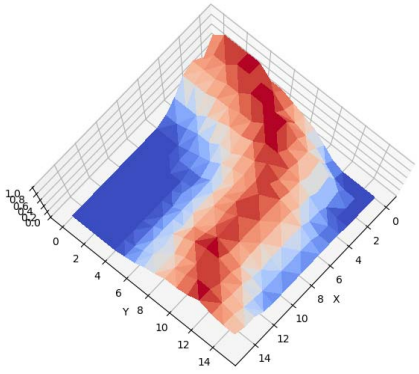


Fig. 8. Representation of the forest fire front

IEEE Third International Conference on Data Stream Mining & Processing
August 21-25, 2020, Lviv, Ukraine

Computer Vision in Control and Optimization of Road Traffic

Vladyslav Zinchenko
Intelligent Information Systems Department
Petro Mohyla Black Sea National University
Mykolaiv, Ukraine
gerogo333@gmail.com

Galyna Kondratenko
Intelligent Information Systems Department
Petro Mohyla Black Sea National University
Mykolaiv, Ukraine
halyna.kondratenko@chmnu.edu.ua

Ievgen Sidenko
Intelligent Information Systems Department
Petro Mohyla Black Sea National University
Mykolaiv, Ukraine
ievgen.sidenko@chmnu.edu.ua

Yuriy Kondratenko
Intelligent Information Systems Department
Petro Mohyla Black Sea National University
Mykolaiv, Ukraine
yuriy.kondratenko@chmnu.edu.ua

***Abstract*— This paper presents the approaches and methods used in computer vision for the detection of moving objects and their tracking. Ways of using computer vision methods to optimize and control traffic are investigated. Every day, the number of vehicles on the roads is increasing and congestion problems are getting worse. The main goal in the development of modern traffic management systems is to create effective traffic management mechanisms in accordance with dynamic traffic conditions. Nowadays, the systems that regulate traffic have many drawbacks. The main ones, that such systems are working according to a predefined program and not being aware of the proper real-time data. This paper focuses on a novel approach to road traffic management by incorporating an intelligent traffic light controlling system using an algorithm that consumes real data from closed-circuit television (CCTV) cameras. As part of the solution, have been developed a program using a popular programming platform that would calculate sets of drive orders for traffic signal lights. The main goal of the proposed system is to provide better results in terms of reduced waiting delay for pedestrians and vehicles, shorter travel time, and increased average velocity of vehicles.**

Keywords*— *traffic control, computer vision, decision-making, artificial neural network

I. Introduction

Commonly used road-level control systems have a significant disadvantage, they do not handle real-time data and work on predefined programs. Therefore, the main meta-developers of modern traffic management systems are to create efficient traffic management enterprises that are applied to dynamic results [1, 2].

Every day, the number of expensive vehicles is rapidly increasing, therefore the problems are very acute. Thus, this paper is dedicated to mathematical, algorithmic, software tools that offer the ability to automate for computer use those that work for people who remotely search for entire tasks.

Computer vision is currently a well-developed area of information technology, despite this recognition system, classification and tracking facilities are not widely implemented in everyday aspects of life [3-5]. Therefore, now we do not yet have an intellectual automated traffic management system in the urban roads, which could react to changes in urban traffic in automatic mode. The actuality of this work consists of the development of a system that will reduce time Waiting at the intersection of adjustable intersections using computer vision techniques [3].

The purpose of work is to develop and describe a system of traffic control and optimization, which will choose the mode of operation of the light, depending on the road situation.

An object of research is the process of recognition of objects and their tracking on video images. The research subject is neural network architectures used for recognition and classification of objects on images and can be applied to video.

II. Related Works and Problem Statement

Substantial filling of the image is manifested as a large variety of geometric structures. It is extremely difficult to classify the objects present in the image. The detection and identification of many types of such objects, such as buildings and roads in aerial photographs, have even turned into separate areas of research. Thus, only the problems of the selection of buildings on images have been addressed in recent years by several major international conferences, which raised only several new problems in addition to the existing ones [4, 6].

Machine vision (MV) theory offers many different model descriptions of observable objects that can be used to detect and measure them. The literature describes a wide range of such models - from simple character-based descriptions to highly specialized and sophisticated structural models [7-9].

During the last ten to fifteen years, the algorithmic aspect of the image processing has been considered in accordance with the so-called modular paradigm [5, 8]. After a long study of the mechanisms of human visual perception, D. Marr suggested a paradigm that states that image processing should

978-1-7281-3214-3/20/$31.00 ©2020 IEEE

be based on several consecutive levels of the ascending information line: from "Iconic" representation of objects (bit-map, unstructured information) - to their symbolic representation (vector and attribute data in structured form, relational structures, etc.) [9].

In recent years, in the algorithm aspect of the sequence of actions to process the image is considered in harmony with the so-called modular paradigm. Based on this, in the field of Machine vision adopted highlight the following main stages of data processing: image processing; segmentation; allocation of geometric structure; definition of relative structure and semantics [6, 8, 9].

Based on these stages we can distinguish the following levels of data processing: lower level processing, intermediate level, high level [4].

In recent years, considerable progress has been made on the problems of matching points and matching fragments, the selection of features inside small fragments, the high accuracy of three-dimensional positioning of points, which implies appropriate modeling and calibration of sensors and their combinations, selection of simple geometric structures such as "point" or "edge" [3-5, 7].

Currently, there are several basic algorithmic approaches and mathematical formalisms used in the development of practical image analysis systems. These are histogram transformation, projection analysis, linear and nonlinear image filtering, brightness and textural segmentation, correlation detection and consistent filtering, Yu. P. Pitiev's morphological approach, Serra's mathematical morphology, Haf's transformation method, a structural and linguistic approach and many others [8, 9].

Accordingly, nowadays, the systems that regulate traffic have many drawbacks. The main ones, that such systems are working according to a predefined program and not being aware of the proper real-time data. This paper focuses on a novel approach to road traffic management by incorporating an intelligent traffic light controlling system using an algorithm that consumes real data from closed-circuit television (CCTV) cameras.

III. Overview of Methods for Object Recognition

Let's formulate the main requirements methods and algorithms of machine vision on the example of the most specific group of algorithms for identifying objects in images. At the same time, the authors consider three main types of requirements: stability, accuracy, computational implementation.

Pattern recognition is a very broad scope, many tasks from this area are already selected. Firstly, it is necessary to understand the basic concepts of this process. Class is the set of objects that have common properties. Classification is the process of distributing objects by class, according to their characteristics. Classifier is a tool that the input receives a set of properties of the object, and the output issue class to which the object should be assigned. Verification is the process of matching a specific object with the object model or with the description of a class. The symptom is the quantitative description of the properties of the investigated object or phenomenon [10, 11].

In general, distinguish 3 groups of recognition methods by tools. *Pattern Comparison* group includes methods and techniques that use the distance category and the approach (classification with the closest average, the distance to the nearest neighboring value, etc.). *Statistical methods* based on probability calculation (for example, the Bayes's method). *Neural networks* reveal the possibility of recognition during network training [12-15].

Consider also give another alternative classification of methods by the method of submission. *Methods for full collation*. Among the advantages is that a minimum of information may be lost. But the disadvantage is that these methods are quite expensive, because it is necessary to process a large amount of information, so they are not always productive. *Methods of matching signs or structural methods*. The mapping of objects occurs according to certain characteristics, features. The main advantage of such methods is that they are resistant to changes in position, illumination, size, etc. But the disadvantage is that it is quite difficult to automatically distinguish the optimal set of signs for classification as well as high complexity of computing. *Hybrid methods*. Most similar to the principles of human vision: the object is evaluated as a whole and its characteristics [8, 9, 16].

Classification of the nearest average. The object on the image appears as a vector of elementary signs. The set of vectors objects can be known in advance as a result of training or are provided based on any models. The classification algorithm groups by the reference data of the class using a vector of the average value (mathematical expectation). For example, if the object belongs to Class *i*, it is closer to the vector of the mathematical expectation of Class *i* than the vector's mathematical expectations of other classes [11, 17].

Classification by distance to the nearest neighbor. The nearest neighbor (NN) is the simplest algorithm classification [8, 18]. NN training is reduced to the memorization of sampling. The only advantage of this algorithm is the simplicity of implementation. Disadvantages are: instability to the error; no parameters configurable were by the sample.

Statistical methods use statistical information in solving the problem of recognition. The method defines the object's belonging to a specific class based on probability p. The Bayes's approach relies on the theorem that if the distribution density of classes is known, the classification algorithm having the minimum likelihood of errors can be discharged explicitly [19]. Among the signs of naive classification are: it is necessary to define all variables and establish the dependencies between them; all variables are equally important; all variables are statistically independent. Among the disadvantages of naive classification include: only complete statistical independence of all input variables allows finding product probability; variables must be in discrete representation; consider the values of variables independently [9, 11].

Artificial neural network (ANN) is a mathematical model that allows you to collect artificial neurons into one network to solve any problem. ANN consists of layers, there is an input layer where the input signal is given, there is a source layer from which the result of the neural network is removed, and there are hidden layers between them. They can be 1, 2, 3... n. If there are more hidden layers than 1, such a network is considered deep if 1 is shallow [11, 13, 20].

There are many different architectures of deep neural network, which differ from each other by recognition accuracy, speed of recognition, amount of RAM required and other parameters. In this work, only some modern architecture will be considered. It should be noted that as a rule the ResNet and VGG architectures are used as the basis for other networks or recognition methods [8, 9, 11].

For the task discussed in this work, it is necessary to choose the most lightweight classifier, which also will be enough accuracy and speed of recognition.

Table 1 shows information about the number of parameters and layers of the analysis networks.

TABLE I. COMPARE ANNS BY THE NUMBER OF LAYERS AND PARAMETERS

Neural network	Number of parameters	Number of layers
VGG-16	138 357 544	40
VGG-19	143 667 240	46
ResNet-50	25 610 216	177
AlexNet	62 378 344	8
LeNet-5	60 000	7
YOLO	51 000 657	106
VGG-16 + Faster R-CNN	140 757 544	57

The analysis revealed that deep neural networks are suitable for the delivered task. Since the images, objects in which need to be classified, are large enough. Many the convoluted layers can highlight the remarkable features for the problem solved and the work of such a network will be effective. Also, it was decided to use integrated solutions such as YOLO and Faster R-CNN [21-23].

IV. A NEURAL NETWORK MODEL FOR OBJECT DETECTION IN A TRAFFIC CONTROL SYSTEM

Let's describe the mathematical model of neural networks and tools to work with them. The model of an ANN was proposed by Warren McCulloch and Walter Pitts in his work on the simulation of nervous activity [11].

The output signal is determined by the formula:

$$a = \varphi\left(\sum_{i=1}^{n} w_i x_i\right), \quad (1)$$

where φ is an activation function which serves to convert the input pulse to the resulting value; x_i is an input signal; w_i is the weight of the input signal.

Combining individual neurons can be obtained by a neural network. For this initial neuron, signals are fed to the entrance of the next neuron. A neural network consists of several layers, each of which may be several neurons. The layer that takes the signals from the outside world is called input. The layer that gives the signals to the outside world-weekends, other layers are called hidden.

The architecture of the convoluted network is the alternation of the convoluted layers and sub-discretizing layers. It was first proposed in the work. The structure of such a network has no reverse connections and is necessarily layered [8, 21, 24].

Network structure is the unidirectional (no backlinks), fundamentally multilayer. The convoluted layer of the neural network consists of an indication of a map of a characteristic. In such map (card), there is a kernel that is responsible for data filtering.

The number of maps is chosen depending on the task, with a greater quantity of maps, improving the quality of recognition, but it will increase the computational complexity. In most cases, the ratio of the number of maps of signs is offered to choose equal to two [8, 22].

The sizes of the roll layer maps are the same and are defined by the formula:

$$(w,h) = (mW - kW + 1, mH - kH + 1), \quad (2)$$

where w,h is the calculated map size; mW is a width of the previous map; mH is a height of the previous map; kW is a core width; kH is a core height.

The kernel acts as a filter. Its task is to find certain signs over the entire area of the map.

One of the main features of the convoluted neural network is that core also represents a system of shared weights or synapse. Multi-layer network consists of a large number of links between neurons (synapse), which slows down the recognition process considerably. The convoluted network uses the principle of total weights, which makes it possible to reduce the number of neural bonds and allows you to find a certain sign throughout the image area [23, 24].

First, the value of each map is equal to 0. The values of the cores are randomly asked in the region from minus 0.5 to 0.5. The core slips on the previous map and produces the convolution operation, which is described by the following formula:

$$(f \cdot g)[m,n] = \sum_{k,l} f[m-k, n-l] \cdot g[k,l], \quad (3)$$

where f is the source matrix of the image, and g is the convolution core; "$\cdot$" is a convolution operation.

Describe a convolution operation. Window of the kernel dimension g pass with the given step (usually 1) all the image f, at each step, the element multiply the contents of the window on the core g, the result is summed up and written into the result matrix, visually this process is presented in the Fig. 1.

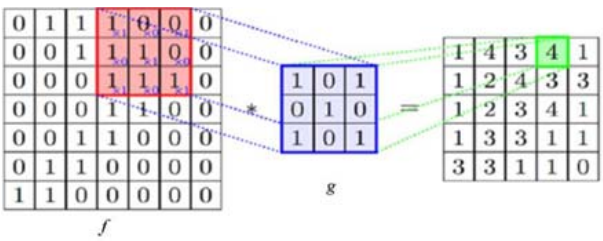


Fig. 1. Convolution operation [8].

Ultimately, the convolution layer can be described as a formula:

$$x^l = f\left(x^{l-1} \cdot k^l + b^l\right), \quad (4)$$

where x^l is the output signal of the layer l; f is an activation function; b^l is a coefficient of the layer l shear.

The scalar result of each convolution enters the activation function, which is a certain nonlinear function. It is necessary in order to bring to the network nonlinearity.

One of the most popular activation features is the ReLU (rectified linear unit) function. Advantages of using this function include: function is not saturated; the use of this function increases the convergence rate of stochastic gradient descent in comparison with Sigma and hyperbolic tangent; does not require complex calculations, as the conversion is performed. Among the disadvantages distinguish insufficient reliability of the function [25].

The DBMS and the scraping layer (the subset layer) are also as well as the con-joined maps, but their number coincides with the previous (conjoined) layer. The purpose of the layer is to reduce the size of the maps of the previous layer. If there were already certain signs in the previous transaction, then for further processing, such a detailed image is no longer required, and it is compacted to less detailed. Besides filtering of unnecessary details prevents retraining

In the process of processing the kernel layer subset (filter) maps of the previous layer, the scan engine does not intersect in contrast to the convoluted layer. Usually, each map has a core size of 2x2, which allows reducing the previous map of the convolution layer by 2 times. The entire feature map is divided into cells of 2x2 elements, of which maximum values are selected.

The ReLU [26] activation function is used normally in the subset layer. In Fig. 2, given operation of subsampling.

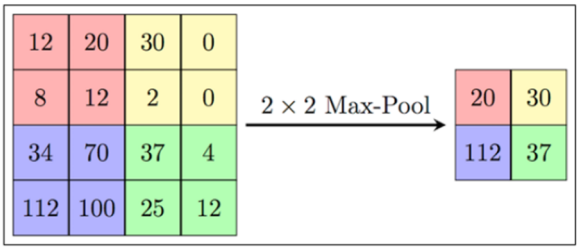


Fig. 2. The operation of creating a subset, forming a new layer [26].

The full layer can be considered as a normal layer of a multilayer perceptron. The purpose of the layer is a classification, simulating complex nonlinear function, which improves OCR quality [23].

The main idea of layer exclusion is an instead of learning one deep network to teach a few deep neural networks and then to average the resulting results.

The layers that implement this method are used to prevent classifier retraining. In practice, dropout is used in conjunction with other regularizations [23, 25].

Teaching neural network to perform the tasks assigned to it consists of setting its architecture and calculating the scales for connections between individual neurons.

Therefore, after the creation of architecture, the network is unable to fulfill the tasks for which was created. The learning process can be presented in the form of a follow-up presentation to the input network of an image, a training set, and a comparison of the output results with the true result. Take two images, for example, the image of the car, and the image of the requested part, which is empty. The response of the answer is the result of the resulting error function (delta error). You want to extend to all neurons on the network.

The basic method used to educate artificial neural networks is the reverse propagation error (backpropagation) algorithm. The essence of the algorithm is that the scales of concealed neurons are altered directly proportionally to the error of neurons with which the specific neuron is associated.

Describe the architecture of the management system and traffic optimization. This architecture has been developed through various methodologies and designs, such as IDEF and UML.

The overall system was divided into several separate modules (Fig. 3): video processing module, for recognition of road users; control module of traffic light regulation; the intersection simulation module.

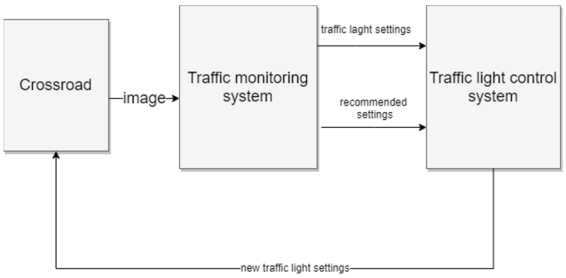


Fig. 3. System overview.

Considering the methods of object classification, convolutional neural networks were selected to create the input video processing module. Since vehicle monitoring is an actively developing area, there are a number of proposed solutions. To compare the performance of each, separate resources were created such as “The KITTI Vision Benchmark Suite” [27].

It was determined that the SSD and YOLO [21, 22] architectures are the most optimal for real-time recognition tasks for video. Determining the effectiveness of recognition by determining the mAP for all classes of objects. Mean Average Precision (mAP) is used to evaluate the accuracy of object classification tasks [28]. To calculate mAP, it is necessary to calculate the accuracy for each class based on the model's expert predictions. An object is considered classified if the area of the expert judgment and the result of the classification intersect more than 50%.

The YOLO type neural network is implemented using a Darknet framework using the OpenCV library [13, 29-31]. The network architecture consists of 24 convolutional layers and two interconnected layers. The output layer 7x7x30 is a multidimensional array. The network architecture is presented in Fig. 4.

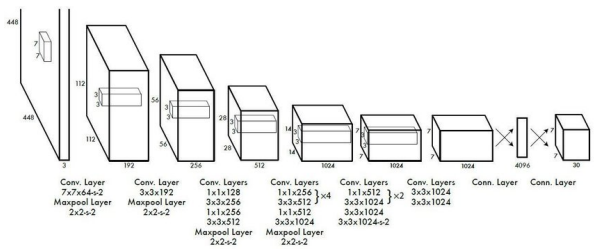


Fig. 4. The architecture of YOLO type neural network [22]

The window of video control (Fig. 5) allows you to set images from CCTV cameras at intersections that are directed in different directions.

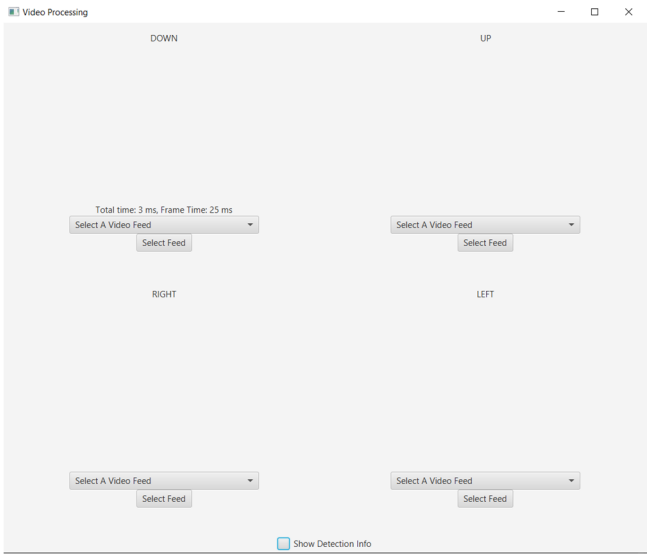


Fig. 5. Window of video stream settings.

In this window, the user can select a video stream from the drop-down list that lists the available cameras.

After all the settings in the system, the detection (recognition) of vehicles on each side of the intersection in real time will start (Fig. 6).

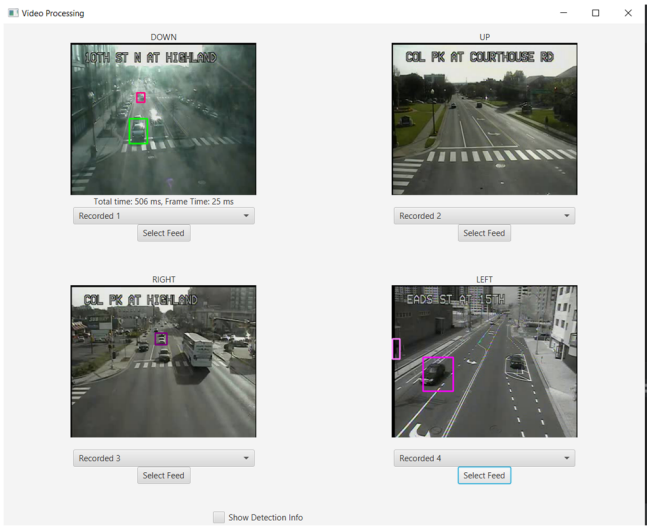


Fig. 6. Vehicle recognition in the frame.

The user is also able to view the classification details using the "Show detected info" checkbox (Fig. 7).

Fig. 7. Detailed information on the results of vehicle recognition.

A module was also developed where the result of modeling the work of a crossroads is displayed (Fig. 8).

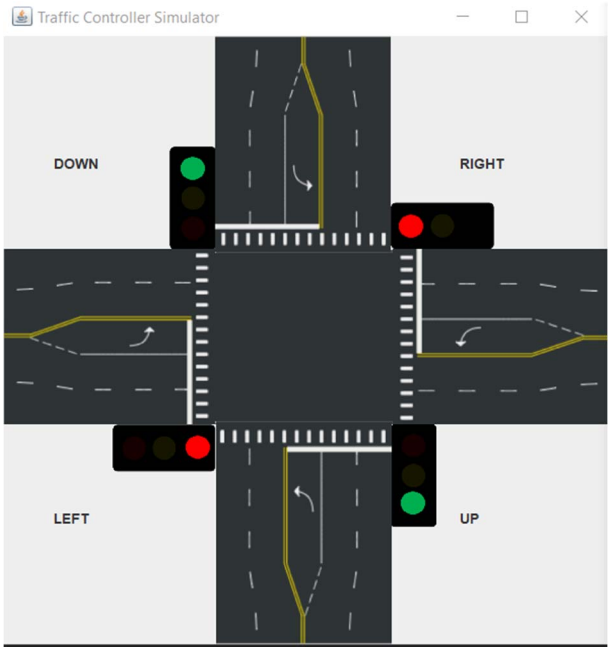


Fig. 8. Window of result of modeling the work of a crossroads.

The input is the result of image processing from the road cameras. After that, similar road circumstances are reproduced, namely, the same number of vehicles is taken as at the intersection, and the traffic light mode of operation for the given road circumstances is simulated (Fig. 9).

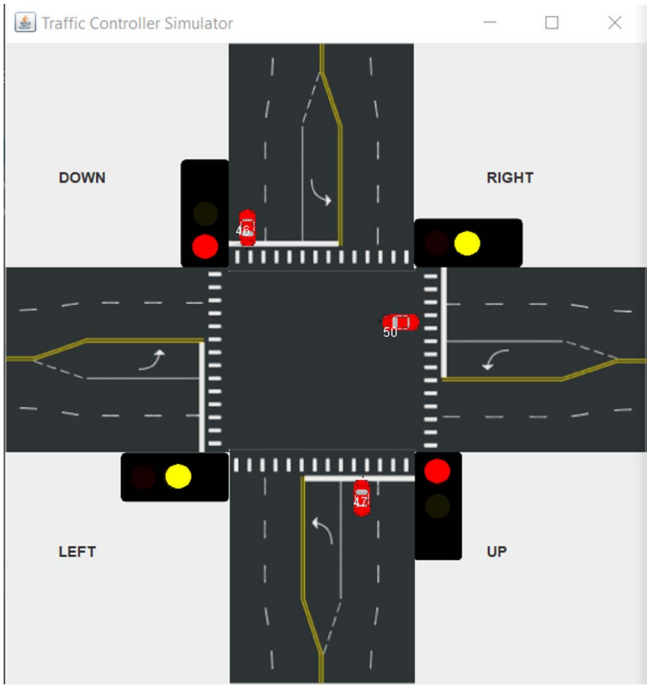


Fig. 9. Modeling of traffic light for given road conditions

The best vehicle recognition result is 94.65%, the worst result is 88.71%. This result satisfies the software requirements.

CONCLUSIONS

In this paper, it was investigated the methods and approaches used in computer vision technologies and ways of their application for traffic monitoring and optimization. Special attention was paid to modern principles and peculiarities of neural net-works used for objects recognition on video. For the task discussed in this work, it is necessary to choose the most lightweight classifier, which also will be enough accuracy and speed of recognition. During the work models of neural networks, principles of construction of layers of a convolutional neural network, their principles of work were investigated.

The analysis revealed that deep neural networks are suitable for the delivered task. Since the images, objects in which need to be classified, are large enough. Many convolutional layers can highlight the remarkable features for the problem solved and the work of such a network will be effective. In addition, it was decided to use integrated solutions such as YOLO and Faster R-CNN. Based on the analyses of recognition methods, YOLO architecture the most accurate recognition of video. As a result, the traffic management system was developed. This system allows adapting the traffic control regimes and the traffic-controlled algorithm.

REFERENCES

[1] D. Levinson and W. Chen, "Traffic management systems," in Assessing the Benefits and Costs of ITS. Transportation Research, Economics and Policy, vol. 10, D. Gillen and D. Levinson, Eds. Boston, MA: Springer, 2004, pp. 263–285.

[2] W. Krautter, D. Manstetten, and T. Schwab, "Traffic simulation for the development of traffic management systems," in Traffic and Mobility, W. Brilon, F. Huber, M. Schreckenberg, and H. Wallentowitz, Eds. Berlin, Heidelberg: Springer, 1999, pp. 193–204.

[3] X. Li and Y. Shi, "Computer vision imaging based on artificial intelligence," International Conference on Virtual Reality and Intelligent Systems (ICVRIS), Changsha, 2018, pp. 22–25.

[4] D. Barik and M. Mondal, "Object identification for computer vision using image segmentation," 2nd International Conference on Education Technology and Computer, Shanghai, 2010, pp. 170–172.

[5] D. Gerónimo, J. Serrat, A. M. López, and R. Baldrich, "Traffic sign recognition for computer vision project-based learning," in IEEE Transactions on Education, vol. 56, no. 3, 2013, pp. 364–371.

[6] B. Dolez and N. Vincent, "Sample selection in textured images," IEEE International Conference on Image Processing, San Antonio, TX, 2007, pp. 221–224.

[7] J. Beyerer, F. P. León, and C. Frese, Machine vision. Automated visual inspection: theory, practice and applications. Berlin, Heidelberg: Springer, 2016.

[8] B. Batchelor and F. Waltz, Intelligent machine vision. Techniques, implementations and applications. London: Springer, 2001.

[9] K. R. Chowdhary, "Machine vision," in Fundamentals of Artificial Intelligence. New Delhi: Springer, 2020, pp. 669–706.

[10] T. K. Ho, J. J. Hull, and S. N. Srihari, "On multiple classifier systems for pattern recognition," 11th IAPR International Conference on Pattern Recognition, vol. II. Conference B: Pattern Recognition Methodology and Systems, The Hague, Netherlands, 1992, pp. 84–87.

[11] M. N. Murty and V. S. Devi, Pattern recognition. An algorithmic approach. London: Springer, 2011.

[12] Y. Kondratenko and E. Gordienko, "Neural networks for adaptive control system of caterpillar turn," Annals of DAAAM for 2011 & Proceeding of the 22th Int. DAAAM Symp. "Intelligent Manufacturing and Automation", Vienna, Austria, 2011, pp. 0305–0306.

[13] I. Sidenko, K. Filina, G. Kondratenko, D. Chabanovskyi, and Y. Kondratenko, "Eye-tracking technology for the analysis of dynamic data," IEEE 9th International Conference on Dependable Systems, Services and Technologies (DESSERT), Kiev, 2018, pp. 479–484.

[14] Y. Pomanysochka, Y. Kondratenko, G. Kondratenko, and I. Sidenko, "Soft computing techniques for noise filtration in the image recognition processes," IEEE 2nd Ukraine Conference on Electrical and Computer Engineering (UKRCON), Lviv, Ukraine, 2019, pp. 1189–1195.

[15] B. Hoppenstedt, K. Kammerer, M. Reichert, M. Spiliopoulou, and R. Pryss, "Convolutional neural networks for image recognition in mixed reality using voice command labeling," in Augmented Reality, Virtual Reality, and Computer Graphics. AVR 2019. Lecture Notes in Computer Science, vol. 11614, L. De Paolis and P. Bourdot, Eds. Cham: Springer, 2019, pp. 63–70.

[16] Z. Wang, W. Lu, Y. He, N. Xiong, and J. Wei, "RE-CNN: a robust convolutional neural networks for image recognition," in Neural Information Processing. ICONIP 2018. Lecture Notes in Computer Science, vol. 11301, L. Cheng, A. Leung, and S. Ozawa, Eds. Cham: Springer, 2018, pp. 385–393.

[17] I. Okfalisa, I. Gazalba, R. Mustakim, and N. G. I. Reza, "Comparative analysis of k-nearest neighbor and modified k-nearest neighbor algorithm for data classification," 2nd International conferences on Information Technology, Information Systems and Electrical Engineering (ICITISEE), Yogyakarta, 2017, pp. 294–298.

[18] B. Wang, Yong Zeng, and Yupu Yang, "Generalized nearest neighbor rule for pattern classification," 7th World Congress on Intelligent Control and Automation, Chongqing, 2008, pp. 8465–8470.

[19] P. Bidyuk, V. Beglytsia, A. Gozhyj, and I. Kalinina, "Using the Metropolis-Hastings algorithm in Bayesian data analysis procedures," 14th IEEE International Conference on Computer Sciences and Information Technologies (CSIT), Lviv, Ukraine, 2019, pp. 98–101.

[20] I. Sidenko, G. Kondratenko, P. Kushneryk, and Y. Kondratenko, "Peculiarities of human machine interaction for synthesis of the intelligent dialogue chatbot," 10th IEEE International Conference on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications (IDAACS), Metz, France, 2019, pp. 1056–1061.

[21] Y. Nie, P. Sommella, M. O'Nils, C. Liguori, and J. Lundgren, "Automatic Detection of Melanoma with Yolo Deep Convolutional Neural Networks," E-Health and Bioengineering Conference (EHB), Iasi, Romania, 2019, pp. 1–4.

[22] R. Dey, D. Bhattacharjee, and M. Nasipuri, "Object detection in rainy condition from video using YOLO based deep learning model," in Advanced Computing and Systems for Security. Advances in Intelligent Systems and Computing, vol. 1136, R. Chaki, A. Cortesi, K. Saeed, and N. Chaki, Eds. Singapore: Springer, 2020, pp. 121–131.

[23] Y. Liu, "An improved faster R-CNN for object detection," 11th International Symposium on Computational Intelligence and Design (ISCID), Hangzhou, China, 2018, pp. 119–123.

[24] P. Kushneryk, Y. Kondratenko, and I. Sidenko, "Intelligent dialogue system based on deep learning technology," 15th International Conference on ICT in Education, Research, and Industrial Applications: PhD Symposium (ICTERI 2019: PhD Symposium), vol. 2403, Kherson, Ukraine, 2019, pp. 53–62.

[25] J. Si, S. L. Harris and E. Yfantis, "A dynamic ReLU on neural network," IEEE 13th Dallas Circuits and Systems Conference (DCAS), Dallas, TX, 2018, pp. 1–6.

[26] A. Mazumdar and A. S. Rawat, "Learning and recovery in the ReLU model," 57th Annual Allerton Conference on Communication, Control, and Computing (Allerton), Monticello, IL, USA, 2019, pp. 108–115.

[27] A. Geiger, P. Lenz and R. Urtasun, "Are we ready for autonomous driving? The KITTI vision benchmark suite," IEEE Conference on Computer Vision and Pattern Recognition, Providence, RI, 2012, pp. 3354–3361.

[28] M. Trott, "Numerical computations," in The Mathematica GuideBook for Numerics. New York: Springer, 2006, pp. 1–967.

[29] S. Gollapudi, Learn computer vision using OpenCV. With deep learning CNNs and RNNs. Berkeley, CA: Apress, 2019.

[30] F. Nelli, "Image analysis and computer vision with OpenCV," in Python Data Analytics. Berkeley, CA: Apress, 2018, pp. 507–535.

[31] X. Farhodov, O. Kwon, K. W. Kang, S. Lee, and K. Kwon, "Faster RCNN detection based OpenCV CSRT tracker using drone data," International Conference on Information Science and Communications Technologies (ICISCT), Tashkent, Uzbekistan, 2019, pp. 1–3.

IEEE Third International Conference on Data Stream Mining & Processing
August 21-25, 2020, Lviv, Ukraine

Methods for Lines and Matrices Segmentation in RNN-based Online Handwriting Mathematical Expression Recognition Systems

Oleg Yakovchuk*‡, Anastasiia Cherneha*†, Dmytro Zhelezniakov*†, Viktor Zaytsev*
* Samsung R&D Institute, 57 Lva Tolstogo str.
Kyiv, Ukraine
o.yakovchuk@samsung.com,
a.cherneha@samsung.com,
d.zheleznyak@samsung.com,
v.zaytsev@samsung.com
† Taras Shevchenko National University of Kyiv
Kyiv, Ukraine
‡ National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute"
Kyiv, Ukraine

***Abstract*—Modern applications for handwritten mathematical expressions recognition are not limited to single expression input and rather provide the possibility to input several expressions at once, complex elements such as equation systems and matrices, and also support edit modifications. In this paper, we examine segmentation of mathematical expressions and present new techniques: multi-line segmentation method using special Dynamic Threshold Distance and matrix segmentation method based on projection profiling modification. The proposed methods are used for the preliminary structure analysis and utilize only geometrical features of the input strokes, therefore they can be used at the first stage of the recognition workflow. By incorporating these methods into our recognition system we make it possible to recognize multi-line expressions and matrices. The evaluation of the segmentation itself and overall recognition accuracy is performed using open benchmark CROHME datasets and in-house datasets and for matrices it demonstrates segmentation success rate 93.66% on CROHME2016. The proposed methods can be applied for other applications such as analysis of handwritten document layout, tables and charts recognition.**

***Index Terms*—handwritten mathematical expressions recognition, mathematical expression segmentation, line segmentation, matrix detection, matrix segmentation**

I. Introduction

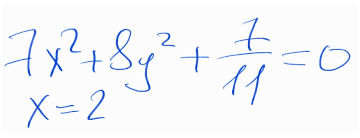
Throughout history, people have developed many specialized languages for communicating with each other in a particular field. Mathematical expression, musical notes, diagrams, electrical schemas are recognized in the world and are understood by experts in their field. Mathematical expressions (MEs) are widely used in various industries such as science, engineering, education, finance, and many others. Mathematical notation (MN) is one of the most ancient [1] languages and one of the most complex. MN has a two-dimensional structure and includes more than 1,500 characters [2]. The alphabet of MN contains many similar characters such as 'x', 'χ', 'v', 'ν', 'n', 'η', '0', 'O', '∞', '$\propto$' etc. Therefore, despite significant advances in the field of computer vision and voice recognition, the recognition of ME is still a challenging task.

"Online" and "Offline" are different subclasses of Handwritten Mathematical Expressions (HME) recognition [3]. Online recognition is a translation of dynamic representation (the traces of pen/finger movement) into the mathematical notation such as LaTeX or MathML. Offline recognition operates with a static representation of the input, such as an image. It is customary to distinguish three epochs in the development of online HME recognition systems [4]: sequential solutions, integrated solutions, and end-to-end neural network-based solutions. Sequential solutions and integrated solutions are characterized by the presence of two stages of recognition [5]: (1) symbol segmentation and classification and (2) structural analysis.

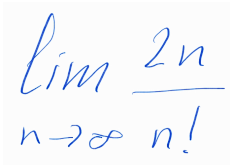
Such approaches for the symbol segmentation have been examined and published in many papers [6]–[10]. But the most research papers on HME direction consider math recognition as single-expression recognition when samples in applied datasets mostly have one line of ME. Special cases like equation systems with separate expressions in multiple lines and matrices, where every matrix cell is a separate ME itself, are represented to a much lesser extent. However, modern systems can not limit the input with a single math instance. Such systems should support any type of ME input, both as a single line expression or multiple lines of independent expressions during one input session, as well as editing modifications to these expressions.

Another important feature of new systems is the focus on mobile devices [11]–[13]. Such systems have limited sizes on the mobile screen, so users can place multiple expressions and their parts very close side by side. On the contrary, some parts of a single ME can be written very far one from another (see examples a, b on Figure 1).

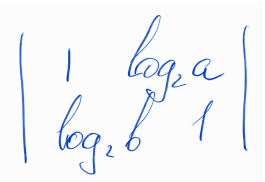
978-1-7281-3214-3/20/$31.00 ©2020 IEEE



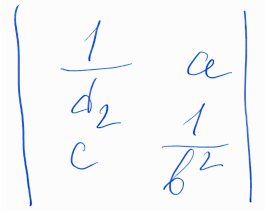
(a) Overlapping areas of two expressions



(b) Large vertical distance between elements of one expression



(c) Overlapping cell projections



(d) Uncertainty of the belonging of individual elements to a particular cell

Fig. 1: Ambiguity during segmentation.

Described requirements lead to the fact that it is necessary to separate the input strokes into different instances of formulas and blocks for the next recognition stages. This segmentation task becomes more complicated, taking into account that one ME isn't limited to a specific line's height. For example, it can have many fractions one in another, different special large symbols of operations (like sum, integral, limit) with upper and lower parts that also can contain ME inside themselves. Other ambiguities can also occur due to personal handwriting styles of different users, for example, far-away standing dot of symbols 'i', 'j' that can also be interpreted as a single-standing dot of separate instance. Figure 1 illustrates the main ambiguities during the segmentation of multi-line expressions and matrix cells. Under such conditions, many approaches [14], [15] for the segmentation of elements cease to work.

In order to solve this segmentation task and overcome described ambiguities, as well as satisfy our recognition system requirement, discussed in section III, we propose new segmentation techniques.

II. Prior Work

Approaches for matrix segmentation were discussed in a few works and targeted different problems: offline printed matrices recognition, offline handwritten matrices recognition, and online handwritten matrices recognition. Many methods perform character recognition first, and then use character type information during segmentation: [16] first forms matrix elements by inflating bounding boxes and uniting intersected characters, [14] forms matrix elements by checking distances to adjacent characters, [17] combines both actions. In [15] adjacent characters are united based on the average character width. They also experiment with clustering elements centroids projected on an axis. In [18] a minimum spanning tree is used and segmentation is first performed by rows, then rows are divided to elements by spaces. The reason we can not use

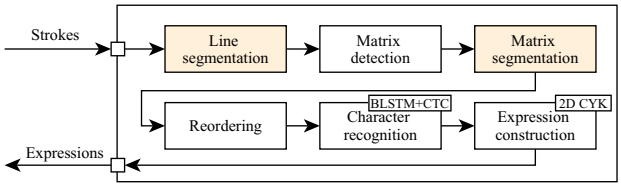


Fig. 2: Recognition workflow.

most of the existing approaches in our recognition system is explained in Section III.

Our solution for matrices is based on projection profiling. Projection profiling is extensively used in document structure analysis [19], [20], specifically interesting is its usage for tabular data segmentation [21], since matrices have grid structure alike tables.

III. Overview of the Recognition System

The main modules and overall workflow are illustrated in Figure 2. Line segmentation and Matrix segmentation modules are described in Section IV. The character recognition module performs character segmentation and classification of mathematical symbols. It's based on a Bidirectional Long Short-Term Memory (BLSTM) neural network with with Connectionist Temporal Classification (CTC) [22], [23] output layer. The expression construction is implemented using Probabilistic Context-Free Grammar (PCFG) [24], [25] and is also equipped with the N-gram language model.

An LSTM model is presented by the next compact form of the equations:

$$\begin{aligned} f_t &= \sigma_g(W_f x_t + U_f h_{t-1} + b_f) \\ i_t &= \sigma_g(W_i x_t + U_i h_{t-1} + b_i) \\ o_t &= \sigma_g(W_o x_t + U_o h_{t-1} + b_o) \\ \tilde{c}_t &= \sigma_h(W_c x_t + U_c h_{t-1} + b_c) \\ c_t &= f_t \circ c_{t-1} + i_t \circ \tilde{c}_t \\ h_t &= o_t \circ \sigma_h(c_t)\,, \end{aligned} \tag{1}$$

where x_t – input vector at the time stamp t; f_t – forget activation vector; i_t – input activation vector; o_t – output activation vector; h_t – hidden state vector; $\tilde{c}_t$ – cell input activation vector; c_t – cell state vectora; W – weight matrices; b – bias vectors; σ_g – logistic sigmoid; σ_h – hyperbolic tangent function. The input of BLSTM is a normalized vector of features: Δx, Δy, and pen-up/pen-down.

CTC is associated as a scoring function and can be presented by the next formula:

$$O(S) = -\ln\left(\prod_{(x,z)\in S} p(z|x)\right) = -\sum_{(x,z)\in S} \ln p(z|x)\,, \tag{2}$$

where x – the input sequence and z – the labels from the ground truth.

The approach based on the BLSTM in connection with CTC tends to find the symbols bounds at the corresponding

input sequence. However, this approach requires very strict observance of the characters order, which has to correspond to the order during the training process. The following order was chosen for the learning of character recognition model: from top to bottom and from left to right. The strict strokes order in the input sequence allows the recurrent neural network to provide a more accurate classification of characters based on context. And on the other hand, a sequence of input strokes that do not correspond to the required order leads to a very significant decrease in the quality of character segmentation and classification.

Considering there aren't any strict rules of symbols writing order in math, users are able to input strokes in arbitrary order, and even alternately in separate expressions. Segmentation by expression instances and by matrix cells is designed to ensure the required strokes order for BSLTM input, and also provides a preparatory analysis of the spatial relations structure for the expression construction module.

IV. Proposed methods for Segmentation

In this section, we will describe the first and third steps of our system recognition workflow - line segmentation and matrix segmentation. At this stage, we don't have any typographic information of the characters (ascent, descent, mean line, etc.) and can utilize only geometry features of input strokes.

To deal with both multi-line and matrix expressions that have various layouts with their specific features, we will use different segmentation approaches.

A. Multi-line segmentation

Multi-line segmentation is used to split all input sample strokes into separate segments for every math instance. First, we receive a sequence of input strokes and perform common preprocessing for our segmentation approaches - calculate strokes bounding boxes and save the order index of every stroke. The purpose of the last step is the necessity to save the initial strokes' order inside every segment for the next recognition stages.

In order to overcome ambiguous situations described previously we use a special Dynamic Threshold Distance (DTD) to distinguish far-written parts of one expression from single standing MEs. We have performed empirical measurements of users' preferable average line spacing for writing MEs. Accordingly to our observations, DTD is also strongly associated with user interface preferences, specifically with the font size, that is related to the device screen resolution. So we calculate DTD taking into account the median stroke of the whole sequence with the next formula:

$$DTD = C_{ls} * \frac{fontSize + H_m}{2} * ppi\,, \tag{3}$$

where:

C_{ls} - line spacing empirical constant (chosen as 0.42),
H_m - height of the strokes sequence median stroke,
ppi - pixels per inch value of the device screen.

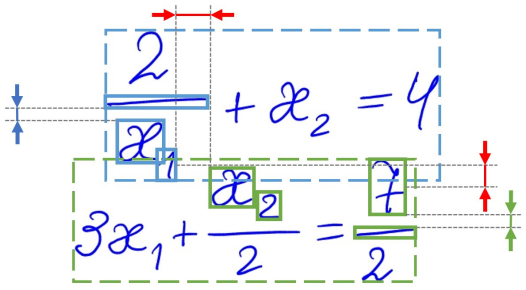
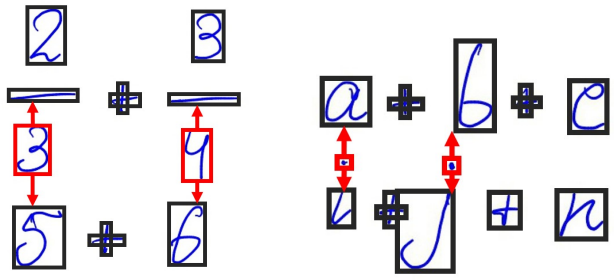


Fig. 3: Successful multi-line segmentation of basic geometry method. Some critical distances between potentially neighboring strokes are greater than DTD, so these strokes belong to different segments. Such values are marked with red arrows.



(a) Large vertical distance (b) Far-away standing parts of symbols between fraction elements

Fig. 4: Examples that require to find special cases during multi-line segmentation.

The backbone of the segmentation method is based on the geometry relations of strokes bounding boxes. Its pseudo-code is presented as Algorithm 1. There we consider two strokes are in one line when either one of their centers hits in another stroke bounding box, but resized to 50% of its height. And we confirm the vertical relation of two strokes if they can have subscript/superscript or upper/lower relative positioning. Merging process includes insertion of the stroke index to other segment strokes indexes to preserve the initial strokes order, and calculation of the union of their bounding boxes.

We can use the described algorithm as a basic segmentation method itself. There is an example of its usage in the Figure 3. This basic geometry method is doing well with simple multi-line expressions and with instances that have a low amount of horizontal intersections. But there are more ambiguous cases, like ones on the Figure 4, where some strokes have similar distances to multiple segments. Such special cases are mostly related to vertical-ordered blocks and symbols, one of the most frequent of them is a fraction. Other specific symbols are common math characters 'i' and 'j' with their upper dots. And considering special geometry features of such symbols we can try to detect them during this stage to analyze preliminary expression structure, that can help to solve this issue.

The main improvement of this method is a special strokes detector that finds cases of interest - fractions and small

Algorithm 1 Multi-line segmentation method

```
function LineSegmentation(S)
    Input: S - list of input strokes
    Output: R - list of segments (strokes lists)
    R[0].insert(S[0])
    for s in S do
        M.clear()        ▷ M - segments for current stroke s
        for seg in R do
            for s_seg in seg do
                if isOneLine(s, s_seg) then
                    M.insert(seg)
                else if isVerticalRelation(s, s_seg)) then
                    if distance(s, s_seg) < DTD then
                        M.insert(seg)
                    end if
                end if
            end for
        end for
        if M.empty() then
            R.insert(s)                ▷ create new segment
        else
            MergeToSegments(s, M)
        end if
    end for
    return R
end function
```

symbols like dots and commas. We perform this detection step for every stroke of the input sequence. If stroke sizes in comparison to the sequence median stroke satisfy some conditions, we then perform stroke approximation using the Ramer–Douglas–Peucker algorithm and calculate stroke angles deviations by all approximated stroke points. Basing on the acquired features we use a decision tree algorithm to detect fraction-like symbols and dots/commas-like symbols. Using the information about detected special cases we make an assumption about possible expression structure and use it in the main segmentation stage. Now the appropriate segments are searched more effectively by giving more attention to the bottom direction for dots-like strokes, and by increasing vertical DTD value for the bounding boxes of fractions-like strokes.

By applying special cases detection we can overcome more complex segmentation ambiguities, still without getting any context information about ME symbols.

B. Matrix segmentation

Matrices recognition begins with matrices detection performed by a special module (second module in Figure 2). It detects matrices by finding round, square, and straight-line brackets using similar algorithms from the special cases detector described in the previous part. When this step is completed, we have a collection of strokes that are assumed to be parts of a matrix. The goal of segmentation is to unite them in matrix elements and reconstruct matrix rows and columns

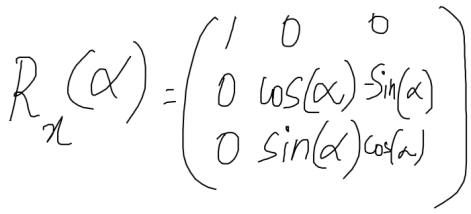


Fig. 5: Matrix, where structure is better expressed in the first row and the first column.

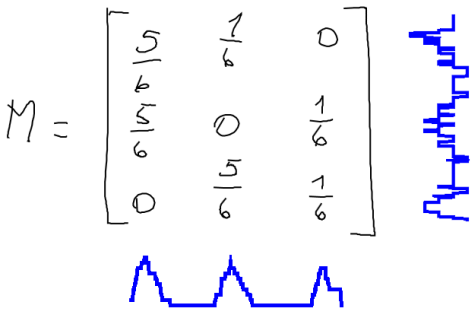


Fig. 6: Matrix, where columns are more clearly separated than rows.

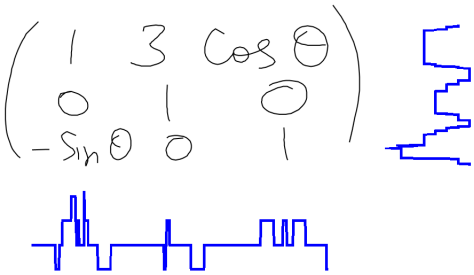


Fig. 7: Matrix, where rows are more clearly separated than columns.

structure. Handwritten matrices have some specific problems: uneven spacing and writing, insufficient spacing, lines going up or going down. Often, matrix structure is guessed from certain row or column, written best or having better spacing. For example, in the Fig. 5 matrix structure is much better seen in the first row and the first column. This observation is used in the proposed algorithm by choosing a certain row or column as a segmentation template.

The basic idea of the algorithm is taken from projection profiling method. First, bounding boxes of each stroke are projected on X and Y axis simultaneously. Because of matrices grid structure, projections on axis are seemingly more intensive next to rows and columns. To extract segments of higher concentration, an intensity threshold is chosen based on maximum projection concentrations: the more strokes are projected to the same axis point, the higher the threshold is.

Using the extracted segments, we make an assumption about which axis projection is more clear. Here we choose projection

TABLE I: Multi-line segmentation results

	Basic geometry method		Special cases detection method	
	Segmentation succ. rate (%)	Recog. rate (%)	Segmentation succ. rate (%)	Recog. rate (%)
Single line (100 samples)	100	84	**96**	82
Multi-line (100 samples)	42	33	**75**	61
Total	71	58.5	**85.5**	71.5

with more similar segments length by calculating segments length variance. Then, accordingly to the assumption, it is decided to separate coordinate ranges either for rows or for columns. In the Figure 6 is an example of a matrix where columns are better separated than rows, so the matrix will be first separated by columns. In the Figure 7 we can see the opposite situation: the matrix will be first divided into rows.

Now we temporarily assign strokes to nearest row (column). Then one row or column is chosen to deduce a number N of element in all other rows or columns. One of the ways we tried to chose one row (column) is finding the one that contains least amount of strokes. Another way is to pick the one containing longest space. Then all other rows (columns) are segmented based on the deduced quantity of elements in each. This can be done by separating row (column) strokes by $N-1$ largest spaces.

V. Evaluation

By applying segmentation methods we have to consider their efficiency and influence on whole recognition system, so we measure both segmentation accuracy and overall recognition accuracy.

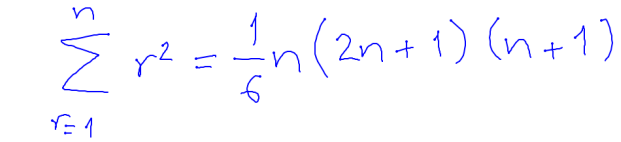
A. Results

1) Multi-line segmentation evaluation: There are no well-known open datasets for the evaluation of HME lines segmentation so for this task we created a small in-house dataset by sampling multi-line instances of different types. We manually provided ground-truth as a relation of every stroke to specific expression. Also, single line samples from CROHME 2014 and CROHME 2016 open datasets were added. As a result, the dataset for evaluating line segmentation methods contains 200 samples - 100 multi-line samples with more than 200 MEs and 100 single line samples. The successful segmentation of multi-line sample is interpreted as a correct appropriation of every stoke to its segment. We use single line expressions for segmentation methods evaluation to be sure that the whole recognition system would not get worse with 'traditional' samples. And we can see one of some rare cases of wrong single line sample segmentation in Figure 8a. Table I shows multi-line segmentation results.

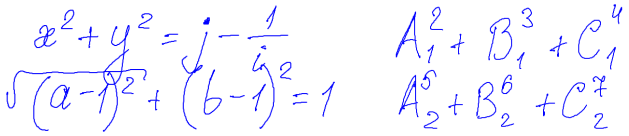
2) Matrix segmentation evaluation: No specific metrics were earlier presented for matrix segmentation quality evaluation, and rates for total recognition performance are often presented when evaluating matrix segmentation. So, let us specify

TABLE II: Matrix segmentation results

	Our method CROHME 2014	Our method CROHME 2016	MyScript CROHME 2016	Wiris CROHME 2016
Cell Recall Rate	90.51	94.37	87.49	84.68
Row Recall Rate	91.53	94.82	95.61	87.16
Column Recall Rate	89.16	95.30	90.71	82.22
Matrix Recall Rate	97.71	98.07	97.52	85.67
Segmentation Success Rate	**86.28**	**93.66**		



(a) Wrong segmentation of single line sample - two segments found



(b) Wrong segmentation of multi-line sample - one segment found

(c) Correct segmentation of multi-line sample

Fig. 8: Multi-line segmentation examples.

matrix segmentation success rate on a dataset containing N matrices $\{m_1, m_2, ..., m_N\}$ as:

$$Rate = \frac{\sum_{i=1}^{N} f(m_i)}{N}, \tag{4}$$

$$f(m) = \begin{cases} 1, \text{each stroke is in proper matrix cell} \\ 0, \text{otherwise} \end{cases}$$

Matrix segmentation is evaluated on CROHME 2014 [26] matrix dataset and CROHME 2016 [27] matrix dataset. In samples that contain more than one matrix, we consider each matrix separately. Cell, row, column and matrix recall rates are calculated on strokes segmentation level by LgEval tool used in CROHME competitions [28]. The results and comparison are presented in Table II. A few examples of correctly segmented matrices are shown in the Figure 9.

B. Limitations and Future work

Through our explorations, we have found that the proposed approaches for segmentation have limitations. Using only geometry information of the strokes, our multi-line segmentation method isn't able to resolve complex uncertain cases with vertical-order symbols (like sigma, integral, limit) and close-standing MEs with a big amount of strokes intersections.

One of the difficulties for matrix segmentation is finding "valleys" in projections when bounding boxes of different rows or columns overlap too much (Figure 10a). Another situation

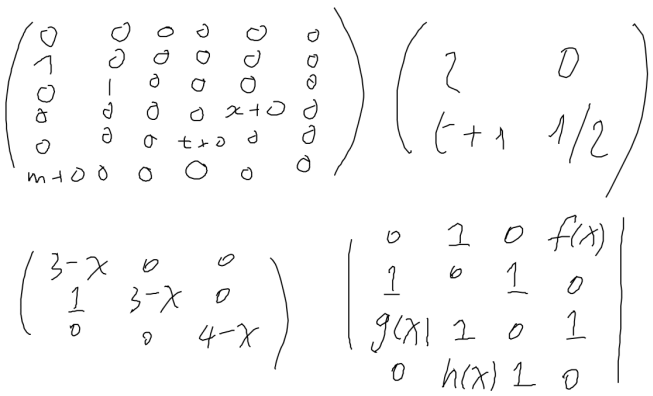
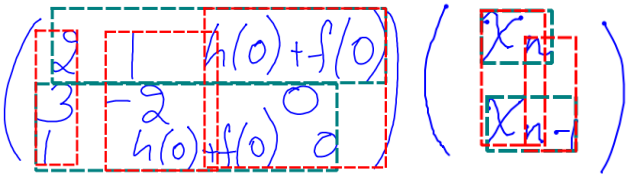


Fig. 9: Examples of correctly segmented matrices.



(a) Wrong segmentation due to rows bounding boxes overlap

(b) Wrong segmentation in $n \times 1$ sized matrix

Fig. 10: Examples of matrices segmented with errors.

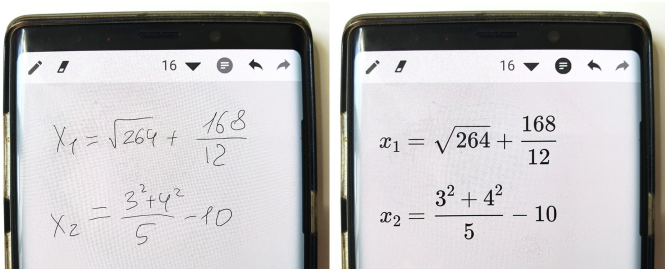
where a mistake can occur is when some rows or columns of the matrix contain many strokes, and others contain just a few. Such a projection segment may be mistakenly treated as "valley". The method also needs special improvements to deal with matrices containing rows with only one element in each. That is because small "valleys" that in most cases are not real separations between rows or columns are eliminated based on the length of the maximum valley. But in $n \times 1$ matrices the maximum space cannot be used. A simple example of this problem is in Figure 10b.

So, due to the fact some multi-line samples and matrices can't be segmented without semantic information, we are going to improve our special cases detection to have more context knowledge on the segmentation stage. It can be possibly done using some basic symbol classification before the segmentation, more advanced analysis of spatial relations, and also with grammatical analysis.

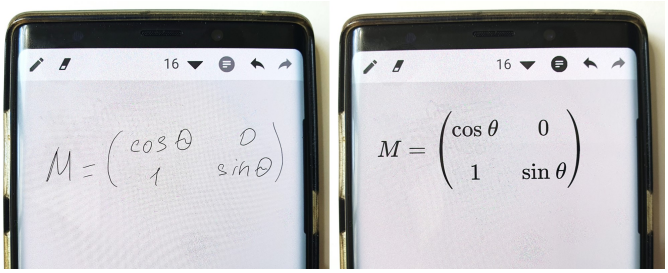
VI. Conclusion

In this work, we demonstrated different segmentation tasks in HME recognition, described the related issues, and proposed methods for multi-line segmentation of ME and matrices elements.

The proposed methods can be used on the first stage of HME recognition workflow and are focused on the analysis of ME structure, which ensures the correct preprocessing of input strokes to take into account the requirements of the following recognition steps. Such structural analysis also can significantly reduce the execution time on the final stage of recognition of ME. By incorporating these methods into our



(a) Multi-line segmentation



(b) Matrix detection and segmentation

Fig. 11: Incorporation of proposed methods to the developed recognition system. User's input on the left and the results of segmentation and recognition on the right.

recognition system [29] we made it possible to recognize multi-line expressions and matrices (Figure 11). We achieved high segmentation accuracy, dealing with most cases of average complexity.

Besides the recognition of ME, the explained methods can potentially be applied for other applications with handwriting input tasks such as analysis of handwritten document layout, table recognition, charts, and other complex document elements. We also plan to evaluate these methods for equation systems recognition.

Another important direction for further research is more deep clarification of the ME structure based on the results of both character segmentation and classification, which, according to our observations, should significantly improve the quality of the proposed approaches.

References

[1] F. Cajori, *A history of mathematical notations*. Courier Corporation, 1993, vol. 1.

[2] B. Beeton, A. Freytag, and M. Sargent III, "Unicode support for mathematics," The Unicode Consortium, Tech. Rep., 2017.

[3] D. Blostein and A. Grbavec, "Recognition of mathematical notation," in *Handbook of character recognition and document image analysis*. World Scientific, 1997, pp. 557–582.

[4] T. Zhang, "New architectures for handwritten mathematical expressions recognition," Ph.D. dissertation, 2017.

[5] K.-F. Chan and D.-Y. Yeung, "Mathematical expression recognition: a survey," *International Journal on Document Analysis and Recognition*, vol. 3, no. 1, pp. 3–15, 2000.

[6] M. Okamoto, "Recognition of mathematical expressions by using the layout structure of symbols," in *Proc. 1st Int. Conf. Document Analysis and Recognition, 1991*, 1991, pp. 242–250.

[7] H.-J. Winkler, H. Fahrner, and M. Lang, "A soft-decision approach for structural analysis of handwritten mathematical expressions," in *1995 International Conference on Acoustics, Speech, and Signal Processing*, vol. 4. IEEE, 1995, pp. 2459–2462.

[8] A. Kosmala and G. Rigoll, "On-line handwritten formula recognition using statistical methods," in *Proceedings. Fourteenth International Conference on Pattern Recognition (Cat. No. 98EX170)*, vol. 2. IEEE, 1998, pp. 1306–1308.
[9] N. E. Matsakis, "Recognition of handwritten mathematical expressions," Ph.D. dissertation, Massachusetts Institute of Technology, 1999.
[10] V. Volkova, I. Deriuga, V. Osadchyi, and O. Radyvonenko, "Improvement of character segmentation using recurrent neural networks and dynamic programming," in *2018 IEEE Second International Conference on Data Stream Mining & Processing (DSMP)*. IEEE, 2018, pp. 218–222.
[11] MyScript, "Nebo," "https://www.nebo.app/", 2019, accessed: 2019-12-14.
[12] Microsoft, "Microsoft Math Solver," "https://math.microsoft.com/", 2019, accessed: 2019-12-08.
[13] Samsung, "Samsung S Note," "https://play.google.com/store/apps/details?id=com.samsung.android.snote", 2019, accessed: 2019-12-08.
[14] K. Toshihiro and S. Masakazu, "A recognition method of matrices by using variable block pattern elements generating rectangular area," in *International Workshop on Graphics Recognition*. Springer, 2001, pp. 320–329.
[15] D. Tausky, G. Labahn, E. Lank, and M. Marzouk, "Managing ambiguity in mathematical matrices," in *Proceedings of the 4th Eurographics workshop on Sketch-based interfaces and modeling*, 2007, pp. 115–122.
[16] C. Li, R. Zeleznik, T. Miller, and J. J. LaViola, "Online recognition of handwritten mathematical expressions with support for matrices," in *2008 19th International Conference on Pattern Recognition*, 2008, pp. 1–4.
[17] T. Kanahori and M. Suzuki, "Detection of matrices and segmentation of matrix elements in scanned images of scientific documents," vol. 2003, 09 2003, pp. 433 – 437 vol.1.
[18] E. Tapia and R. Rojas, "Recognition of on-line handwritten mathematical expressions using a minimum spanning tree construction and symbol dominance," 07 2003, pp. 329–340.
[19] J. Ha, R. Haralick, and I. Phillips, "Document page decomposition by the bounding-box project." 01 1995, pp. 1119–1122.
[20] R. Ptak, B. Żygadło, and O. Unold, "Projection–based text line segmentation with a variable threshold," *International Journal of Applied Mathematics and Computer Science*, vol. 27, 03 2017.
[21] S. Chandran and R. Kasturi, "Structural recognition of tabulated data," in *Proceedings of 2nd International Conference on Document Analysis and Recognition (ICDAR '93)*. Los Alamitos, CA, USA: IEEE Computer Society, oct 1993, pp. 516,517,518,519. [Online]. Available: https://doi.ieeecomputersociety.org/10.1109/ICDAR.1993.395683
[22] M. Liwicki, A. Graves, S. Fernàndez, H. Bunke, and J. Schmidhuber, "A novel approach to on-line handwriting recognition based on bidirectional long short-term memory networks," in *Proceedings of the 9th International Conference on Document Analysis and Recognition, ICDAR 2007*, 2007.
[23] A. Graves, "Connectionist temporal classification," in *Supervised Sequence Labelling with Recurrent Neural Networks*. Springer, 2012, pp. 61–93.
[24] F. Á. Muñoz, "Mathematical expression recognition based on probabilistic grammars," Ph.D. dissertation, Universitat Politécnica de Valéncia, 2015. [Online]. Available: https://pdfs.semanticscholar.org/5605/a2d4196c8388a0d4bbf03cfbb3f6d847cee4.pdf
[25] D. Zhelezniakov, V. Zaytsev, and O. Radyvonenko, "Acceleration of online recognition of 2d sequences using deep bidirectional lstm and dynamic programming," in *International Work-Conference on Artificial Neural Networks*. Springer, 2019, pp. 438–449.
[26] H. Mouchère, C. Viard-Gaudin, R. Zanibbi, and U. Garain, "Icfhr 2014 competition on recognition of on-line handwritten mathematical expressions (crohme 2014)," vol. 2014, 09 2014.
[27] H. Mouchère, C. Viard-Gaudin, R. Zanibbi, and U. Garain, "Icfhr2016 crohme: Competition on recognition of online handwritten mathematical expressions," in *2016 15th International Conference on Frontiers in Handwriting Recognition (ICFHR)*. IEEE, 2016, pp. 607–612.
[28] Document and Pattern Recognition Lab, "LgEval Evaluation Tool," "https://www.cs.rit.edu/ dprl/Software.html".
[29] D. Zhelezniakov, A. Cherneha, V. Zaytsev, T. Ignatova, O. Radyvonenko, and O. Yakovchuk, "Evaluating new requirements to pen-centric intelligent user interface based on end-to-end mathematical expressions recognition," in *Proceedings of the 25th International Conference on Intelligent User Interfaces*. ACM, 2020, pp. 212–220.

IEEE Third International Conference on Data Stream Mining & Processing
August 21-25, 2020, Lviv, Ukraine

Recognition of High Dimensional Multi-Sensor Remote Sensing Data of Various Spatial Resolution

Volodymyr Hnatushenko
Dep. of Information Systems and Technologies
Dnipro University of Technology
Dnipro, Ukraine
vvgnat@ukr.net

Viktoriia Hnatushenko
Dep. of Information Technologies and Systems
National Metallurgical Academy of Ukraine
Dnipro, Ukraine
trip@ukr.net

***Abstract*— High-dimensional multi-sensor image data open new possibilities in remote sensing image analysis. One of the main problems in the field of automated high-dimensional data processing is interpretation and recognition of geometric forms in images. The use of traditional moment invariants in object recognition is limited to simple geometric transforms. In this study, a novel recognition method of high-dimensional multi-sensor remote sensing data of various spatial resolution was proposed. The proposed method of photogrammetric image recognition represented in raster formats of computer graphics enables identification of objects without implementation of complex algorithms involving transformation of digital images. This, in turn, enables improvement of the recognition accuracy. The experiments have shown that the proposed algorithm provides a useful alternative approach to feature reduction in high-dimensional digital image data.**

Keywords*— high-dimensional data, recognition, remote sensing image, invariant, resolution

I. INTRODUCTION

One of the main problems in the field of automated high-dimensional data processing is interpretation and recognition of geometric forms in images. Object recognition from remote sensing acquired data has been an important research topic in computer vision for many years [1-3]. Some useful applications of this subject include map creation, urban city planning, and land use analysis [4, 5]. For instance, building extraction and delineation is one of the most important tasks in map feature extraction. Obviously, the automation of such tasks can lead to greater productivity, resulting in reduced timelines for map production [6]. Its significance is determined by the fact that visualization of physical fields of material objects is one of the main ways of studying their characteristics. Since the geometric form of the image of an object depends on the positional conditions of the image acquisition, the main requirement to the methods of automated image analysis lies in the invariability to geometric transformations stemming from the variations of those conditions. We shall concentrate on processing images obtained with the methods of remote sensing of Earth (RSE). Recognizing structures is difficult in remote sensing imagery. For these high-dimensional data, a typical representation is raster formats, where each pixel of the image represents a particular elementary area on the surface of Earth; and the value of this pixel corresponds to the amount of energy reflected or radiated from this area [7]. In many cases, it is necessary to process images of the same territory collected from various sources. But spatial resolution of satellite and aerial images may vary from several kilometers to centimeters. For comparison of individual images pixelwise, the pixel meshes of the compared images should be in conformity. For transformation of individual images into a common coordinate mesh, rectification methods are used. Even though some rectification algorithms are rather efficient, in this operation, certain spectral information is nevertheless lost. Therefore, it is important to develop such recognition algorithms that do not require transformation of the analyzed images.

II. STATEMENT OF THE RESEARCH PROBLEM

Despite a tremendous effort and huge number of published papers, there are still open problems to be resolved. The use of traditional moment invariants in object recognition is limited to simple geometric transforms, such as rotation, scaling and affine transformation of the image [8]. Several papers studied recognitive and reconstruction aspects, noise tolerance, discretization errors and other numerical properties of various kinds of moment invariants [9, 10]. In [7, 12], multidimensional information-geometric models of projective nature images were developed that are invariant as for their affine transformations. In the framework of these models, an image is represented as a countable set of normed semi-invariants of paired order of the function of brightness (or indicator functions) that forms a linear space with a pseudo-Euclidean metrics. However, these models do not take into account the specific of morphogenesis of satellite scanner images, in particular, their multicomponent nature. Other authors propose algorithms of image processing (classification, rectification, referencing of images etc.) that change either the geometry of the image as a whole or the values of some of the pixels [13-17].

The aim of this work is development of a new method of high dimensional multi-sensor remote sensing data recognition that would enable identification of objects without loss of spectral information.

III. THE PROPOSED METHODOLOGY

Of a substantial importance in modern ways of formation and reproduction of visual information are multicomponent images that are considered here in the framework of the following definition.

Definition 1. A multicomponent image is such that is composed of a finite set of fragments with their own brightness functions, each of the former having information value in interpretation of the image as a whole.

978-1-7281-3214-3/20/$31.00 ©2020 IEEE

Thus, raster images can be considered as multicomponent, the individual components of which are pixels (or groups thereof) with the same level of brightness. To the images of this type belong those acquired with image sensors in the form of discrete sensor surfaces, in the capacity of which matrices of charge-coupled devices are widely used. Each element of such a matrix registers an individual component of an image as a pixel (Fig.1).

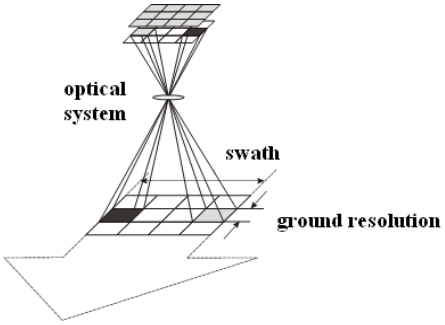


Fig. 1. Formation of remote sensing high-dimensional data

The indicator function of a composed image that includes M fragments (further, M-image), in accordance with its definition, is represented by the sum of the indicator functions of its components:

$$f_M(\mathbf{r}) = \sum_{\alpha=1}^{M} f_0\left(\hat{T}_\alpha \mathbf{r}\right), \quad (1)$$

where $\hat{T}_\alpha$ is the operator of geometric transformation that formed the α -th fragment of the image, where the fragments do not overlap pairwise. Assuming a square being a forming image, in the morphogenesis of individual fragments of such images, only homothetic transformation and parallel translation participate:

$$\hat{T}_\alpha \mathbf{r} = \mathbf{A}_\alpha\left(\mathbf{r} - \mathbf{b}_\alpha\right), \quad (2)$$

where $\mathbf{A}_\alpha$ is the homothety matrix with k_α coefficients and $\mathbf{b}_\alpha$ is the translation vector for the α fragment.

Thus, the geometric form of a multicomponent images is composed of the aggregate of geometric forms of individual fragments and of location of the set of points with radius-vectors $\mathbf{b}_\alpha, \alpha = \overline{1,M}$ on the surface of the image.

The indicator function of the mentioned discrete set of points can be represented as

$$f_p(\mathbf{r}) = \sum_{\alpha=1}^{M} \delta\left(\mathbf{r} - \mathbf{b}_\alpha\right), \quad (3)$$

where $\delta(\mathbf{r})$ is the Dirac delta function.

It can be shown that the relation between the Descartes moments of an M-image $(\boldsymbol{M}_{nm})$ and of its individual fragments $M_{nm}^{(\alpha)}$ looks as:

$$\boldsymbol{M}_{nm} = \sum_{\alpha=1}^{M}\sum_{j=0}^{n}\sum_{l=0}^{m} \frac{1}{k_\alpha^2} C_n^j C_m^l M_{nm}^{(\alpha)} \left(b_1^{(\alpha)}\right)^n \left(b_2^{(\alpha)}\right)^m \quad (4)$$

where $C_n^j = \frac{n!}{j!(n-j)!}$ are binomial coefficients.

On the basis of this relation (4), one can obtain semi-invariants of the indicator function of the M-image, which are the basis of the affine-invariant model of image interpretation [7]. Consider a particular case of morphogenesis of M-images on the basis of homotheties with the same k coefficients for all fragments, which is typical for images of raster formats of computer graphics. In a particular case of the second-order model $(n+m=2)$ we have

$$S_{02} = \frac{1}{M\sigma_0}\left[s_{02} + k^2\left\langle\left(b_2 - \langle b_2\rangle\right)^2\right\rangle\right], \quad (5)$$

$$S_{11} = \frac{1}{M\sigma_0}\left[s_{11} + k^2\left\langle\left(b_1 - \langle b_1\rangle\right)\left(b_2 - \langle b_2\rangle\right)\right\rangle\right], \quad (6)$$

$$S_{20} = \frac{1}{M\sigma_0}\left[s_{20} + k^2\left\langle\left(b_1 - \langle b_1\rangle\right)^2\right\rangle\right], \quad (7)$$

where s_{nm} are semi-invariants of a pixel; S_{nm} are semi-invariants of the raster image and σ_0 is the area of the forming image $(\sigma_0 = M_{00})$ [7]:

$$S_{nm} = \frac{1}{j^{n+m}} \frac{\partial^{n+m}}{\partial u_1^n \partial u_2^m} \ln F(\mathbf{u})\Big|_{\mathbf{u}=0}, \quad (8)$$

where $j = \sqrt{-1}$ is imaginary unit; $F(\mathbf{u})$ is integral Fourier transform of image brightness with parameter vector $\mathbf{u} = (u_1, u_2)$.

The semi-invariant is normalized by the rule:

$$Z_k = S_{k-1,N+1-k} \exp\left(-\frac{N \cdot S_{00}}{2}\right), \; k = \overline{1, N+1}. \quad (9)$$

In (5)-(7), the angle brackets indicate the operation of averaging of the corresponding factors over the set of fragments.

The connection between zero-order semi-invariants for a raster image and an individual pixel is established as:

$$S_{00} = \ln s_{00} + k^{-2} \ln M. \quad (10)$$

Basing on statements (5)-(10), we can formulate components of representation of the M-image vector in space $\mathbf{R}_1^3$ according to its definition above:

$$Z_1 = \frac{z_1}{M} + \frac{k^2}{M\sigma_0}\left\langle\left(b_2 - \langle b_2\rangle\right)^2\right\rangle, \quad (11)$$

$$Z_2 = \frac{z_2}{M} + \frac{k^2}{M\sigma_0}\left\langle\left(b_1 - \langle b_1\rangle\right)\left(b_2 - \langle b_2\rangle\right)\right\rangle, \quad (12)$$

$$Z_3 = \frac{z_3}{M} + \frac{k^2}{M\sigma_0}\left\langle\left(b_1 - \langle b_1\rangle\right)^2\right\rangle, \quad (13)$$

where $\mathbf{Z}, \mathbf{z}$ are feature vectors of the raster and forming images in the model space $\mathbf{R}_1^3$.

As follows from the results of [7], the equation of the quadrature of the representation of a multicomponent image in space $\mathbf{R}_1^3$ looks as follows:

$$M^2 \cdot (\mathbf{Z}-\mathbf{u})^T \mathbf{G}(\mathbf{Z}-\mathbf{u}) = F_2^2, \qquad (14)$$

where $\mathbf{G}$ is the matrix of the metric tensor of space $\mathbf{R_1^3}$. In equation (14), $|F_2|$ designates the radius of the affine-invariant pseudosphere, on which the point of the representation of the form of the forming image is located; $\mathbf{u}$ is the feature vector of the point set of locations of the fragments on the image plane. Considering the definition of the metric tensor of space $\mathbf{R_1^3}$, equation (13) can be written in terms of radiuses of affine-invariant pseudospheres that represent in it the geometric forms of the forming and raster images and the set of points (3). After transition to the coordinate reference related to the main directions of the mentioned pseudospheres, the equation looks as follows:

$$F_2^2 = M^2 (\mathbf{Y}-\mathbf{v})^T \tilde{\mathbf{G}} (\mathbf{Y}-\mathbf{v}), \qquad (15)$$

where $\tilde{\mathbf{G}} = \begin{pmatrix} -1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$;

$\mathbf{y}, \mathbf{Y}, \mathbf{v}$ are, respectively, feature vectors of the forming and M images and the set of points $\{b_j : j = \overline{1,2}\}$ with components

$$v_i = \left\langle (b_1 - \langle b_1 \rangle)^{i-1} (b_2 - \langle b_2 \rangle)^{3-i} \right\rangle, i = \overline{1,3}. \qquad (16)$$

In the general case, components of vector $\mathbf{v}$ averaged over the set of fragments of the multicomponent image depend on the number M of these fragments. Let us take into consideration that, in practical situations of photogrammetric images of raster formats, the number of these fragments is large, which gives the ground to use results of the probability theory as for statistical characteristics of large samples for evaluation of the dependence $\mathbf{v} = \mathbf{v}(M)$. The central limit theorem of the probability theory states that the dependence of the results of averaging across such samples reduces with the increase of the cardinal number of the sample. Applying this statement to the problem under consideration, let us assume that the dependence of vector $\mathbf{v}$ on the number of fragments can be neglected. Considering invariability of the geometric model to the homothety of tie images, we can assume $\sigma_0 = 1, k = 1$ in (11)-(13). As a result, we obtain the correspondence of feature vectors of two multicomponent images $\mathbf{Z}_1, \mathbf{Z}_2$ with different numbers of fragments M_1, M_2:

$$\mathbf{Z}_2 = \mathbf{Z}_1 \left(\frac{M_1 - M_2}{M_1 M_2} \right) + \mathbf{Z}_1 \left(\frac{M_1}{M_2} \right). \qquad (17)$$

For images with large numbers of fragments, the first summand in (17) is substantially smaller than the second, due to which we have

$$\mathbf{Z}_2 \cong \mathbf{Z}_1 \frac{M_1}{M_2}, \qquad (18)$$

that is, the feature vectors of two multicomponent images that represent a fixed class of geometric forms are practically collinear. Within the accuracy of equation (16), the feature vectors of the raster image and the discrete set of location of its components in the plane are also collinear. However, existing systems of information characteristics in the form of a set of invariant features are characterized by information incompleteness, and augmentation of their ensemble is associated with the increase of the space of the model, that is, with calculations of invariants of higher orders. One should note that the research we have carried out has shown that there also exist manifolds in the space of the model when normed semi-invariants of an odd order are used. Thus, for instance, in calculation of semi-invariants up to the fourth order, an image can be described by a six-parameter vector of invariant features, and, in calculation of semi-invariants of up to the fifth order, the system of information characteristics may have ten affine-invariant features. For example,

$$\begin{aligned}
\mathbf{I}_1 &= Z_{20}(Z_{21}Z_{03} - Z_{12}^2) - Z_{11}(Z_{03}Z_{30} - Z_{21}Z_{12}) + \\
&+ Z_{02}(Z_{30}Z_{12} - Z_{21}^2); \\
\mathbf{I}_2 &= Z_{20}^3 Z_{03}^2 - 6Z_{20}^2 Z_{11}Z_{12}Z_{03} - 6Z_{20}^2 Z_{21}Z_{02}Z_{03} + \\
&+ 9Z_{20}^2 Z_{02}Z_{12}^2 + 12Z_{20}Z_{11}^2 Z_{03}Z_{21} - \\
&- 18Z_{20}Z_{11}Z_{02}Z_{21}Z_{12} - 8Z_{11}^3 Z_{03}Z_{30} - \\
&- 6Z_{02}^2 Z_{20}Z_{30}Z_{12} + 9Z_{02}^2 Z_{20}Z_{21}^2 + Z_{02}^3 Z_{30}^2 + \\
&+ 12Z_{02}Z_{11}^2 Z_{30}Z_{12} - 6Z_{02}^2 Z_{11}Z_{30}Z_{21} + \\
&+ 6Z_{20}Z_{11}Z_{02}Z_{30}Z_{03}; \\
\mathbf{I}_3 &= Z_{50}^2 Z_{05}^2 - 10Z_{50}Z_{41}Z_{14}Z_{05} + 4Z_{50}Z_{32}Z_{23}Z_{05} + \\
&+ 16Z_{50}Z_{32}Z_{14}^2 - 12Z_{50}Z_{23}^2 Z_{14} + 16Z_{41}^2 Z_{23}Z_{05} + \\
&+ 9Z_{41}^2 Z_{14}^2 - 12Z_{32}^2 Z_{41}Z_{05} - 76Z_{41}Z_{32}Z_{23}Z_{14} + \\
&+ 48Z_{41}Z_{23}^3 + 48Z_{14}Z_{32}^3 - 32Z_{32}^2 Z_{23}^2; \\
\mathbf{I}_4 &= Z_{30}^2 Z_{12}Z_{05} - Z_{30}^2 Z_{14}Z_{03} - 2Z_{30}Z_{21}Z_{12}Z_{14} + \\
&+ 4Z_{30}Z_{21}Z_{03}Z_{23} + 2Z_{30}Z_{12}^2 Z_{23} - 4Z_{30}Z_{12}Z_{03}Z_{32} + \\
&+ Z_{03}^2 Z_{30}Z_{41} + 3Z_{21}^3 Z_{14} - 6Z_{21}^2 Z_{12}Z_{23} - 2Z_{21}^2 Z_{03}Z_{32} + \\
&+ 2Z_{03}Z_{21}Z_{12}Z_{41} - Z_{21}Z_{03}^2 Z_{50} - 3Z_{12}^3 Z_{41} + \\
&+ 6Z_{12}^2 Z_{21}Z_{32} + Z_{12}^2 Z_{03}Z_{50} - Z_{21}^2 Z_{30}Z_{05}; \\
\mathbf{I}_5 &= 2Z_{30}Z_{12}Z_{41}Z_{05} - 8Z_{30}Z_{12}Z_{32}Z_{14} + 6Z_{23}^2 Z_{30}Z_{12} - \\
&- Z_{30}Z_{03}Z_{50}Z_{05} + 3Z_{30}Z_{41}Z_{03}Z_{14} - 2Z_{41}Z_{21}^2 Z_{05} - \\
&- 2Z_{30}Z_{32}Z_{03}Z_{23} + 8Z_{21}^2 Z_{32}Z_{14} - 6Z_{21}^2 Z_{23}^2 - \\
&- 3Z_{21}Z_{12}Z_{41}Z_{14} + 2Z_{21}Z_{12}Z_{32}Z_{23} + 2Z_{21}Z_{03}Z_{50}Z_{14} - \\
&- 8Z_{21}Z_{03}Z_{41}Z_{23} + 6Z_{32}^2 Z_{21}Z_{03} - 2Z_{12}^2 Z_{50}Z_{14} + \\
&+ 8Z_{12}^2 Z_{41}Z_{23} + Z_{21}Z_{12}Z_{50}Z_{05} - 6Z_{12}^2 Z_{32}^2; \\
\mathbf{I}_6 &= 10Z_{33}^2 - 15Z_{24}Z_{42} + 6Z_{15}Z_{51} - Z_{06}Z_{60}; \\
\mathbf{I}_7 &= 35Z_{44}^2 - 56Z_{35}Z_{53} + 28Z_{26}Z_{62} - 8Z_{17}Z_{71} + Z_{08}Z_{80}; \\
\mathbf{I}_8 &= 126Z_{55}^2 - 210Z_{46}Z_{64} + 120Z_{37}Z_{73} - \\
&- 45Z_{28}Z_{82} + 10Z_{19}Z_{91} - Z_{0,10}Z_{10,0}.
\end{aligned} \qquad (19)$$

IV. Recognition Experiments

Testing of equations (17) and (18) was carried out for a photogrammetric image of an area of the Earth surface obtained from a spacecraft and shown in Fig. 2a. The image in Fig. 2b is the result of extraction of the contour, and its dimensions are 256×256 . In the framework of the employed model of morphogenesis of M-images this corresponds to a forming image as an individual pixel, its homothety with the coefficient $k = 1$ and parallel translation. The image shown in Fig. 2c was obtained on the basis of the homothety of a pixel with the coefficient $k = 0.5$, and the image in Fig. 2d, with $k = 0.25$. The numeric values of characteristics of the images are given in Table 1. Analysis of the respective data confirms satisfaction of the condition of collinearity of the feature vectors of the images.

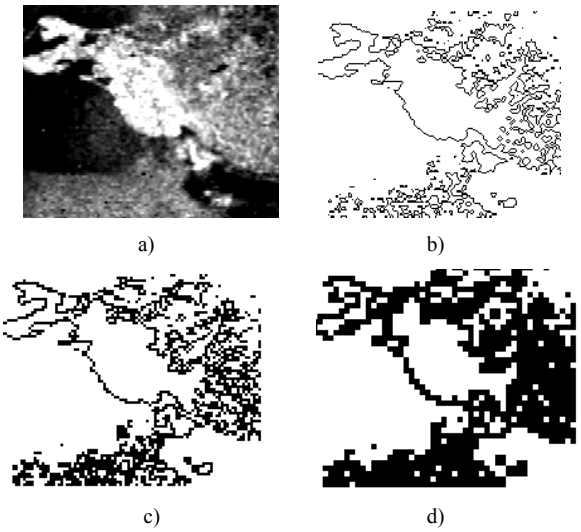


Fig. 2. Test results

TABLE I. Image Characteristics

Homothety Coefficient k	Z_1	Z_2	Z_3	Number of fragments M
k=1 (Fig. 2b)	0.840	-0.060	0.871	4576
k=0.5 (Fig. 2c)	1.541	-0.116	1.575	2523
k=0.25 (Fig. 2d)	3.316	-0.218	3.438	1155

The presented data confirm observation of equation (17), which, therefore, can be accepted as the necessary condition of belonging of two multicomponent images with feature vectors $\mathbf{Z}_1, \mathbf{Z}_2$ to the same class of geometric shapes, which is determined by the vector $\mathbf{v}$ of positioning of the fragments.

V. Conclusions and Future Work

In this study, a novel recognition method of remote sensing high-dimensional data of various spatial resolution was proposed. The proposed new method of photogrammetric image recognition represented in raster formats of computer graphics enables identification of objects without implementation of complex algorithms involving transformation of images. This, in turn, enables improvement of the recognition accuracy up to 96%. The hyperspectral data are now more often high dimensional not only spectrally, but also spatially and temporally. Further research will be related to consideration of the possibilities of application of the developed model to processing hyperspectral images.

References

[1] H. Taud, J. C. Herrera-Lozada, J. A. álvarez-Cedillo, M. Marciano-Melchor, R. Silva-Ortigoza and M. Olguín-Carbajal, "Circular object recognition from satellite images", 2012 IEEE Intern. Geoscience and Remote Sensing Symposium, Munich, 2012, pp. 2324-2327.

[2] Y.I. Shedlovska, V.V. Hnatushenko. "Shadow detection and removal using a shadow formation model". 2016 IEEE First International Conference on Data Stream Mining & Processing (DSMP), 2016. doi:10.1109/dsmp.2016.7583537.

[3] A. Ghandour, A. Jezzini. "Autonomous Building Detection Using Edge Properties and Image Color Invariants". Buildings, 8(5), 65, 2018. doi:10.3390/buildings8050065.

[4] D.K. Mozgovoy, V.V. Hnatushenko and V.V. Vasyliev. "Automated recognition of vegetation and water bodies on theterritory of megacities in satellite images of visible and IR bands", ISPRS Ann. Photogramm. Remote Sens. Spatial Inf. Sci., 2018. IV-3, 167–172, doi:10.5194/isprs-annals-IV-3-167-2018.

[5] Dmitriy Mozgovoy, Volodymyr Hnatushenko, and Volodymyr Vasyliev "Accuracy evaluation of automated object recognition using multispectral aerial images and neural network", Proc. SPIE 10806, Tenth International Conference on Digital Image Processing (ICDIP 2018), 108060H (9 August 2018);https://doi.org/10.1117/12.2502905.

[6] N. Ali, B. Zafar, MK. Iqbal, M. Sajid, MY. Younis, SH. Dar, et al. "Modeling global geometric spatial information for rotation invariant classification of satellite images". 2019, PLoS ONE 14(7): e0219833. doi:10.1371/journal.pone.0219833

[7] V. Gnatushenko "The use of geometrical methods in multispectral image processing". Journal of Automation and Information Sciences, Volume 35 (12), 1-8, 2003. doi:10.1615/JAutomatInfScien.v35.i12.10

[8] J. Flusser, J. Kautsky, and F. Šroubek. "Implicit Moment Invariants". International Journal of Computer Vision, 2009, 86(1), 72–86. doi:10.1007/s11263-009-0259-4 .

[9] M. Pawlak. "Image analysis by moments: reconstruction and computational aspects". Wroclaw: Wroclaw University of Technology Press, 2006.

[10] R.R. Bailey, M. Srinath. "Orthogonal moment features for use with parametric and non-parametric classifiers". IEEE Transactions on Pattern Analysis and Machine Intelligence, 1996, 18, 389–398.

[11] N. Tatar, H. Arefi. "Stereo rectification of pushbroom satellite images by robustly estimating the fundamental matrix". International Journal of Remote Sensing, 2019, 1–20. doi:10.1080/01431161.2019.1624862

[12] V.I. Voloshin, V.M. Korchinsky and M.M. Kharitonov. "A Novel Method For Correction Of Distortions And Improvement Of Informational Content In Satellite Acquired Multispectral Images". Advances and Challenges in Multisensor Data and Information Processing. - Vol. 8. – 2007. - P. 315-323.

[13] D. Zeng, S. Chen, B. Chen, S. Li. "Improving Remote Sensing Scene Classification by Integrating Global-Context and Local-Object Features". Remote Sensing. 2018;10(5):734.

[14] N. Ali, B. Zafar, F. Riaz, et al. "A Hybrid Geometric Spatial Image Representation for scene classification". PloS one. 2018; 13(9):e0203339. pmid:30208096

[15] L. Ma M. Li, X. Ma, L. Cheng, P. Du, Y .Liu. "A review of supervised object-based land-cover image classification". ISPRS Journal of Photogrammetry and Remote Sensing. 2017;130:277–293.

[16] X. Chang, S. Du, Y. Li, S. Fang. "A Coarse-to-Fine Geometric Scale-Invariant Feature Transform for Large Size High Resolution Satellite Image Registration". Sensors, 2018, 18(5), 1360. doi:10.3390/s18051360

[17] V.J. Kashtan, V.V. Hnatushenko and Y.I. Shedlovska. "Processing technology of multispectral remote sensing images". 2017 IEEE International Young Scientists Forum on Applied Physics and Engineering (YSF), 2017. doi:10.1109/ysf.2017.8126647.

Topic #3

Dynamic Data Mining & Data Stream Mining

IEEE Third International Conference on Data Stream Mining & Processing
August 21-25, 2020, Lviv, Ukraine

The Definition of Influence Different Drug Exposure Types to Medical Indicators of White Rats

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

Olha Lalymenko
Department of Hygiene and Ecology No. 2
Kharkiv National Medical University
Kharkiv, Ukraine
yaloposta@gmail.com

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

Viktor Reshetnik
Systems Engineering Department
Kharkiv National University of Radio Electronics
Kharkiv, Ukraine
viktor.reshetnik@nure.ua

Nelia Miroshnychenko
Biomedical Engineering Department
Kharkiv National University of Radio Electronics
Kharkiv, Ukraine
nelamiroshnichenko3@gmail.com

***Abstract*—At this paper the definition of influence by five different drug exposure types (three intragastrical and two inhalation types) to medical blood and urine indicators of white rat are investigated. Authors propose a method based on Manhattan metrics and aimed at comparing with special control group without any exposures.**

Keywords— principal component analysis, pharmaceutical companies, drug exposure, membership level, Manhattan metrics

I. INTRODUCTION

Nowadays the problem of environmental pollution by chemicals is a well-founded concern for the world and in particular the European community. The chemical-pharmaceutical industry according to the international classification (US Environmental Protection Agency) is referred to the group of environmentally hazardous enterprises [1]. Approximately 4.5 thousand chemical enterprises are under sanitary and epidemiological surveillance in Ukraine; up to 940 chemicals used within industry have been registered [2]. Chemical-pharmaceutical companies are sources of toxic compounds emissions into the production medium and environment [3]. High manufacturability and growth rate of the industrial complex of chemical-pharmaceutical enterprises are leading to an increase in release as well as use of new chemicals, including medications, consequently increasing the likelihood of chemical pollution of industrial sites and the environment. Such events, in turn, contribute to increased risks of occupational pathology of enterprise staff and population morbidity [4-5].

Most working chemical-pharmaceutical companies are daily exposed to chemical compounds during manufacture of medicines. The leading adverse factor, in particular within the manufacture of medicines, is the contamination of harmful organic and inorganic substances in the air of working premises, clothing and skin of workers, surfaces of equipment, building structures, industrial areas and environment [6-7]. If the sanitary and hygienic requirements are not complied with, the release of harmful chemicals takes place due to insufficient sealing of the equipment within the course of technological processes, communication, transportation, loading and unloading of raw materials, sampling of air, technological process. The person who works under such conditions receives an unknown amount of toxic compound through the respiratory system, mucous membranes and skin. The toxicant eventually spreads throughout the human body and exhibits its toxic effect [8].

Within industrial production medicines and their active components are potentially hazardous to the health of workers, as they can enter the working area and directly cause different manifestations of abnormalities or pathological processes in the human body [9].

The prevalence of production-related pathology caused by chemical etiology among other work-related diseases remains significant and constitutes 23%; occupational intoxication among workers in such field is ranked fourth in the list of general and occupational morbidity [3].

The mentioned data conditions the need to develop hygienic standards for the content of chemical compounds in the production environment at all stages of the technological process, as well as the need to develop methods for their biological control [10].

One of the important measures for the prevention of the harmful effects of chemical substances on working chemical and pharmaceutical enterprises is the systematic control of the content of harmful substances in the air of the working area (chemical monitoring). However, the determination of chemical compounds in the air of the working area allows estimating their concentration only at a specific time moment and at a certain place [11]. Unfortunately, the control of the content of toxic compounds in the air of the working area cannot completely exclude the effect of toxic substances on the body of the working person. Therefore, priority should be accorded to introduce methods of biological monitoring of the effects of toxic compounds on humans in order to achieve more careful control of the effects of toxicants on the human body.

In accordance with WHO regulations, biological monitoring involves the identification and measurement of concentrations of toxic chemicals in human biological environments (blood plasma, urine, etc.) followed by further comparison of these data with the levels of the specified toxicant in the air of the working area/atmosphere air and developed pathological conditions in human body [12]. Biomonitoring is intended to eliminate uncertainty when assessing the degree of exposure of a toxicant, which inevitably arises in the case of using indirect methods, which rely on estimated approaches and are unable to take into account the individual doses absorbed by the organism, to evaluate the impact of the toxicant that comes from many sources and via different ways. Therefore, biomonitoring can be considered as the main tool for determining the severity and nature of the effect of an unfavorable factor on the human body and undoubtedly complementing the sanitary-chemical control of the air in the working environment area [13].

Hygienic identification of hazardous chemical compounds and assessment of their long-term effects is an important component of the system of evidence of cause-and-effect relationships between workers' / population's health effects and the formation of unwanted health abnormalities. Such aspect, in turn, allows deeply assessing the real risk of exposure of toxicants to the body of enterprise personnel [14]. The methodology of human biological monitoring has been recently widely used to establish cause-and-effect relationships between the action of chemical compounds, in particular medications, and environmental pollution and / or the occurrence of occupational pathology. Many authors emphasize the need for in-depth study and development of methodological approaches in this area in order to implement them in the practice of occupational health and environmental protection [15].

The term "Human biomonitoring" was initially proposed in 1980 at a seminar organized by the European Economic Community (EEC) together with the US National Institute for Occupational Safety and Health (NIOSH) and the Occupational Safety and Health Administration (OSHA) in Luxembourg [16].

Biological monitoring can be defined as a method of assessing the effects of hazardous chemical compounds on human health by measuring the content of these substances in samples of human biological material. With regard to production facilities, the proposed approach involves establishing a relationship between the toxicant level in the air of working area and the substance content in human materials, thus the use of exposure tests. Exposure test (biomarker of exposure) is the content of the toxicant in the biological substrate (plasma and serum, urine, saliva, hair, etc.), which depends on the dose level of the substance that has entered the human body. Biological control of the industrial exposure of hazardous chemicals allows us to estimate the integral dose of a harmful compound in the body, regardless of the route of entry, with the determination of the real risk to humans, to identify individuals with increased individual sensitivity and early signs of intoxication. The application of this approach within occupational health field requires regular monitoring of exposure test values and markers of the effect of the compound on the body of workers, which are established taking into account the occupation, working experience, production process and other specific conditions [17].

The conduct of biological monitoring of humans as a mandatory component of the health care in most industrialized countries, such as countries of European Union, the United Kingdom, the United States and Russia, is regulated by a number of international legal instruments [18-20]. In accordance with WHO regulations, human biomonitoring enables the detection of toxicity and contributes to a more complete assessment of the risk of effect of exogenous chemical agents on the health of workers and public health in general.

The literature analysis shows that adequate biomedical criteria for early diagnosis of disease are needed under the conditions of anthropogenic chemical factors on human health; such criteria first of all must reflect the relationship between the occurrence of the disease and the action of relevant toxicants and to enable the identification and evaluation of changes at the molecular, subcellular, cellular, organ and organismic levels [21]. Secondly, these indicators should identify the negative effects, specific disorders in the body and various aspects of damaging effects of the toxin followed by identification of the most vulnerable target organs with the ability to differentiate the processes of physiological adaptation, compensatory-adaptive reactions and pathological conditions. Within prophylaxis medicine these criteria must be defined at different exposure lines and different dose levels in order to ensure the safety of the population and the working population who are in the face of prolonged exposure to dangerous chemical factors [21-23].

II. BIOLOGICAL MARKERS FOR TOXICOLOGY AND RISK ASSESSMENT

According to the results of studies of biochemical, immunological, hematological, oxidative-antioxidant status, determination of the state of neuroendocrine regulation, molecular-genetic parameters, chemical-analytical studies, a comparative assessment of the degree of indicators deviations from the norm in humans is performed. This, in turn, contributes to the selection of clinical laboratory diagnostic indicators as informative and highly sensitive markers [24].

According to the WHO definition, "biomarker is basically any quantitative indicator or system of indicators that reflects the interaction between a biological system (a human body) and a potentially dangerous environmental factor (physical, chemical, biological)". Such indicator could be functional, physiological and biochemical and must necessarily reflect the interaction at the cellular or molecular level. Biomarkers allow:

- to identify individuals with high levels of hazardous chemical agents in the biological substrate (blood, saliva, urine);
- to establish (deny) the presence of a disease caused by the action of a toxic compound;
- to evaluate specific effects of the toxicant on the human body;
- to determine the presence of hypersensitivity reaction of the human body under the action of a specific toxicant;
- to check possible mechanisms of action of the toxicant;
- to control the influence of hazardous chemical compounds on people at risk (population living in

hazardous areas, children from 0 to 14 years, pregnant and fertile women, professional contingent who have contact with hazardous production factors);

- to evaluate the effectiveness of preventive measures for the population / professional contingent [25-27].

Taking into account the mentioned data, such area of research allows more careful monitoring of the effects of chemical compounds (toxicants) in workers involved in the production process, in particular production of medications as well as to prevent the effects of chemical compounds on their body. At the same time, the studies provide an individualized approach, identify at-risk individuals, identify diseases that have emerged from such impact, and develop preventative measures at a particular pharmaceutical enterprise.

III. Experimental Results

The study involved the experimental determination of a chemical compound (succinic acid derivative – a medication) in the blood of laboratory animals, followed by a comparative study of these data with health disorders, which were determined by the study of physiological, biochemical parameters in male rats. The representative sample was formed by the method of random selection of animals from the general population, their distribution into experimental groups – by the method of randomization. All experimental studies were conducted in accordance with the "General Ethical Principles of Animal Experiments". Toxicological studies included the determination of drug concentrations in the blood plasma of rats under the conditions of a 30-day intragastric exposure to a compound at a dose of 100/mg/kg at the stages of 5th, 15th, and 30th day of experiment and modeling inhalation (through respiratory tract) at doses of 1 and 7 mg/ml, respectively, at the stage of 21 days.

So, input dataset presented by 5 data frames, each contains of 7 (for intragastric exposure) or 10 (for inhalation exposure) rats described by 14 features (malondialdehyde, lipid hydroperoxides, diene conjugates, superoxide dismutase, catalase, glutathione peroxidase, blood plasma nitrites, nitrates, urine nitrites and nitrates, Nitric oxide synthases) (Fig.1). Each group can be associated with a particular control group (rats without any exposures). Number of rats in control groups is 42.

Index	f1	f2	f3	f4	f5	f6	f7	f8	f9	f10	f11	f12	f13	f14
IN0001	15.9	0.30	0.38	5.0	141.3	2.56	154.3	26.2	4.11	16.1	8.7	53.0	2.40	124.5
IN0002	19.0	0.46	0.56	4.7	115.9	2.31	144.2	25.7	3.90	13.1	11.2	52.0	1.70	120.3
IN0003	17.9	0.38	0.32	5.4	131.0	2.87	139.2	22.4	4.10	14.2	10.1	54.7	1.89	115.0
IN0004	23.6	0.55	0.44	4.1	46.3	2.20	133.7	34.1	4.66	12.3	9.5	53.2	1.64	99.2
IN0005	27.9	0.81	0.49	4.6	39.4	1.56	98.8	33.9	3.98	10.2	8.2	48.7	1.43	114.1
IN0006	26.3	0.71	0.50	4.5	84.7	1.65	122.8	24.8	4.00	10.9	8.0	49.4	1.50	121.7
IN0007	25.2	0.68	0.55	5.7	40.5	2.02	126.5	22.1	3.76	11.0	7.6	46.3	1.57	85.4

Fig. 1. Data Frame of intragastric exposure of drug at the stages of 5th day

The principal goal is to find the group of biggest influence of drugs in ratio to common control group.

At first step we can provide data visualization using PCA approach [28] for all dataset and centers of cluster (Fig. 2). Different data frames are marked by circles (red for intragastric exposure at the stages of 5th day, green for intragastric exposure at the stages of 15th day, blue for intragastric exposure at the stages of 30th day, magenta for inhalation at doses of 1 mg/ml, orange for inhalation at doses of 7 mg/ml and black for control group), cluster centers are marked by squares. For calculation the positions of cluster centers we should calculate mean in each group.

At second step we should calculate distances in Manhattan metrics [29]-[31] between centers of each group in ratio to control group (marked by lines on Fig. 3) to determine the influence of drugs using formula:

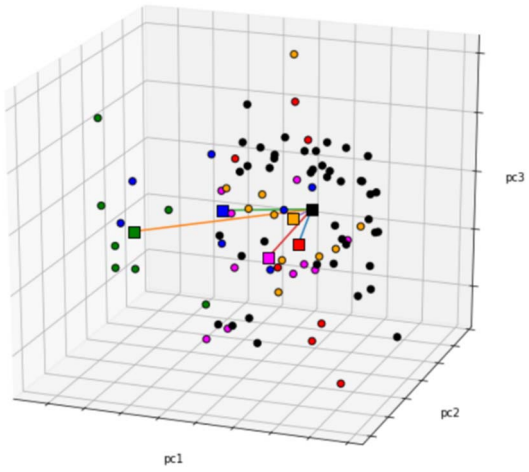


Fig. 2. PCA-visualization (red for intragastric exposure at the stages of 5th day, green for intragastric exposure at the stages of 15th day, blue for intragastric exposure at the stages of 30th day, magenta for inhalation at doses of 1 mg/ml, orange for inhalation at doses of 7 mg/ml and black for control group)

$$d_q = \sum_{i=1}^{n} \left| c_{iq} - c_{im} \right|$$

where $q = 1 \,..\, m$ is the number of group (m=6 is a control group), $n = 14$ is a number of medical features.

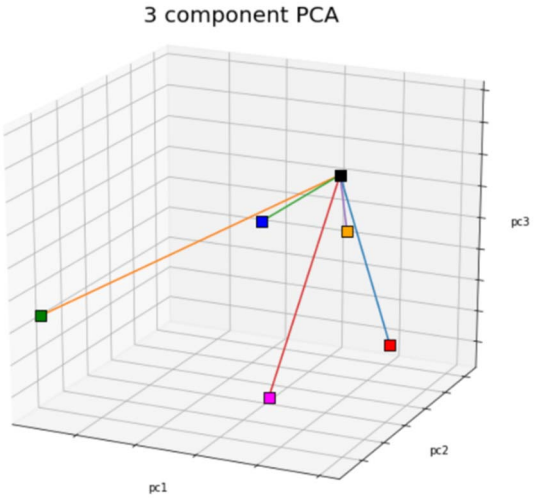


Fig. 3. The distances between clusters in ratio to control group

The last step is a calculation of values of membership functions for finding the group of biggest influence of drugs in ratio to common control group using following way:

$$mf_l = \frac{\left(\frac{1}{d_l}\right)}{\sum_{l=1}^{m-1}\left(\frac{1}{d_l}\right)}$$

The values of membership functions (*mf*1 for intragastric exposure at the stages of 5th day, *mf2* for intragastric exposure at the stages of 15th day, *mf3* for intragastric exposure at the stages of 30th day, *mf4* for inhalation at doses of 1 mg/ml, *mf5* for inhalation at doses of 7 mg/ml) are:

mf1 = 0.17;

mf2 = 0.10;

mf3 = 0.12;

mf4 = 0.28;

mf5 = 0.33.

To interpret the result it is necessary to take into account the fact that the membership function are minimized (low values correspond to high levels of influence and high values correspond to low level of influence). So, the values of membership functions show the biggest influence of intragastric drugs exposure at the stages of 15th day (*mf*2 = 0.1) and low level of influence during inhalation at doses of 7 mg/ml (*mf*5 = 0.33).

CONCLUSION

The definition of influence by five different drug exposure types like intragastric exposure at the stages of 5th day, intragastric exposure at the stages of 15th day, intragastric exposure at the stages of 30th day, inhalation at doses of 1 mg/ml and inhalation at doses of 7 mg/ml to medical blood and urine indicators of white rat were investigated. Method include a calculation of distances in Manhattan metrics between all cluster's centers and control group and definition of influence's level of drugs exposure.

REFERENCES

[1] P. Whitaker "Occupational allergy to pharmaceutical products," Curr Opin Allergy Clin Immunol, vol. 16(2), 2016, pp. 101-106.

[2] A. Nagorna, I. Kononova, N. Grechkivska. "The current state and ways of improving the organization and control over medical examinations of workers in hazardous and dangerous conditions," Ukrainian Journal of Occupational Health, vol. 3(40), 2014, pp. 10–19 (in Ukrainian)

[3] P. Apostoli, R. Cornelis, J. Duffus, P. Hoet, D. Lison, D. Templeton, S. Hahn, J. Kielhorn, M. Nordberg, V. Riihimäki, A. "Aitio Environmental Health Criteria 234: Elemental speciation in Human Health Risk Assessment," Geneva: WHO/IPCS, 2006, 256 p. Available from: http://www.who.int/ipcs /publications/ehc/ehc234.pdf

[4] G. Mosconi, G. Bartolucci, P. Apostoli "The role of occupational physician in the risk assessment," G. Ital. Med. Lav. Ergon., vol. 36(2), 2014, pp. 69–77.

[5] K. Fulcher, H. Gibb "Setting the research agenda on the health effects of chemicals," Int J Environ Res Public Health, vol. 11(1), 2014, pp. 1049-57.

[6] K. Gielen, A. Goossens "Occupational allergic contact dermatitis from drugs in healthcare workers," Contact Dermatitis, vol. 45(5), 2001, pp. 273-279.

[7] S. Araya, E. Lovsin-Barle, S. Glowienke "Mutagenicity assessment strategy for pharmaceutical intermediates to aid limit setting for occupational exposure," Regul Toxicol Pharmacol, vol. 73(2), 2015, pp. 515-20.

[8] S.P. Binks "Occupational toxicology and the control of exposure to pharmaceutical agents at work," Occup Med., London, vol. 53(6), 2003, pp. 363-370

[9] R.J. Heron, F.C. Pickering, Health effects of exposure to active pharmaceutical ingredients (APIs). Occup Med London, vol. 53(6), 2003, pp. 357-362.

[10] M. Bader, T. Van Weyenbergh, E. Verwerft, J. Van Pul, S. Lang, C. Oberlinner "Human biomonitoring after chemical incidents and during short-term maintenance work as a tool for exposure analysis and assessment," Toxicol Lett, vol. 231(3), 2014, pp.328-336.

[11] Committee on Human Biomonitoring for Environmental Toxicants. Human Biomonitoring for Environmental Chemicals, Washington: National Academies Press. NRC, 2006. Available from: https://www.nap.edu/catalog/11700/human biomonitoring for environmental chemicals

[12] Principles for the assessment of risk to human health from exposure to chemicals, Geneva: WHO/IPCS, 1999, 91 p. Available from: http://www.pic.int/Portals/5/secEdoc/Environmental%20Health%20Criteria%20210.pdf

[13] Y. Wu, J. Gu, Y. Huang, Y. Duan, R. Huang, J. Hu "Dose-Response relationship between cumulative occupational lead exposure and the associated health damages: a 20-year cohort study of a smelter in China," Int J Environ Res Public Health, vol. 13(3), 2016, pp. 356–359.

[14] R. Stahlmann, A. Horvath "Risks, risk assessment and risk competence in toxicology," Ger Med Sci, 2015, pp. 9-13.

[15] WHO, "Principles for evaluating health risks in children associated with exposure to chemicals (Environmental Health Criteria 237)," Geneva: WHO, 2006 [cited 2014 March 19]. Available at: http://www.who.int/ipcs/publications/ehc/ehc237.pdf

[16] R. Persoons, J. Richard, C. Herve, S. Montlevier, M. Marques, A. Maitre "Biomonitoring of styrene occupational exposures: Biomarkers and determinants," Toxicol Lett, 2018 [cited 2016 Feb 12]; 291: S1066–9. Available from: https://www.ncbi.nlm.nih.gov.

[17] Council Directive 98/24/EC. "On the protection of the health and safety of workers from the risk related to chemical agents at work," Brussels: European commission, 1998, 23 p. Available from: http://www.reach-compliance.eu

[18] E. Richter "Biomonitoring of human exposure to arylamines," Front Biosci (Elite Ed), vol. 1(7), 2015, pp. 193-207.

[19] S. Ling, D. McD Taylor, J. Robinson "Workplace chemical and toxin exposures reported to a Poisons Information Centre: a diverse range causing variable morbidity," Eur J Emerg Med, vol. 25(2), 2018, pp. 134–139.

[20] ACGIH. "Documentation of the Threshold Limit Values and Biological Exposure Indices", 7th Edition, Denver: ACGIH, 2017. Available from: Available from: https://www.acgih.org

[21] J. Edward, "Calabrese biological effects of low level exposures to chemical and radiation," Reissued: CRC Press, 2018, 155 p.

[22] A. Alves, A. Kucharska, C. Erratico "Human biomonitoring of emerging pollutants through non-invasive matrices: state of the art and future potential," Anal Bioanal Chem, vol. 406(17), 2014, pp. 63-88.

[23] Europian Comission. "Methodology for the Derivation of Occupational Exposure Limits: Key Documentation, " Brussels: SCOEL, 2009. Available from: http://www.sesric.org/imgs/news/Image/671-a-7.pdf

[24] L. Pralong, A. Berthet, D. Vernez, N. Hopf, L. Benaroyo "Biomonitoring information management and communication: an ethical and interdisciplinary perspective," Rev Med Suisse, vol. 11(499), 2015, pp.2400-2403.

[25] L.K. Lowry "How to promote the use of biological monitoring," Toxicol. Lett, vol. 231(2), 2014, pp. 289–290.

[26] WHO. "Biomarkers and risk assessment: concepts and principles," Geneva, 1993, 80 p.

[27] L. Mortimer, E. Mendelsohn "Biomarkers and occupational health. Progress and perspectives," Washington: Josef Henry Press, 1995, 65p.

[28] Ye. Bodyanskiy, I. Perova, P. Zhernova Online fuzzy clustering of high-dimensional data based on ensembles in data stream mining tasks, The current state of research and technology in industry, vol. 1(7), 2019, pp. 16-24.

[29] P. Zhernova, A. Deyneko, Zh. Deyneko, I. Pliss, V. Ahafonov "Data stream clustering in conditions of an unknown amount of classes," Vol. 754, pp. 410-419. In: Kacprzyk J. Advances in Intelligent Systems and Computing, Warsaw, Poland, Springer, 2019, 773 p.

[30] P. Zhernova, A. Deyneko, Ye. Bodyanskiy, V. Riepin "Adaptive Kernel Data Streams Clustering Based on Neural Networks Ensembles in Conditions of Uncertainty About Amount and Shapes of Clusters," IEEE Second International Conference on Data Stream Mining & Processing, 21-25 August, 2018, Lviv, Ukraine, pp. 7-12.

[31] I. Perova, Ye. Bodyanskiy "Adaptive fuzzy clustering based on manhattan metrics in medical and biological applications," Bulletin of National University "Lviv Polytechnic", vol. 826, 2015, pp. 8–12.

IEEE Third International Conference on Data Stream Mining & Processing
August 21-25, 2020, Lviv, Ukraine

The Combined Time Series Forecasting Model

Fedir Geche
Department of Cybernetics and Applied Mathematics
Uzhhorod National University
Uzhhorod, Ukraine
fgeche@hotmail.com

Anatoliy Batyuk
ACS Department
Lviv Polytechnic National University
abatyuk@gmail.com

Oksana Mulesa
Department of Cybernetics and Applied Mathematics
Uzhhorod National University
Uzhhorod, Ukraine
Oksana.Mulesa@uzhnu.edu.ua,

Veronika Voloshchuk
Uzhhorod National University
Uzhhorod, Ukraine
veronika.smolanka@uzhnu.edu.ua

***Abstract*—This article presents the method of constructing a combined forecasting model based on known time series forecasting models. At the forecasting stage, each basic model "produces" its weighting coefficient with which it is included in the combined model. At the training stage of the combined model, these weighting coefficients are refined, the most influential basic models for the given time series are determined, and a combined forecasting model is constructed with respect to these models.**

Keywords—functional, least squares method, trend, weighting coefficient, step of forecast, step of prehistory, model training.

I. INTRODUCTION

The development of effective methods for predicting time series is a topical and practically important task. Time series makes it possible to set different indicators in the field of economy, in the social sphere, in medicine, etc. The qualitative forecast of economic indicators enables one to find additional resources of enterprises, to develop substantiated strategic plans of their economic activity. In this regard, the development of effective methods for predicting the financial condition of enterprises has been the subject of scientific researches by lots of scientists, such as Blank I.A. [1], Heyets V.M. [2], Zaychenko Y.P. [3], Ivakhnenko V.M. [4], Tkachenko R. [5], Bodyanskiy Y. [6], Teslyuk V. [7], Tsmots I. [8].

It is impossible to single out the "best" method of the forecast of the economic indicators used to determine the financial state or the efficiency of the utilization of production resources of an enterprise because of the diversity of internal laws (trends) of different systems. Therefore, there exists a problem of choosing an effective forecast method.

The main purpose of this article consists in reducing the risk of obtaining the forecast of poor quality in the way of combining different forecasting models with appropriate weighting coefficients so that they would take into account each other's disadvantages.

II. CONSTRUCTION OF THE COMBINED TIME SERIES FORECASTING MODEL

Let $v_1, v_2, ...v_t, ..., v_n$ be a time series and covariance of v_t with each element from (v_{t-1}, v_{t-2} …) is not equals to zero. To construct a combined forecasting model using the autoregression method, one shoud determine the optimal prehistory step for a given time series with a fixed forecast step. The autoregressive forecast value is determined as follows:

$$\tilde{v}_{n+\tau} = a_1^{(\tau)} v_n + a_2^{(\tau)} v_{n-1} + ... + a_{k_\tau}^{(\tau)} v_{n-k_\tau+1} \ . \qquad (1)$$

Let $a_1^{*(\tau)}, ..., a_{k_\tau}^{*(\tau)}$ be optimal parameters of the model (1), i.e. ones that minimize the functional

$$L\left(a_1^{(\tau)}, ..., a_{k_\tau}^{(\tau)}\right) = \sum_{t=k_\tau+\tau}^{T} \left(v_t - a_1^{(\tau)} v_{t-\tau} - ... - a_k^{(\tau)} v_{t-\tau-k_\tau+1}\right)^2 .$$

In accordance with (1):

$$\tilde{v}_{n+\tau} = a_1^{*(\tau)} v_n + a_2^{*(\tau)} v_{n-1} + ... + a_{k_\tau}^{*(\tau)} v_{n-k_\tau+1} , \qquad (2)$$

where $t \geq k + \tau$.

Obviously, the value $\tilde{v}_t$ at the fixed $\tau = \tau_0$ depends on the parameter $k_\tau (1 \leq k_\tau \leq n - \tau)$. To find the optimal value of the prehistory parameter k_τ (at $\tau = \tau_0$) for the given time series v_t, we consider the following variables:

$$\delta_1 = \frac{1}{n-\tau} \sum_{t=\tau+1}^{n} (v_t - a_1^{*(\tau)} v_{t-\tau})^2 ,$$

$$\delta_2 = \frac{1}{n-\tau-1} \sum_{t=\tau+2}^{n} (v_t - a_1^{*(\tau)} v_{t-\tau} - a_1^{*(\tau)} v_{t-\tau-1})^2 ,$$

....

$$\delta_{n-\tau} = (v_n - a_1^{*(\tau)} v_{t-\tau} - ... - a_{n-\tau}^{*(\tau)} v_1)^2$$

and calculate $\min\{\delta_1, \delta_2, ..., \delta_{n-\tau}\} = \delta_{k_\tau^*}$. The value k_τ^* determines the optimal value of the autoregression model prehistory parameter at fixed $\tau = \tau_0$.

After determining k_τ^* ($\tau = \tau_0$), let us consider different time series forecasting models $M_1, M_2, ..., M_q$ with the forecast step τ at the following time points $n - k_\tau^* + 1, n - k_\tau^* + 2, ..., n$. Based on the results of the forecast using the above mentioned methods $M_1, M_2, ..., M_q$, the following table is constructed:

TABLE I. FORECAST VALUES OF THE ELEMEVYS OF A TIME SERIES WITH RESPECT TO DIFFERENT MODELS

For e-casting models	**Real values of time series elements at time points** $n-k_\tau^*+1$	$n-k_\tau^*+2$	...	n
	$v_{n-k_\tau^*+1}$	$v_{n-k_\tau^*+2}$	...	v_n
M_1	$\tilde{v}_{n-k_\tau^*+}^{(1)}$	$\tilde{v}_{n-k_\tau^*+2}^{(1)}$	...	$\tilde{v}_n^{(1)}$
M_2	$\tilde{v}_{n-k_\tau^*+}^{(2)}$	$\tilde{v}_{n-k_\tau^*+2}^{(2)}$	...	$\tilde{v}_n^{(2)}$
⋮	⋮	⋮	...	⋮
M_q	$\tilde{v}_{n-k_\tau^*+}^{(q)}$	$\tilde{v}_{n-k_\tau^*+2}^{(q)}$	...	$\tilde{v}_n^{(q)}$

In each column $v_{n-k_\tau^*+1}, v_{n-k_\tau^*+2}, ..., v_n$ of the table 1, the least square deviation between the predicted and the real values of the corresponding terms of the time series is determined. Mathematically, this can be written as follows:

let $j_1 = n - k_\tau^* + 1$ and

$$\varepsilon_1 = \min\left\{(v_{j_1} - \tilde{v}_{j_1}^{(1)})^2, (v_{j_1} - \tilde{v}_{j_1}^{(2)})^2, ..., ((v_{j_1} - \tilde{v}_{j_1}^{(q)})^2)\right\},$$

$j_2 = n - k_\tau^* + 2$ and

$$\varepsilon_2 = \min\left\{(v_{j_2} - \tilde{v}_{j_2}^{(1)})^2, (v_{j_2} - \tilde{v}_{j_2}^{(2)})^2, ..., ((v_{j_2} - \tilde{v}_{j_2}^{(q)})^2)\right\},$$

.....

$j_{k_\tau^*} = n$ and

$$\varepsilon_{k_\tau^*} = \min\left\{(v_n - \tilde{v}_n^{(1)})^2, (v_n - \tilde{v}_n^{(2)})^2, ..., ((v_n - \tilde{v}_n^{(q)})^2)\right\},$$

Let us define the sets $I_1, I_2, ...I_{k_\tau}$ as follows:

$$I_1 = \{\, i \in \{\, 1,2,...,q \,\} \mid \varepsilon_1 = (v_{j_1} - v_{j_1}^{(i)})^2 \,\},$$

$$I_2 = \{\, i \in \{\, 1,2,...,q \,\} \mid \varepsilon_2 = (v_{j_2} - v_{j_2}^{(i)})^2 \,\},$$

.....

$$I_{k_\tau^*} = \{\, i \in \{\, 1,2,...,q \,\} \mid \varepsilon_{k_\tau^*} = (v_n - v_n^{(i)})^2 \,\}$$

and construct the following table:

TABLE II. PARAMETERS OF THE FORECASTING SCHEMES

Fore-casting models	j_1	j_2	...	$j_{k_\tau^*}$	**Result column**
M_1	$a_{11}(\beta)$	$a_{12}(\beta)$	...	$a_{1k_\tau^*}(\beta)$	$S_1(\beta)$
M_2	$a_{21}(\beta)$	$a_{22}(\beta)$	...	$a_{2k_\tau^*}(\beta)$	$S_2(\beta)$
⋮	⋮	⋮	⋮	⋮	⋮
M_q	$a_{q1}(\beta)$	$a_{q2}(\beta)$	...	$a_{qk_\tau^*}(\beta)$	$S_q(\beta)$

where

$$a_{ps}(\beta) = \begin{cases} \beta^{k_\tau^* - s} & if\ s \in I_s, \\ 0 & if\ s \notin I_s, \end{cases}$$

$$S_p(\beta) = \sum_{j=1}^{k_\tau^*} a_{pj}(\beta), \quad 0 < \beta < 1 \ (p = 1,2,...,q; s = 1,2,...,k_\tau^*).$$

By means of $S_p(\beta)$ and $S(\beta) = \sum_{p=1}^{q} S_p(\beta)$, the weighting coefficients of the forecasting models M_p $(p = 1,2,...,q)$ are determined, with which they are included in the following forecasting model:

$$\tilde{v}_{n+\tau} = \frac{S_1(\beta)}{S(\beta)}\tilde{v}_{n+\tau}^{(1)} + \frac{S_2(\beta)}{S(\beta)}\tilde{v}_{n+\tau}^{(2)} + ... + \frac{S_q(\beta)}{S(\beta)}\tilde{v}_{n+\tau}^{(q)}, \quad (3)$$

where $\tilde{v}_{n+\tau}^{(i)}$ is the forecast value of a time series according to the M_i model with the forecast step τ.

Before using the model (3), let us carry out its training with respect to β. This means that the model (3) is transformed depending on the value of β. Some models remain with modified weighting coefficients and some are excluded from the model.

To evaluate the forecasting quality of the model (3), the following functional is used:

$$H(\beta) = \sum_{i=1}^{k_\tau^*} \left(v_{j_i} - \frac{S_1(\beta)}{S(\beta)}\tilde{v}_{j_i}^{(1)} - \frac{S_2(\beta)}{S(\beta)}\tilde{v}_{j_i}^{(2)} - ... - \frac{S_q(\beta)}{S(\beta)}\tilde{v}_{j_i}^{(q)} \right)^2 \quad (4)$$

where $(j_i = n - k_\tau^* + i)$.

The model (3) is trained according to the following algorithm:

Step 1. We set m points $\beta_r = \frac{r}{m}$ $(r = 1,2,...,m)$ on the interval (0,1] and for each fixed $\beta = \beta_r$ of the nonzero elements of the last column of the table 2 we construct the ordered set $U_r = \left(S_{r_1}(\beta_r) \geq S_{r_2}(\beta_r) \geq ... \geq S_{r_{q_r}}(\beta_r)\right)$.

Step 2. We define the set of most influential models U_r^* with respect to $\beta = \beta_r$. Initially, we assume that $U_r^* = \{M_{r_1}\}$. The equality (3) with respect to one of the models M_{r_1} is written as follows: $\tilde{v}_{n+\tau} = \tilde{v}_{n+\tau}^{(r_1)}$.

According to (4), we determine the value of the functional $H_1(\beta_r)$:

$$H_{r_1}(\beta_r)=\sum_{i=1}^{k_\tau^*}\left(v_{j_i}-\widetilde{v}_{j_i}^{(r_1)}\right)^2.$$

We establish the rule of including the model M_{r_2} in the combined model. Let us consider the model

$$\widetilde{v}_{n+\tau}=\frac{S_{r_1}(\beta_r)}{S(\beta_r)}\widetilde{v}_{n+\tau}^{(r_1)}+\frac{S_{r_2}(\beta_r)}{S(\beta_r)}\widetilde{v}_{n+\tau}^{(r_2)}, \quad (5)$$

where $S(\beta_r)=S_{r_1}(\beta_r)+S_{r_2}(\beta_r)$.

We find the value of the functional $H_2(\beta_r)$ with respect to the model (5)

$$H_{r_2}(\beta_r)=\sum_{i=1}^{k_\tau^*}\left(v_{j_i}-\frac{S_{r_1}(\beta_r)}{S(\beta_r)}\widetilde{v}_{j_i}^{(r_1)}-\frac{S_{r_2}(\beta_r)}{S(\beta_r)}\widetilde{v}_{j_i}^{(r_2)}\right)^2.$$

If $H_2(\beta_r)<H_1(\beta_r)$, then the model M_{r_2} is included in the set U_r^*, i.e.

$$U_r^*=\{M_{r_1}\}\cup\{M_{r_2}\}.$$

The process of constructing the set U_r^* for each fixed r we continue until the conditions for the inclusion of models from U_r holds. After the set U_r^* $(r=1,2,...,q_r)$ is constructed, we proceed to the step 3.

Step 3. Let $U_r^*=\{M_{r_1},M_{r_2},...,M_{r_{h_r}}\}$ and $H_{r_{h_r^*}^*}(\beta_{r^*})=\min\{H_{r_{h_r}}(\beta_r)\,|\,r=1,2,...,m\}$. Then the combined forecasting model with respect to the sets $U_{r^*}^*=\left\{M_{r_1^*},M_{r_2^*},...,M_{r_{h_r^*}^*}\right\}$ can be determined as follows:

$$\widetilde{v}_{n+\tau}=\frac{S_{r_1^*}(\beta_{r^*})}{S(\beta_{r^*})}\widetilde{v}_{n+\tau}^{(r_1^*)}+\frac{S_{r_2^*}(\beta_{r^*})}{S(\beta_{r^*})}\widetilde{v}_{n+\tau}^{(r_2^*)}+...+\frac{S_{r_{h_r^*}^*}(\beta_{r^*})}{S(\beta_{r^*})}\widetilde{v}_{n+\tau}^{(r_{h_r^*}^*)}, \quad (6)$$

where $S(\beta_{r^*})=S_{r_1^*}(\beta_{r^*})+S_{r_2^*}(\beta_{r^*})+...+S_{r_{h_r^*}^*}(\beta_{r^*})$.

An important stage in forecasting is verification of the forecast, i.e. estimating its accuracy and validity. At the verification stage, a set of criteria are used to assess the quality of the forecasting.

The effectiveness of the forecasting model (6) is illustrated by passenger rail transportations data in Ukraine (the source of input data is [9]).

The criterion MRE (Mean Relative Error) was used to evaluate the quality of the combined forecasting model. MRE is determined by the following formula

$$MRE=\frac{1}{n}\sum_{t=1}^{n}\left|\frac{v_t-\widetilde{v}_t}{v_t}\right|,$$

where v_t is the value of a time series at the moment t; $\widetilde{v}_t$ is the forecasting value v_t.

The mean relative error (MRE) can be used to compare two (or more) different forecasts of the same time series: the one with the smaller MRE value is preferable.

According to the criterion of mean relative error, let us evaluate the quality of the forecast of the combined forecasting model by comparing its results with the results of the classical forecasting models which are the basis for the combined one: autoregression, M_1; the least squares method with weighting coefficients, M_2; the first-order Brown method, M_3; the second-order Brown method, M_4.

The volume of passenger transportation by rail in Ukraine for the period of 1980-2013 is given in [9]. The results of the forecast of the models M_1, M_2, M_3, M_4, and the combined model with the corresponding parameters are given in the following tables:

TABLE III. RAIL TRANSPORT (QUALITY OF FORECASTING):

	Autoreg-ression model M_1	**Least squares method with weight-ings M_2**	**First-order Brown method M_3**	**Second-order Brown method M_4**	**Combi-ned model**
Mean relative error	0.0041	0.015	0.0358	0.0159	0.0039

TABLE IV. RAIL TRANSPORT (QUALITY OF FORECASTING):

	Autoreg-ression model M_1	**Least squares method with weight-ings M_2**	**First-order Brown method M_3**	**Second-order Brown method M_4**	**Combi-ned model**
Mean relative error	0.0045	0.0048	0.0585	0.0041	0.0031

Parameters of the combined model:

$n=34;\widetilde{v}_{34+1}=0{,}48*M_1+0{,}009*M_2+0{,}511*M_4;\beta_{r^*}=0{,}7;k_\tau^*=13.$

Having analyzed the data of the last tables, we can see that the combined forecasting model has the smallest mean relative error. Therefore, the constructed combined forecasting model is the most effective one among the methods on which it is based.

III. CONCLUSIONS

The presented forecasting model is flexible and can be successfully used to develop strategic plans in various fields of human activity. The final forecasting model is obtained by training the basic model. The training is based on the competition between input (basic) models. The final forecasting model includes the most influential basic models for the time series being under investigation. The most influential models for the given time series are determined at

the stage of training the combined model. In the learning process, the forecasting model automatically adjusts to the time series under investigation, and as a result, it will be a convex linear combination of the most influential basic models. This approach gives us an opportunity to construct an effective forecasting model that can be successfully used to forecast different types of time series in the fields of economics, medicine, social sphere, etc.

REFERENCES

[1] I.A. Blank, "*Strategy and Tactics of Financial Management*". The Item LTD, Kyiv (1996) (in Ukrainian)

[2] V.M. Heyets, "*Iinstability and Economic Growth. Institute of Economic Forecasting of National*", Academy of Sciences of Ukraine, Kyiv (2002) (in Ukrainian)

[3] Y.P. Zaychenko, M. Moamed, N.V. Shapovalenko, "*Fuzzy Neural Networks and Genetic Algorithms in Problems of Macroeconomic Forecasting*." Science news of "Kyiv Polytechnic Institute", 4, 20-30. Kyiv (2002) (in Ukrainian)

[4] V. Ivakhnenko, "*Course of Economic Analysis. Znannya Press*", Kyiv (2000). (in Ukrainian}

[5] R. Tkachenko, I. Izonin, P. Vitynskyi, N. Lotoshynska, and O. Pavlyuk. "*Development of the Non-Iterative Supervised Learning Predictor Based on the Ito Decomposition and SGTM Neural-Like Structure for Managing Medical Insurance Costs*", Data, vol. 3, no. 4, рз. 1-14, 2018. DOI:10.3390/data3040046

[6] Y. Bodyanskiy, A. Dolotov, D. Peleshko, Y. Rashkevych, O. Vynokurova, "*Associative probabilistic Neuro-Fuzzy system for data classi_cation under short training set conditions*". Advances in Intelligent Systems and Computing 761, 56-63 (2019).

[7] T. Teslyuk, I. Tsmots, V. Teslyuk, M. Medykovskyy, 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: Shakhovska N., Stepashko V. (eds) Advances in Intelligent Systems and Computing II. CSIT 2017. Advances in Intelligent Systems and Computing, vol. 689, Springer, Cham., pp. 538 – 557.

[8] I. Tsmots, et. al. "*Basic Components of Neuronetworks with Parallel Vertical Group Data Real-Time Processing.*" Advances in Intelligent Systems and Computing 689, Springer, Cham.: 2018, pp. 558–576.

[9] Transport and connection of Ukraine 2013 /Government service of statistics. Statistical collection. 2013. 552p

IEEE Third International Conference on Data Stream Mining & Processing
August 21-25, 2020, Lviv, Ukraine

Model of Evaluation and Selection of Start-up Projects by Investor Goals

Mykola Malyar
Department of cybernetics and applied mathematics
Uzhgorod National University
Uzhgorod, Ukraine
malyarmm@gmail.com

Miroslav Kelemen
Department of flight preparation
Technical University of Kosice
Kosice, Slovak Republic
miroslav.kelemen@tuke.sk

Andriy Polishchuk
Department of cybernetics and applied mathematics
Uzhgorod National University
Uzhgorod, Ukraine
andriy.v.polishchuk@gmail.com

Volodymyr Polishchuk
Department of software systems
Uzhgorod National University
Uzhgorod, Ukraine
volodymyr.polishchuk@uzhnu.edu.ua

Marianna Sharkadi
Department of cybernetics and applied mathematics
Uzhgorod National University
Uzhgorod, Ukraine
marianna.sharkadi@uzhnu.edu.ua

***Abstract* — The actual problem of developing a fuzzy mathematical model of evaluation and selection of start-up projects for the purposes of investors has been conducted. The model developed will be a useful tool to substantiate and increase the security of investors' choice of alternative start-up project financing, using their own targeted needs.**

Keywords — start-up, a decision maker (DM), linguistic evaluation, multicriteria, investor.

I. INTRODUCTION

Today, the importance of the tasks in which compromise decisions have to be made in the process of researching complex social objects has increased to a great extent, in the case of fuzzy or incomplete information. In the period of globalization and innovative development, we see a large number of innovative and start-up projects in all sectors. Yesterday's futuristic projects are now a reality. At such a dynamic pace, innovative start-up projects are increasingly emerging outside of companies. Successful companies carefully consider the issue of purchasing and implementing technology through innovative projects. At the same time, new investment opportunities and support for such projects are emerging. As a consequence, new technological ideas that grow into real projects are possible, as a rule, under external financing. Therefore, it becomes necessary to finance such projects for their implementation on the market and in the future for conquering the market.

Financing start-up projects is a complex, risky, fuzzy and unpredictable activity. Such a task is borne out by the fact that the financing of projects depends largely on the opportunities, wishes, considerations, and opinions of investors [1]. And in any decision support system, the final decision is made by the person. Such systems should help, advise and raise the level of validity of decision-making. Therefore, decision support is taken to a whole new level with the formalization of expertise.

Recent scientific studies indicate the need to develop new models for evaluating innovative or start-up projects, for investors to choose for their own goals. The relevance and need of these models is indicated by an increase in the number of start-ups at university / regional / state level competitions; the emergence of specialized crowdfunding platforms (support for projects in a particular area) and venture funds; a large number of grant and innovation funding programs; the trend of development and implementation of own innovations in the enterprises with the involvement of employees [2-3].

Summarize of the above, there is an urgent task of multicriteria assessment of start-up (innovative) projects for investors to choose for the linguistic goals of the idea prospect, analysis of project implementation risks and competencies of project developers. To do this, the problem of evaluating alternatives when there are multiple goals, each with its own set of criteria should be solved.

II. LITERATURE REVIEW

Will analyze the scientific approaches related to this issue. Note the following methods of multicriteria selection: methods based on quantitative variables, multicriteria utility theory [4]; methods based on qualitative characteristics, the results of which are translated into quantitative form (analytical hierarchy methods) [5]; methods based on fuzzy set theory [6]; methods based on qualitative variables without moving to quantitative ones [7], and others.

Let's analyze the sources, related to the use of fuzzy mathematics apparatus to create support systems for managerial decision-making for various sectors of the economy. In works [8-9] are reviewing general ideas and advantages that underpin modern views on the use of fuzzy logic in decision support systems. In [10-11] presents computational algorithms and procedures for solving practical problems of system analysis in various fields of human activity. The application of paired comparison methods and the consistency of expert assessments are presented in [12]. Thus, the general ideas and benefits of using fuzzy logic in decision support systems are discussed in [13-14]. There are a number of works [15-17] that demonstrate models of enhancing the security of choice of alternatives by groups of goals, but the model is tested in the choice of banking institutions by the entity.

Concerning sources on the topic of evaluation of start-up projects, we conclude: there are currently a number of methods based on simulation and expert models using economic quantitative indicators [18-19]. The problem of

978-1-7281-3214-3/20/$31.00 ©2020 IEEE

evaluating start-up projects is raised in [20-21], where the apparatus of fuzzy sets is used. In works [22-24] uses the apparatus of fuzzy sets, fuzzy logic, and a systematic approach to assessing the risk of financing start-up projects. In [25-26], the problem of constructing an information model for evaluating and rating the start-up teams is raised.

Thus, the issues of evaluating start-up projects, assessing risks, and evaluating development teams with the use of fuzzy sets have already been raised. The issue of multicriteria selection of alternatives by the target group was also raised. But there are no integrated, multi-criteria systematic approach for evaluating and selecting start-up projects, given the targeted needs of investors' wishes, which is so relevant in today's context.

Therefore, the purpose of this paper is to develop a fuzzy mathematical model for evaluating and selecting start-up projects for investors.

III. Data and Methodology

Formulating the task of evaluating the object of study as follows. Set multiple alternatives (start-up s or innovative projects) $P=\{p_1, p_2, \dots, p_n\}$, which need to be evaluated according to goals $G=\{G_1, G_2, \dots, G_g\}$ and sort by some rule. Each of the goals of G has some model of evaluating alternatives. According to the condition of the problem, alternatives $P=\{p_1, p_2, \dots, p_n\}$, it is necessary to evaluate by models of estimation and to build a ranking series of choosing the best start-up project, depending on the following goals of the investor: needs for a prospect start-up project, risk assessment of the project implementation and assessment of the competence of the team of start-up project developers. The block diagram of the solution of the problem can be presented as a follow, fig. 1.

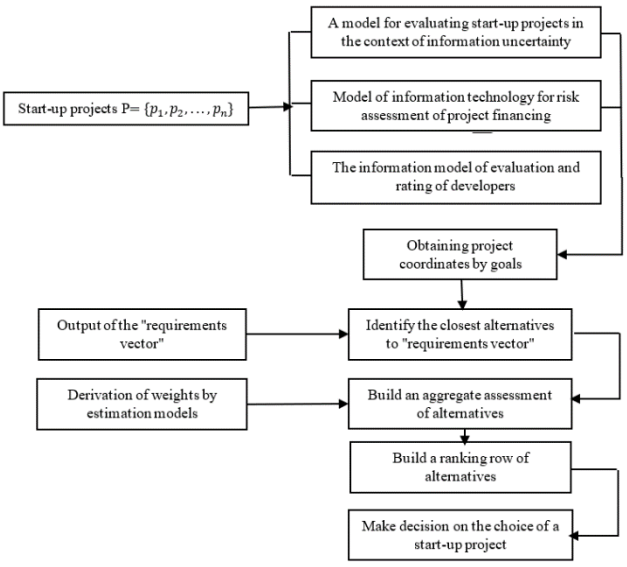


Fig. 1. Structural diagram of the solution of the problem

The model for obtaining an aggregate estimate is represented as:

$$M(G_1(s_1, \dots, s_n); G_2(r_1, \dots, r_n); G_3(t_1, \dots, t_n)) \to P^*. \quad (1)$$

As a result, for each alternative $P=\{p_1, p_2, \dots, p_n\}$, there are finding of normalized estimates that determine the best alternative P^* ; $s_1, s_2, \dots, s_n$ – assessing the relevant alternatives $P=\{p_1, p_2, \dots, p_n\}$ for a goal G_1; $r_1, r_2, \dots, r_n$ – estimates for the goal G_2 and $t_1, t_2, \dots, t_n$ for G_3.

If we have many start-up projects $P=\{p_1, p_2, \dots, p_n\}$, which need to be evaluated and selected by investors to fund them. Each project is evaluated using models that output estimates from the interval [0; 1]. The following goal evaluation models are proposed:

- model of estimation of start-up projects in the conditions of information uncertainty of vectors [21], as a result we get a set $S=\{s_1, s_2, \dots, s_n\}$;
- model of information technology for project risk assessment [23] – $R=\{r_1, r_2, \dots, r_n\}$;
- an information model of evaluating and rating teams of start-up development [26] – $T=\{t_1, t_2, \dots, t_n\}$.

Because we have the task of evaluating alternatives consisting of three goals, then vectors $S=\{s_1, s_2, \dots, s_n\}$, $R=\{r_1, r_2, \dots, r_n\}$ and $T=\{t_1, t_2, \dots, t_n\}$ we design on a three-dimensional coordinate system where the values set by S – is the value plotted on the x-axis, R – axis y, T – axis z. For each alternative, we will get the coordinates by goals G_1, G_2, G_3 which we present in the form: (s_1, r_1, t_1), (s_2, r_2, t_2), …, (s_n, r_n, t_n).

Next, we introduce the three-dimensional "satisfaction vector" $T^*=(A_1, A_2, A_3)$, which takes into account the wishes of a DM regarding the value of the alternatives for the goals G_1, G_2, G_3.

Definition. "Vector of satisfaction of requirements" is an imaginary alternative in which estimates of coordinates by purpose could satisfy a decision-maker [17].

We describe the model of the "requirements vector" as follows. Let the object with 3 inputs and one output be analyzed:

$$U=(A_1, A_2, A_3), \quad (2)$$

where U – is the vector of the initial estimate (u_1, u_2, u_3), whose components take one of the values {0,2; 0,4; 0,6; 0,8; 1}, A_1, A_2, A_3 – are input linguistic variables.

To evaluate the linguistic variables A_1, A_2, A_3 we use qualitative terms from the following term sets:

$$A_1=(a_{11}, a_{12}, \dots, a_{1t}),\ A_2=(a_{21}, a_{22}, \dots, a_{2t}),$$

$$A_3=(a_{31}, a_{32}, \dots, a_{3t}). \quad (3)$$

The knowledge of the "requirements vector" $T=(t_1, t_2, t_3)$ defines a base of fuzzy knowledge in the form of a system of logical statements – "If - Then, Else", which associates the values of input variables A_1, A_2, A_3 with one of the possible values U.

$$\text{If } A_1=a_{1t} \text{ and } A_2=a_{2t} \text{ and } A_3=a_{3t}$$

$$\text{Then } U=(u_1, u_2, u_3) \text{ Else…}. \quad (4)$$

Thus, DM sets the linguistic desire for the "satisfaction vector", which we transfer into the vector of initial quantitative and normalized estimation(u_1, u_2, u_3), denoted respectively $(u_1, u_2, u_3) = (t_1, t_2, t_3)$.

Therefore, the fuzzy knowledge base can be formulated as follows:

IF we have goals:

project start-up prospects (group of indicators G_1):

- a_{11} there is a need for a promising concept then $u_1 = 0{,}2$;
- a_{12} there is a priority need for a promising concept then $u_1 = 0{,}4$;
- a_{13} there is a need for a strong idea and a finished product then $u_1 = 0{,}6$;
- a_{14} the significant need for a promising, strong idea and finished product then $u_1 = 0{,}8$;
- a_{15} the priority needs for a promising, strong idea and finished product then $u_1 = 1$.

AND project implementation risk (group of indicators G_2):

- a_{21} high then $u_2 = 0{,}2$;
- a_{22} average then $u_2 = 0{,}4$;
- a_{23} low then $u_2 = 0{,}6$;
- a_{24} very low then $u_2 = 0{,}8$;
- a_{25} minimum then $u_2 = 1$.

AND Competencies of the start-up project team (group of indicators G_3):

- a_{31} very low then $u_3 = 0{,}2$;
- a_{32} low then $u_3 = 0{,}4$;
- a_{33} average the $u_3 = 0{,}6$;
- a_{34} above average then $u_3 = 0{,}8$;
- a_{35} high then $u_3 = 1$.

THEN logical statement can be formulated as follows:

If we need the prospect of A_1 start-up and the risk of A_2 and the competence of the A_3 start-up team then $U = (u_1, u_2, u_3)$.

Next we find the values of the quantities $\mu(f_{1i})$, $\mu(f_{2i})$, $\mu(f_{3i}), i = \overline{1,n}$, which will allow us to determine the closest alternatives to the "requirements satisfaction vector":

$$\mu(f_{1i}) = \frac{|u_1 - s_i|}{max\left\{u_1 - \min\limits_i s_i; \max\limits_i s_i - u_1\right\}}, \quad (5)$$

$$\mu(f_{2i}) = \frac{|u_2 - r_i|}{max\left\{u_2 - \min\limits_i r_i; \max\limits_i r_i - u_2\right\}}, \quad (6)$$

$$\mu(f_{3i}) = \frac{|u_3 - t_i|}{max\left\{u_3 - \min\limits_i t_i; \max\limits_i t_i - u_3\right\}}, i = \overline{1,n}. \quad (7)$$

After that, we calculate the values $Z_i = (z_{1i}, z_{2i}, z_{3i}), i = \overline{1,n}$, which characterize the relative estimates of the proximity of alternatives to the "requirements satisfaction vector" for each individual objective G_1, G_2, G_3, which removes the question of different rating scales [17]:

$$Z_i = (1;1;1) - (\mu(f_{1i}); \mu(f_{2i}); \mu(f_{3i})), i = \overline{1,n}. \quad (8)$$

Let the decision maker set the weights for each estimation model$\{\alpha_1, \alpha_2, \alpha_3\}$, for example from the interval [1,10]. For further calculations we carry out their normalization:

$$w_1 = \frac{\alpha_1}{\alpha_1+\alpha_2+\alpha_3}, w_2 = \frac{\alpha_2}{\alpha_1+\alpha_2+\alpha_3}, w_3 = \frac{\alpha_3}{\alpha_1+\alpha_2+\alpha_3}. \quad (9)$$

Next, to construct an aggregate estimate, we use one of the convolutions, for example, take a weighted average [17]:

$$Z_i^* = w_1 \cdot z_{1i} + w_2 \cdot z_{2i} + w_3 \cdot z_{3i}, i = \overline{1,n}. \quad (10)$$

Based on the estimates obtained, we select the best start-up project considering the goals of investors:

$$P^* = \max\limits_i Z_i^*, \quad i = \overline{1,n}. \quad (11)$$

Therefore, the best alternative solution will be closest to the "requirements vector" for the goals G_1, G_2, G_3.

IV. EXPERIMENTS

Research results will be tested on the following problem. Consider some start-up projects $P = \{p_1, p_2, \ldots, p_5\}$ (taken from the University Science Park TECHNICOM ecosystem in Kosice and the start-up incubator at the Uzhhorod National University), which should be evaluated on the proposed models and selected for financing according to the following investor goals: G_1 – perspective of the start-up project, G_2 – project implementation risk, G_3 – the competencies of the start-up project development team.

Consider the proposed assessment models and their criteria established by the group of experts.

To evaluate start-up projects, we use an evaluation model in the context of information uncertainty [21]. The model reduces the subjectivity of expert judgment, shows the place of the "idea" among others, allows to set the level of its risk and to take into account the wishes of the decision maker [21]. The evaluation criteria are as follows:

SK_1 – type of goods; SK_2 – field of application; SK_3 – social significance; SK_4 – the power of the idea; SK_5 – strategic partners; SK_6 – the value of the percentage growth of the market for this start-up.

As a result of the estimation, by the model given in [19], we obtain a set of normalized estimates $S = \{s_1, s_2, \ldots, s_5\}$.

To assess the risks of start-up projects, we use the information technology model to evaluate the risk of project financing [22]. Model: Increases the objectivity of expert judgment in project risk assessment, using input linguistic variables and the credibility of expert judgment regarding their assignment; allows changing the levels of decision-

making in the knowledge base, depending on the liquidity of the investment institution; integrates opinions by criteria groups into the final assessment and degree of risk of the project, based on a two-level fuzzy mathematical model [23]. The evaluation criteria in this model are as follows:

RK_1 – risk of the client base loss; RK_2 – risk of supplier loss; RK_3 – the risk of reducing processes quality; RK_4 – the risk of reduced productivity; RK_5 – risk of resource insecurity; RK_6 – risk of inefficient investment; RK_7 – risk of disruption of terms of creation of production funds; RK_8 – risk of exceeding the amount of start-up investment; RK_9 – risk of investor loss.

As a result of the estimation, by the model given in [23], we obtain a set of normalized estimates $R = \{r_1, r_2, \dots, r_5\}$.

To evaluate start-up project developers, we use an information model to evaluate and derive the rating of start-up project development teams [26]. The proposed model: increases the objectivity of peer review in evaluating development teams, using input linguistic variables and the "confidence coefficient" of expert judgment on their assignment; is based on a neuro-fuzzy network that can change the synaptic weight setting; has the ability to train the neuro-fuzzy network by complementing the knowledge base and adjusting the rankings of start-up project development teams [26]. The evaluation criteria in this model are as follows:

TK_1 – successful experience in related or close to the topic; TK_2 – successful management experience; TK_3 – education of leaders; TK_4 – successful experience in large or similar projects; TK_5 – professional education of team members; TK_6 – participation of the team in professional conferences, investment sessions or profile events; TK_7 – publications in the media or professional sources for the project; TK_8 – availability of social networking and messaging with the team; TK_9 – having links with social media advisors.

As a result of the estimation, according to the model given in [26], we obtain a set of normalized estimates $T = \{t_1, t_2, \dots, t_5\}$.

Projects were modeled and aggregated, Table I.

TABLE I. INPUTS FOR START-UP PROJECTS

Models of evaluation	p_1	p_2	p_3	p_4	p_5
S	0.87	0.82	0.6	0.77	0.69
R	0.66	0.83	0.71	0.98	0.91
T	0.78	0.4	0.54	0.85	0.82

Let the investor express wishes regarding the start-up project as follows:

"The prospect of a start-up project $A_1 = \{$there is a priority need for a promising concept$\}$ and the risk of project implementation $A_2 = \{$low$\}$ and the competencies of the start-up project development team $A_3 = \{$above average$\}$".

Then accordingly $(t_1, t_2, t_3) = (0{,}4;\ 0{,}6;\ 0{,}8)$. Find the values of quantities $\mu(f_{1i}),\ \mu(f_{2i}),\ \mu(f_{3i}), i = \overline{1,5}$, that will determine the closest alternatives to the "requirements vector" according to formulas (5)-(7), for example: $\mu(f_{11}) = \frac{|0.4-0.87|}{max\{0.4-0.6;0.87-0.4\}} = 1;$ $\mu(f_{12}) = \frac{|0.4-0.82|}{max\{0.4-0.6;0.87-0.4\}} = 0.894$. All results of calculations will be presented in the Table II.

TABLE II. RESULTS OF CALCULATION VALUES $\mu(f)$

Models of evaluation	p_1	p_2	p_3	p_4	p_5
S	1.000	0.894	0.426	0.787	0.617
R	0.158	0.605	0.289	1.000	0.816
T	0.050	1.000	0.650	0.125	0.050

Next, we calculate the values $Z_i = (z_{1i}, z_{2i}, z_{3i}), i = \overline{1,5}$, according to formula (8), Table III.

TABLE III. RESULTS OF CALCULATION VALUES Z

Models of evaluation	p_1	p_2	p_3	p_4	p_5
S	0.000	0.106	0.574	0.213	0.383
R	0.842	0.395	0.711	0.000	0.184
T	0.950	0.000	0.350	0.875	0.950

Let the decision-maker determine the weighting coefficients for the estimation models – $\{10{,}8{,}9\}$. Normalize them by the formula (9): $w_1 = 0.37;\ w_2 = 0.3;\ w_3 = 0.33$.

Calculate the aggregate estimate by the formula (10): $Z_1^* = 0.37 \cdot 0 + 0.3 \cdot 0.842 + 0.33 \cdot 0.95 = 0.566;$ $Z_2^* = 0.156;\ Z_3^* = 0.54;$ $Z_4^* = 0.37;\ Z_5^* = 0.513.$

Based on our estimates, we build a ranking of alternatives – $\{p_1, p_3, p_5, p_4, p_2\}$. As a result, we conclude that the best start-up project considering the goals of investors is – p_1 with the highest score of 0.566.

V. RESULTS AND DISCUSSION

Built-in model for evaluating and selecting start-up s for investors' goals has several advantages, namely: it increases the objectivity of evaluating alternative options; allows to solve the problem of evaluating alternatives to goals and models of evaluation; builds a ranking number of start-up projects represented by evaluation vectors on different valuation models and improves security of choice of alternatives; investors' wishes are set in natural language, which allows for rapid adaptation to various financial institutions and design contests; the model allows to work with different rating scales that boil down to comparatives.

The disadvantages of this approach can be attributed to the use of different convolution models to obtain an aggregate estimate, which may lead to ambiguity in the final results.

The result of the study is a model of choice of start-up projects for the purposes of investors, the output of which is the overall aggregated assessment of start-up projects and

their ranking. The rationality of the obtained estimation is proved the advantages of the developed model.

VI. Conclusion

Therefore, investigation results as follow:

- For the first time, a mathematical model is presented for solving the problem of evaluating start-up projects according to the investor's goals: the need for prospects for a start-up project, an assessment of the risk of project implementation, and an assessment of the competence of the start-up project team. For these purposes, start-up project evaluation models based on fuzzy set theory and neuro-fuzzy networks have been proposed to reduce the uncertainty of project appraisals and the subjectivity of evaluators.
- For the first time, a "requirements vector" is proposed for the model for solving the multicriteria choice of start-up projects using investor wishes in the form of linguistic considerations of this task.
- Tested model to evaluate and build a ranking of five start-up projects. For example, the investor's linguistic desire was considered for the following purposes: "Priority need for a promising concept for a start-up project and low risk of project implementation and competence of the above-average start-up project team".

The model developed will be a useful tool to substantiate and increase the security of investors' choice of alternative start-up project financing, using their own targeted needs. Further study of the problem can be seen in the approbation of the developed model for the university incubator in the competitive selection of start-up projects for their financial support.

Acknowledgment

The work was carried out within the framework of the state budget research topics of Uzhhorod National University: "Development of mathematical models and methods for information processing and data mining" and "Software engineering methods and tools for the implementation of big data analytics processes on the basis of Information-Electronic Science Platforms".

References

[1] G. Wei, P. Sun, Z. Zhang, X. Ouyang, "The Coordinated Relationship between Investment Potential and Economic Development and Its Driving Mechanism: A Case Study of the African Region", *Sustainability,* 12, 442, 2020.

[2] I. Scholz, Reflecting on the Right to Development from the Perspective of Global Environmental Change and the 2030 Agenda for Sustainable Development. In Sustainable Development Goals and Human Rights, Springer: Cham, Switzerland, 2020.

[3] F. Jovanović, N. Milijić, M. Dimitrova, I. Mihajlović, "Risk Management Impact Assessment on the Success of Strategic Investment Projects: Benchmarking Among Different Sector. Companies", *Acta Polytechnica Hungarica,* Vol. 13, No. 5, pp. 221 – 241, 2016.

[4] R. Kyny, KH. Rayfa, Prynyattya rishen′ pry bahat′okh kryteryyakh: prypushchennya ta zamishchennya, Moscow: Radio i zvyaz, 1981.

[5] T. Saaty. Prynyatye reshenyy. Metod analizu ierarkhiy, Moscow: Radio i zvyaz, 1993.

[6] L. Zade, Ponyatiye lingvisticheskoy peremennoy i yego primeneniye k prinyatiyu priblizhennykh resheniy, Moscow: Mir, 1976.

[7] O. Larychev, Verbal′nyy analiz vyrishenyy: monohrafiya, Moscow: Nauka, 2006.

[8] A. Kofman, KH. Khil Alukha, Vvedennya teoriy nechyslennykh mnozhyn v upravlinni pidpryyemstvam, Minsk: Vysh. shk., 1992.

[9] YU.P. Zaychenko, Nechetkiye modeli i metody v intellektualnykh sistemakh: navchalny posibnyk, Kiyev: Slovo, 2008.

[10] M. Zgurovsky, N. Pankratova, Osnovy systemnoho analizu, Kyyiv: VNV, 2007.

[11] M. Zgurovsky, Yu. Zaychenko, Big Data: Conceptual Analysis and Applications, Springer: New York, 2020

[12] N. Pankratova, N. Nedashkovskaya, "The Method of Estimating the Consistency of Paired Comparisons", *Information Technologies and Knowledge,* 7, № 4., pp. 347-361, 2013.

[13] K. V. Kumar, Neural networks and fuzzy logic. S. K. Kataria & Sons: New Delhi, 2009.

[14] Ye. Bodyanskiy, Yu. Zaychenko, E. Pavlikovskaya, M. Samarina, Ye. Viktorov, "The neo-fuzzy neural network structure optimization using the GMDH for the solving forecasting and classification problems", *Proc. Int. Workshop on Inductive Modeling*, pp. 77-89, 2009.

[15] B. Gavurova, F. Janke, M. Packova, M. Pridavok, "Analysis of impact of using the trend variables on bankruptcy prediction models performance", *Ekonomický časopis,* Vol. 65, No. 4, pp. 370-383, 2017.

[16] J. Belás, M. Mišanková, J. Schonfeld, B. Gavurová, "Credit risk management: Financial safety and sustainability aspects", *Journal of Security and Sustainability Issues*, Vol. 7, No. 1, pp. 79-93, 2017.

[17] V. Polishchuk, "Technology to Improve the Safety of Choosing Alternatives by Groups of Goals", *Journal of Automation and Information Sciences,* Volume 51, Issue 9, pp.66-76, 2019.

[18] A. Damodaran, Valuing young, start-up and growth companies: estimation issues and valuation challenges, New York: Stern School of Business New York University, 2009.

[19] O.D. Zvyahintseva, I.O. Zolotarova, O.V. Shcherbakov, "Intehrovana otsinka startap-proektiv", *Systemy obrobky informatsiyi,* 4, pp. 163-165, 2015.

[20] O.A. Shvetsova, E.A. Rodionova, M.Z. Epstein, "Evaluation of investment projects under uncertainty: multi-criteria approach using interval data", *Entrepreneurship and Sustainability,* Issues 5(4), pp. 914-928, 2018.

[21] M. Kelemen, V. Polishchuk, B. Gavurová, S. Szabo, R. Rozenberg, M. Gera, J. Kozuba, J. Hospodka, R. Andoga, A. Divoková, P. Blišťan, "Fuzzy Model for Quantitative Assessment of Environmental Start-up Projects in Air Transport", *Int. J. Environ. Res. Public Health,* vol. 16, pp. 3585, 2019, [online] Available: https://doi.org/10.3390/ijerph16193585

[22] V. Polishchuk, M. Kelemen, B. Gavurová, C. Varotsos, R. Andoga, M. Gera, J. Christodoulakis, R. Soušek, J. Kozuba, P. Blišťan, S. Szabo Jr., "A Fuzzy Model of Risk Assessment for Environmental Start-up Projects in the Air Transport Sector", *Int. J. Environ. Res. Public Health*, vol. 16, pp. 3573, 2019, [online] Available: https://doi.org/10.3390/ijerph16193573

[23] O. Voloshyn, M. Malyar, V. Polishchuk, M. Sharkadi, "Fuzzy mathematical modeling financial risks", *IEEE Second International Conference on Data Stream Mining & Processing (DSMP),* pp. 65-69, 21-25 August 2018, 2018.

[24] G. Koulinas, O. Demesouka, P. Marhavilas, A. Vavatsikos, D. Koulouriotis, "Risk Assessment Using Fuzzy TOPSIS and PRAT for Sustainable Engineering Projects", *Sustainability,* 11 (3), 615, 2019.

[25] J. C. Mendialdua, "Using fuzzy logic in selecting people and ideas to participate in public programs of support to business start-ups", *Cuadernos de Gestion*, 14(2), pp. 73-98, 2014.

[26] M. Kelemen, V. Polishchuk, "Information Model of Evaluation and Output Rating of Start-up Projects Development Teams", *Proceedings of the Second International Workshop on Computer Modeling and Intelligent Systems (CMIS-2019)*, *CEUR Workshop Proceedings, Vol. 2353,* pp. 674-688, 15-19 April 2019, 2019.

IEEE Third International Conference on Data Stream Mining & Processing
August 21-25, 2020, Lviv, Ukraine

Nonparametric Test for Change-Point Detection in Data Stream

Dmitriy Klyushin, Irina Martynenko
Department of Computer Science and Cybernetics
Taras Shevchenko National University of Kyiv
Kyiv, Ukraine
dokmed5@gmail.com

***Abstract*—We offer a new effective online algorithm for implement the Klyushin–Petunin test on streaming data. The Klyushin–Petunin test is a nonparametric test to evaluate the statistical hypothesis that two samples are drawn from the same distribution. The significance level of the test does not exceed 0.05. The test is based on the only assumption that the underlying distribution is absolutely continuous. We present an algorithm for detecting a change in distribution in a data stream. It allows solving two canonical problems of analyzing statistical data — detection of a change-point and detection of a drift of concepts. We show that proposed algorithm is more effective compared to the alternative Kolmogorov–Smirnov and Wilcoxon tests.**

***Index Terms*—Data Stream, Change-Point Problem, Concept Drift, Nonparametric Test, Similarity Measure**

I. Introduction

One of the main problems of big data is the difficulty of storing huge amount of information. In these situations, many offline algorithms that use full information about general population become infeasible. Thus, we have to use online algorithms that accepts partial information about a data stream and compare data, falling into adjacent sliding windows. There are two problems arise in these approach. First, the size of the sliding window should be small enough so that there are no difficulties with storing data. Second, the small size of the sliding window require that the applied algorithms be sufficiently sensitive and recognize changes using small samples. In this work, we propose such an algorithm and show its advantages over alternative options.

The Klyushin—Petunin test (KP test) is a criterion for testing the statistical hypothesis that two samples belong to the same population, i.e. they are homogeneous. Its advantage is that it does not use any assumptions about a type of distribution, in contrast to parametric tests like the Student t-test and can be applied to any samples drawn from general populations with an absolutely continuous distribution function. The test is based on the measure of similarity between samples (p-statistics), which is used to reject the hypothesis that the distributions coincide at a given significance level.

II. State-of-the-art

The problem of change-point detection is equivalent to identification of change in distribution of data stream. So, it could be reformulated as the two-sample problem of homogeneity and solved by the sliding window approach and a statistical two-sample homogeneity test.

Let $x = (x_1, x_2, ..., x_n)$ and $y = (y_1, y_2, ..., y_m)$ be two samples from general populations G_1 and G_2 with continuous distribution functions F_1 and F_2 respectively. The main hypothesis is $F_1 = F_2$ against the alternative hypothesis $F_1 \neq F_2$. The criteria for testing such hypotheses can be classified in a different way,particularly by the principle of validation: permutation criteria, rank criteria, randomization criteria, and distance criteria. In addition, these tests are divided into universal tests that are valid against any pair of alternatives (for example, Kolmogorov - Smirnov criterion [9], [10]), and criteria that are correct against pairs of different alternatives of a particular class (Dickson [11], Mathisen [12], Wald and Wolfowitz [13], Wilks [14], Wilcoxon [15], Mann-Whitney [16], etc.). Given the recent work that uses mixed principles of hypothesis testing, it is appropriate to generalize the classification and divide these criteria into two large groups, which could be called nonparametric and conditionally nonparametric criteria. The first group includes the criteria for testing the hypothesis of the equivalence of general populations that are nonparametric regardless of whether the distribution is continuous or discrete, and the second - those that depend on the distribution under certain conditions: nonparametric tests [11]– [16] and conditionally non-parametric criteria [17]– [28].

We suppose that data comes in portions of size n. Let us consider a sequence where we select two windows of lengths n and m:

$$\underbrace{x_1, x_2, ..., x_n}_{window\ 1}, \underbrace{x_{n+1}..., x_{n+m}}_{window 2}, ...$$

If the samples $(x_1, x_2, ..., x_n)$ and $(x_{n+1}, x_{n+2}, ..., x_{n+m})$ are heterogeneous, i.e. they have different distributions, then the point $x_n + 1$ is a change point.

The windows may have different sizes and starting points. As an initial point we can select two windows

with the same left end, let one window be fixed and the second window be moving with the step l (for example, $l = 1$). Intuitively clear, that when the sliding window move on homogeneous samples they have a high value of similarity measure, but when the sliding window begins to capture the data with a different distribution the similarity measure will decrease.

III. New test for change-point detection

On 1968 B.M.Hill, an American statistician, made an assumption (Hill's assumption $A_{(n)}$) [1], that if the sample $(x_1, x_2, \ldots, x_n) \in G$ is drawn using random sampling, such that sample values $x_1, x_2, \ldots, x_n$ are exchangeable random values with identical absolutely continuous distribution then

$$P\left(x_{n+1} \in \left(x_{(i)}, x_{(j)}\right)\right) = \frac{j-i}{n+1},$$

where x_{n+1} is a next sample value and $x_{(i)},\ x_{(j)}$ are a order statistics. This assumption was proved in papers of Yu.I.Petunin and their students: for i.i.d. random values in [2] and for exchangeable random values in [3].

Prof. Yu.I.Petunin introduced the p-statistics [4] which is a similarity measure between samples $x = (x_1, x_2, ..., x_n) \in G$ and $y = (y_1, y_2, ..., y_m) \in H$, and does not depends on their distribution. Suppose, that $F_G(x) \equiv F_H(x)$ and construct the variational series $x_{(1)} \leq ... \leq x_{(n)}$. Denote $A_{ij}^{(k)}$ an event that a sample value y_k lays between two order statistic $x_{(i)},\ x_{(j)}$: $A_{ij}^{(k)} = \left\{x_k \in \left(x_{(i)}, x_{(j)}\right)\right\}$. By the Hill's assumption $A_{(n)}$,

$$P(A_{ij}^{(k)}) = P\left(x_k \in (x_{(i)}, x_{(j)})\right) = p_{ij}^{(n)} = \frac{j-i}{n+1}.$$

Let us construct the Wilson confidence interval for an unknown probability of the event A_{ij}:

$$p_{ij}^{(1)} = \frac{h_{ij}^{(n,m)} m + g^2/2 - g\sqrt{h_{ij}^{(n,m)}(1-h_{ij}^{(n,m)})m + g^2/4}}{m+g^2},$$

$$p_{ij}^{(2)} = \frac{h_{ij}^{(n,m)} m + g^2/2 + g\sqrt{h_{ij}^{(n,m)}(1-h_{ij}^{(n,m)})m + g^2/4}}{m+g^2},$$

where $h_{ij}^{(n,m)}$ ai the frequency of $A_{ij}^{(n)}$ in m trials. Then, construct the confidence interval $I_{ij} = \left(p_{ij}^{(1)}, p_{ij}^{(2)}\right)$ with a significance level defined by the parameter g. Note, that when $g = 3$ the significance level of this interval does not exceed 0,05. Also, the value of p-statistics weakly depends on the choice of the confidence intervals for a binomial proportion [5].

Compute $N = n(n-1)/2$, where L is the number of intervals I_{ij} that contain the probabilities p_{ij}, $h = \rho(F_G, F_H) = \rho(x, y) = \frac{L}{N}$ is a similarity measure between samples x and y , i.e. p-statistics. Put $h_{ij} = h, m = N$ and $g = 3$, and construct a confidence interval $I = \left(p^{(1)}, p^{(2)}\right)$ for the probability $p(B)$. Hereinafter, we shall refer to the confidence intervals $I_{ij} = \left(p_{ij}^{(1)}, p_{ij}^{(2)}\right)$ and $I = \left(p^{(1)}, p^{(2)}\right)$ as intervals based on the $3s$-rule.

The scheme of trials where the events $A_{ij}^{(k)}$ can arise when the hypothesis that distributions are identical holds is called a generalized Bernoulli scheme [6]–[8]. If the hypothesis does not hold this scheme is called a modified Bernoulli scheme. In general case, when the hypothesis can be either true or false, i.e. $F_G(u) = F_H(u)$ or $F_G(u) \neq F_H(u)$, the trial scheme is called MP-scheme (Matveychuk–Petunin scheme) [8].

It was proved [4] that if in the generalized Bernoulli scheme holds the conditions 1) $n = m$; 2) $0 < \lim\limits_{n\to\infty} p_{ij}^{(n)} = p_0 < 1$ and 3) $0 < \lim\limits_{n\to\infty} p_q = \lim\limits_{n\to\infty} P(A_{ij}|H_0) = \frac{q}{n+1} = p^* < 1$, then the asymptotic significance level β of a sequence of confidence intervals $I_{ij}^{(n)}$ for the probabilities $p_q^{(n)} = p_{ij}^{(n)} = p\left(A_{ij}^{(n)}\right)$, based on the $3s$-rule, does not exceed 0,05.

Let $B_1, B_2, ..., B_n, ...$ be a sequence of events that can arise in a random experiment E, $\lim\limits_{k\to\infty} p(B_k) = p^*$, $h_{n_1}(B_1),\ h_{n_2}(B_2),\ ..., h_{n_k}(B_k), ...$ be a sequence of frequencies of the events $B_1, B_2, ..., B_n, ...$, correspondingly, and $\frac{k}{n_k} \to 0$ as $k \to \infty$.

We shall call an experiment E a strong random experiment if $h_{n_k}(B_k) \to p^*$ as $k \to \infty$. If in a strong random experiment samples $x = (x_1, ..., x_n) \in G$ and $y = (y_1, ..., y_m) \in H$ have the same size, then the asymptotical significance level of the interval $I^{(n)} = \left(p^{(1)}, p^{(2)}\right)$, based on the $3s$-rule as $g = 3$ using Wilson equations, does not exceeds 0,05.

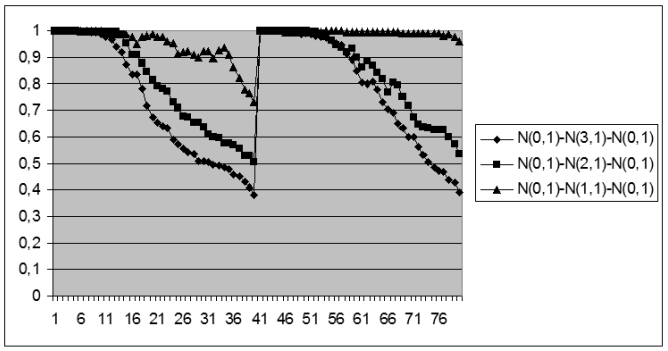
The test for the hypothesis on identity of distribution functions $F_G(u)$ and $F_H(u)$ with a significance level, that does no exceed 0,05 is following: if $I = \left(p^{(1)}, p^{(2)}\right)$ contains 0,95, then we accept the null hypothesis, else we reject the null hypothesis.

IV. Results

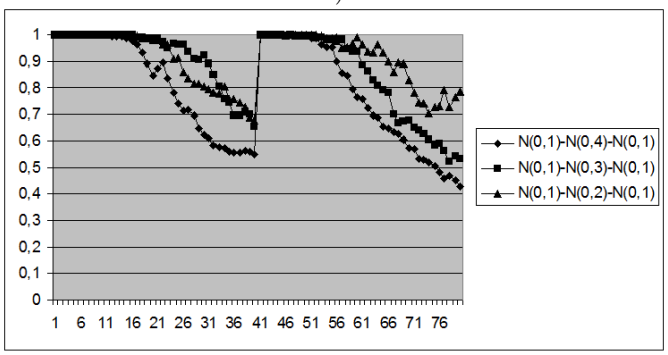
Here we use the following notation: N(m, v) is a normal distribution, where m is the mean and v is the standard deviation, U(a, b) is the uniform distribution on an interval (a, b), LN(m, v) is a lognormal distribution, where m is the mean and v is the standard deviation, P(m) is a Poisson distribution with a parameter m, E(l) is an exponential distribution with a parameter l, G(a, b) is a gamma distribution with parameters a and b.

For comparison we selected p-statistics and wide-used Kolmogorov-Smirnov statistics. The width of sliding window is equal to 40. The step of sliding is equal to 1. The averaging was made on 10 experiments. The results demonstrate the high effectiveness of the proposed method. Below we show the results of experiments with samples of size 40 from different distributions. We selected variants with different degrees of overlapping between samples.

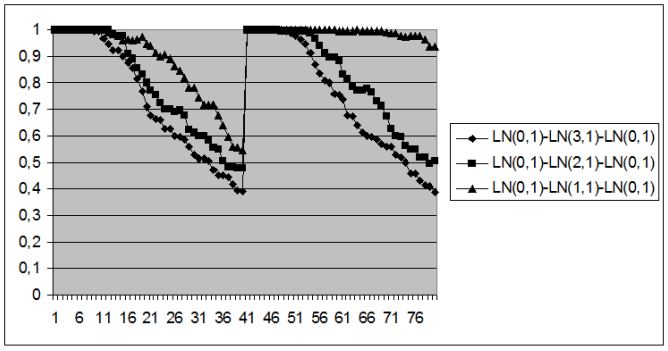
As we see, the Kolmogorov-Smirnov and Wilcoxon statistics as opposed to the p-statistics, in many cases demonstrate non-monotonic and nor-robust behaviour.



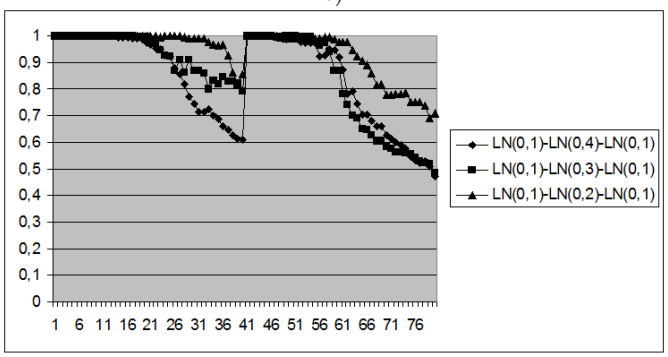
a)



b)

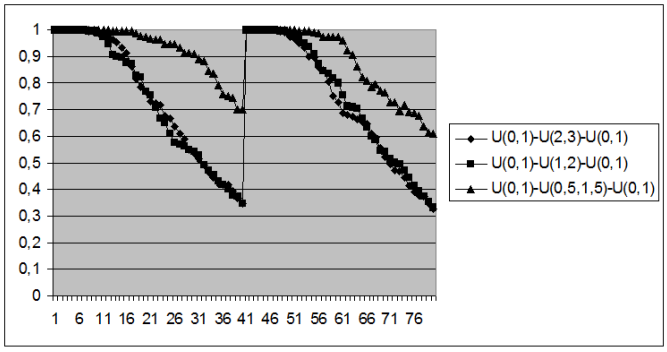


c)

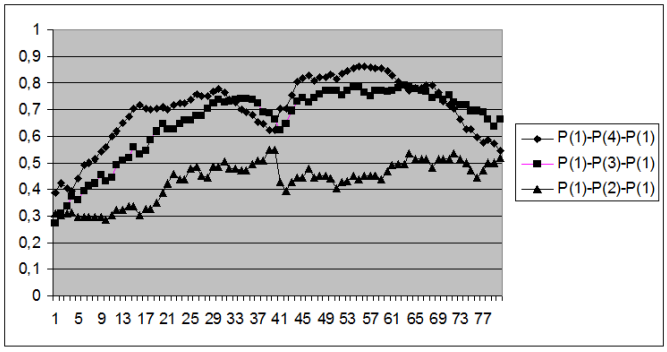


d)

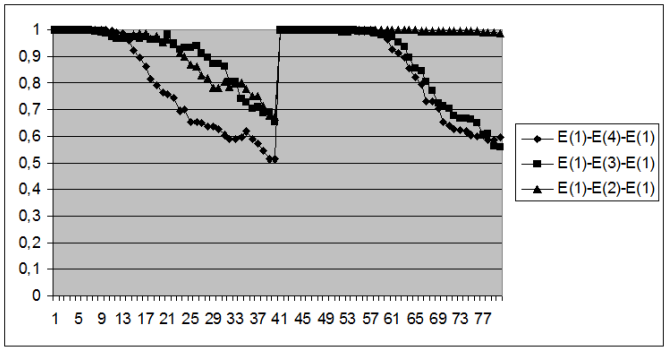
Fig. 1: P-statistics between the samples of size 40 from: a) normal distribution with different mean and the same standard deviation, b) normal distribution with the same mean and different standard deviation, c) lognormal distribution with the same mean and different standard deviation, d) lognormal distribution with the same mean and different standard deviation.



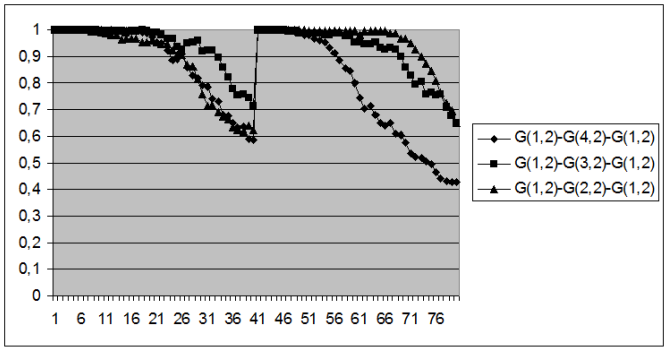
a)



b)

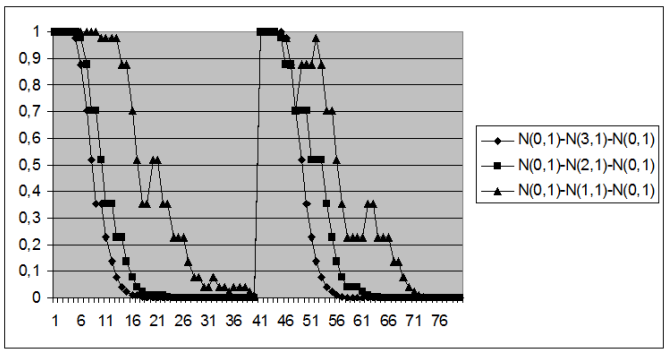


c)

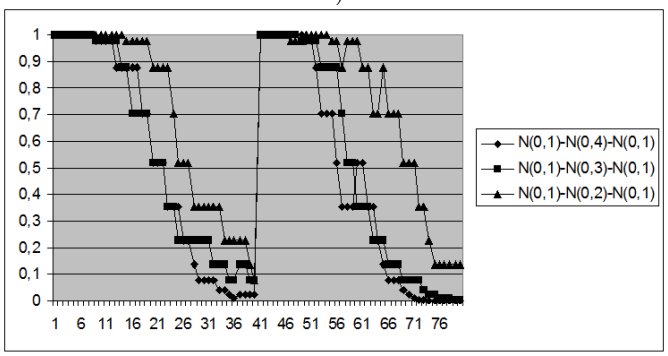


d)

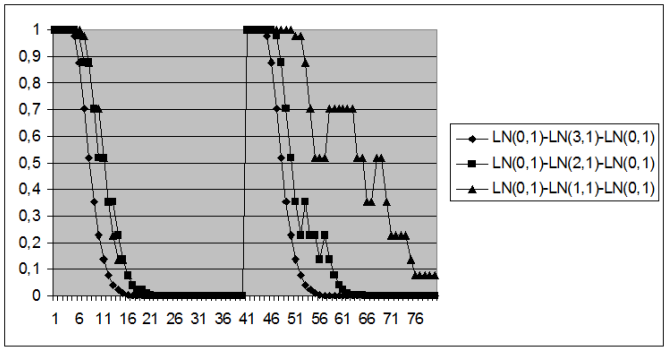
Fig. 2: P-statistics between the samples of size 40 from: a) uniform distribution on different intervals with various degree of overlapping, b) Poisson distribution with different rate parameters, c) exponential distribution with different rate parameters, d) gamma-distribution with the same shape and different scale characterizing different degree of overlapping



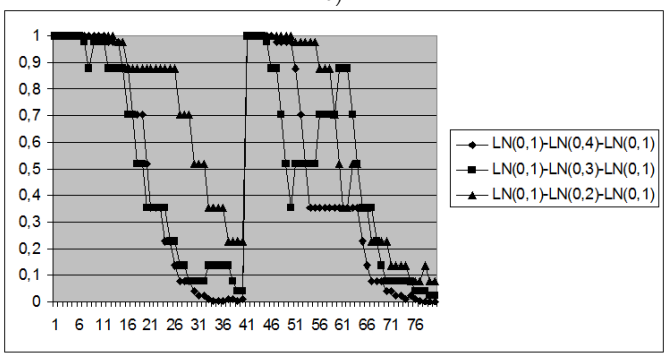
a)



b)

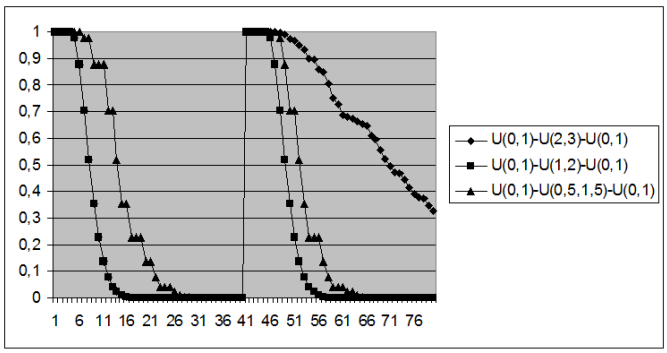


c)

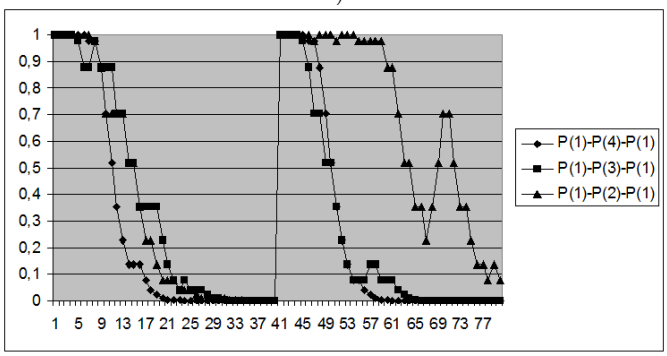


d)

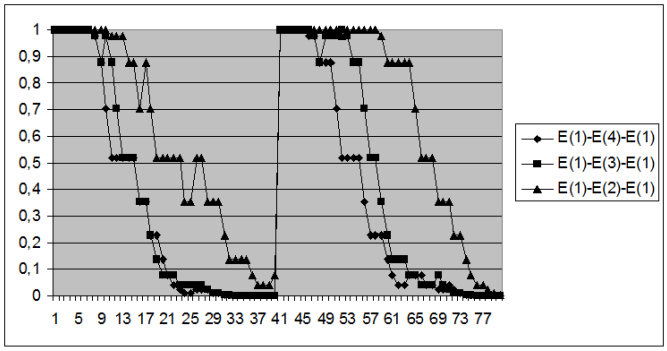
Fig. 3: Kolmogorov-Smirnov statistics between the samples of size 40 from: a) normal distribution with different mean and the same standard deviation, b) normal distribution with the same mean and different standard deviation, c) lognormal distribution with the same mean and different standard deviation, d) lognormal distribution with the same mean and different standard deviation.



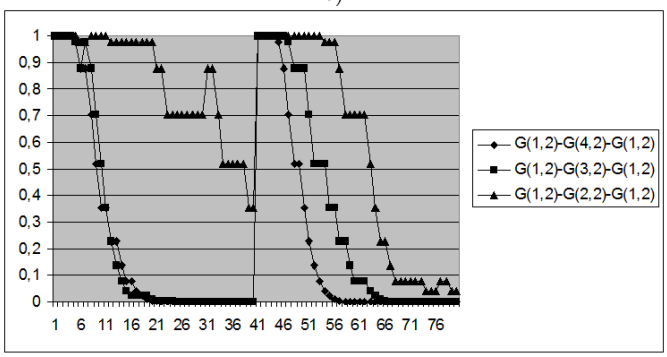
a)



b)

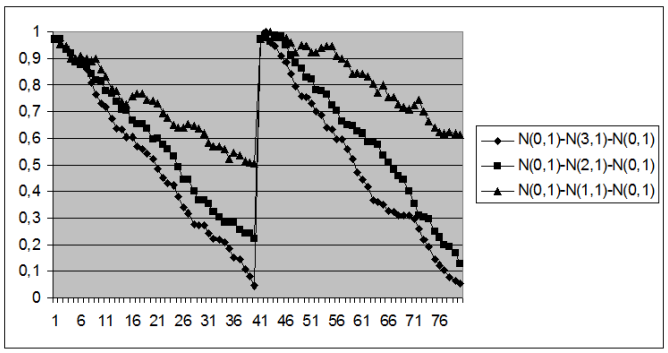


c)

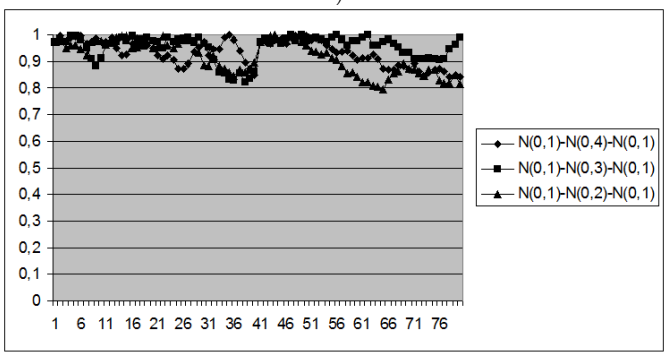


d)

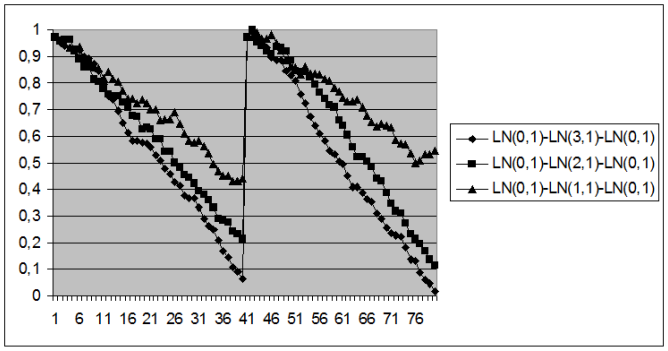
Fig. 4: Kolmogorov-Smirnov statistics between the samples of size 40 from: a) uniform distribution on different intervals with various degree of overlapping, b) Poisson distribution with different rate parameters, c) exponential distribution with different rate parameters, d) gamma-distribution with the same shape and different scale characterizing different degree of overlapping



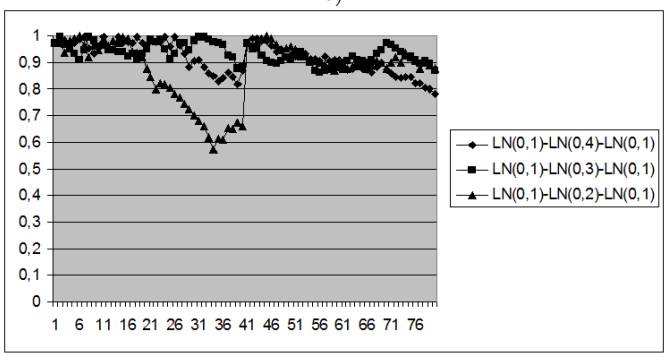
a)



b)

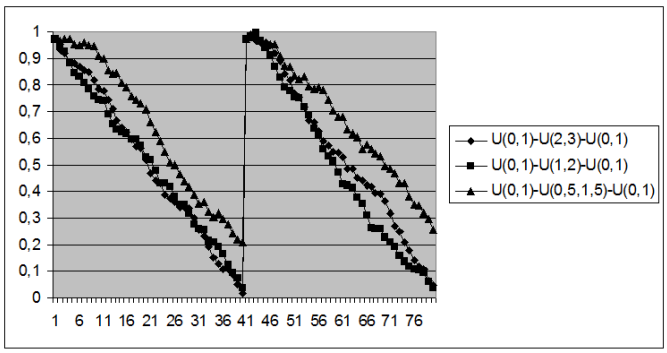


c)

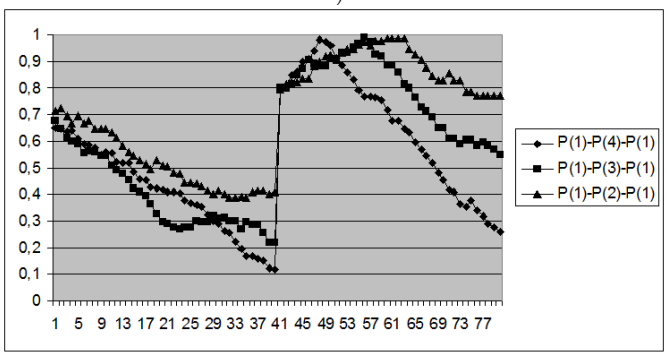


d)

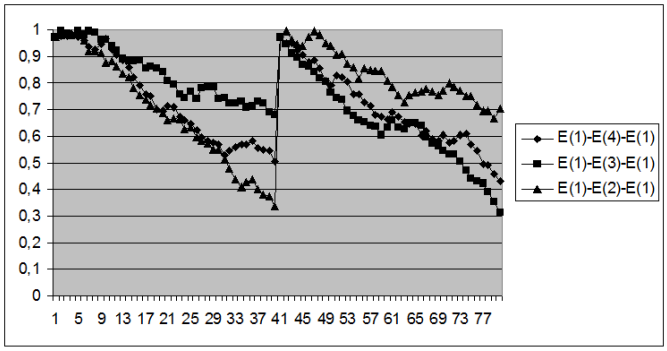
Fig. 5: Normed Wilcoxon statistics between the samples of size 40 from: a) normal distribution with different mean and the same standard deviation, b) normal distribution with the same mean and different standard deviation, c) lognormal distribution with the same mean and different standard deviation, d) lognormal distribution with the same mean and different standard deviation.



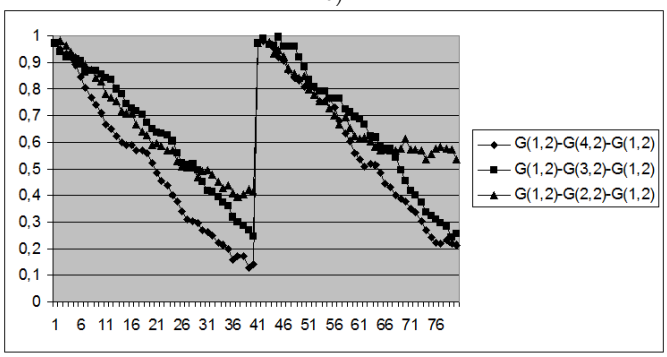
a)



b)



c)



d)

Fig. 6: Normed Wilcoxon statistics between the samples of size 40 from: a) uniform distribution on different intervals with various degree of overlapping, b) Poisson distribution with different rate parameters, c) exponential distribution with different rate parameters, d) gamma-distribution with the same shape and different scale characterizing different degree of overlapping

V. Conclusions

The new online algorithm implementing the Klyushin–Petunin test on streaming data is effective and sensitive. It does not depend on assumptions about distributions of samples except for the assumption that these distributions must be absolutely continuous. The significance level of the test does not exceed 0.05. It does not requires the special conditions for saving the data. The algorithm effectively solve the change-point and is more sensitive than alternative Kolmogorov–Smirnov and Wilcoxon statistics.

The experiments with samples from normal, lognormal, uniform, Poisson, exponential and gamma distribution with parameters describing different degrees of overlapping shows that p-statistics is more stable and monotonic than Kolmogorov-Smirnov and Wilcoxon statistics.

In particular, the Kolmogorov-Smirnov statistics demonstrates non-monotonic behavior due to the sensitiveness to outliers.

The Wilcoxon statistics is based on counting of inversions. So, it is quite effective for shifted samples drawn from the distributions with different means and same variance, but it is absolutely ineffective when the samples have large overlapping as in cases of the normal and lognormal distributions with the same mean and different variance.

Both Kolmogorov-Smirnov and Wilcoxon statistics allow to detect the change-point in data stream, but the p-statictics has better accuracy and robustness, because the p-statistics is very effective both for shifted samples drawn from the distributions with different means and same variance, and for samples with large overlapping (excepting the Poisson distribution).

The main feature of the p-statistics in these examples is the monotonicity depending on the degree of overlapping between the samples (the more overlapping the more p-statistics, and vice versa) and a sharp jump at the change-point.

References

[1] B.M. Hill, “Posterior distribution of percentiles: Bayes’ theorem for sampling from a population”, J Am Stat Assoc, vol. 63, pp.677–691, 1968.

[2] I. Madreimov and Yu. I. Petunin. “Characterization of a uniform distribution using order statistics”, Teor. Ver. Mat. Statist., Issue 27, pp. 96-102, 1982.

[3] R.I. Andrushkiw et al. “Construction of the bulk of general population in the case of exchangeable sample values”, In: Proceedings of the International Conference of Mathematics and Engineering Techniques in Medicine and Biological Science (METMBS’03), Las Vegas, Nevada, USA, June 26-29, 2003, pp.486–489.

[4] D. Klyushin and Yu. Petunin “A Nonparametric Test for the Equivalence of Populations Based on a Measure of Proximity of Samples”, Ukrainian Math. J., vol. 55, no. 2, pp. 181-198, 2003.

[5] D. Klyushin. S. Lyashko, S. Zub “A(n) Assumption in machine learning. In: Computational Linguistics and Intelligent Systems”, Lviv : Lviv Politechnic Publishing House, 2019. Vol 2 : Proceedings of the 3nd International conference, COLINS 2019, Workshop, Kharkiv, Ukraine, April 18-19, 2019, pp. 32-38.

[6] S. A. Matveichuk and Yu. I. Petunin, “Generalization of Bernoulli schemes that arise in order statistics”, I, Ukrainian Math. J., vol. 42, no. 4, pp. 459–466, 1990.

[7] S. A. Matveichuk and Yu. I. Petunin, Generalization of Bernoulli schemes that arise in order statistics”, II, Ukrainian Math. J., vol. 43, no. 6, pp. 728–734, 1991.

[8] N. Johnson amd S. Kotz, “Some generalizations of Bernoulli and Polya-Eggenberger contagion models”, Statist. Paper, vol. 32. p. 1–17, 1991.

[9] N.V. Smirnov, “Estimate of difference between empirical distribution curves in two independent samples”, Byull. Mosk. Gos. Univ, vol. 2, no. 2, pp. 3–14, 1939.

[10] N.V. Smirnov, “On the deviations of an empirical distriburion curve”, Mat. Sb., vol. 6(48), No. 1. pp. 3–26, 1939.

[11] W.G. Dixon, “A criterion for testing the hypothesis that two samples are from the same population”, Ann. Math. Stat., vol. 11, pp. 199–204, 1940.

[12] H.C. Mathisen “A method of testing the hypothesis that two samples are from the same population”, Ann. Math. Stat., vol. 14, pp. 188–194, 1943.

[13] A. Wald., J. Wolfowitz, “On a test whether two samples ate from the same population”, Ann. Math. Stat., vol. 11, pp. 147–162, 1940.

[14] S.S. Wilks, “A combinatorial test for the problem of two samples from continuous distributions”, In: Proc. Fourth Berkeley Symp. on Math. Stat. and Prob., vol. 1, University of California Press. pp. 707–717, 1961.

[15] F. Wilcoxon F. “Individual comparisons by ranking methods”, Biometrika, vol. 1, pp. 80–83, 1945.

[16] H.B. Mann, D.R. Whitney, “On a test of whether one of the random variables is stochastically larger than other”, Ann. Math. Stat., vol. 18, pp. 50–60, 1947.

[17] E.L. Lehmann, “Consistency and unbiasedness of certain nonparametric tests” Ann. Math. Stat., vol. 22, P. 165–179, 1947.

[18] M. Rosenblatt, “Limit theorems associated with variants of the von Mises statistic”, Ann. Math. Stat., vol. 23, pp. 617–623, 1952.

[19] M. Fisz, “On a result be M. Rosenblatt concerning the Mises-Smirnov test”, Ann. Math. Stat., vol. 31, pp. 427–429, 1960.

[20] D.L. Allen, “Hypothesis testing using L_1-distance bootstrap”, The American Statistician, vol. 51, pp. 145-150, 1997.

[21] E.J.G. Pitman, “Significance tests which may be applied to samples from any populations” J. Royal. Stat. Soc. Ser. A., vol. 4, pp. 119–130, 1937.

[22] M. Dwass “Modified randomization tests for nonparametric hypotheses”, Ann. Math. Stat., vol. 28, pp. 181–187, 1957.

[23] B. Efron and R.J. Tibshirani. “An Introduction to the Bootstrap”, vol. 57 of Monographs on Statistics and Applied Probability. - New York: Chapman - Hall, 1993.

[24] G.A. Barnard, “Comment on “The spectral analysis of point processes” by M.S. Bartlett”, J. Royal Stat. Soc. Ser. B., vol. 25, p. 294, 1963.

[25] Z.W. Birnbaum, “Computers and unconventional test-statistics”, In: F. Prochan and R.J. Serfling, eds., “Reliability and Biometry”. SIAM, Philadelphia, PA, pp. 441–458, 1974.

[26] R.W. Foutz, “ A method for constructing exact tests from test statistics that have unknown null distributions” J. Stat. Simul. Comp., vol. 10, pp. 187–193, 1080.

[27] K.-H. Jockel, “Finite sample properties and asymptotic efficiency of Monte Carlo tests”, Ann. Stat., vol. 14, pp. 336–347, 1986.

[28] J.-M. Dufour and A. Farhat, “Exact nonparametric two-sample homogeneity tests for possibly discrete distributions” Center for Interuniversity research in Quantitative Economics (CIREQ). Preprint 2001-23, 2001.

IEEE Third International Conference on Data Stream Mining & Processing
August 21-25, 2020, Lviv, Ukraine

Person Voice Recognition Methods

Oleksandr Tymchenko
Department of Safety Engineering, University of Warmia and Mazury
Olsztyn, Poland
o_tymch@ukr.net

Bohdana Havrysh
Department of Media Technology and Graphic Publishing
Ukrainian Academy of Printing
Lviv, Ukraine
dana.havrysh@gmail.com

Oleksandr O. Tymchenko
Department of Computer Science and Information Technology
Ukrainian Academy of Printing
Lviv, Ukraine
olexandr.tymch@gmail.com

Orest Khamula
Department of Computer science and Information technology
Ukrainian Academy of Printing
Lviv, Ukraine
khamula@gmail.com

Bohdan Kovalskyi
Department of Media Technology and Graphic Publishing
Ukrainian Academy of Printing
Lviv, Ukraine
bkovalskyi@ukr.net

Kateryna Havrysh
IT Step University
Lviv, Ukraine
gavrysh.kateryna@gmail.com

***Abstract*— This article is devoted to the selection and evaluation of language features used in the tasks of automatic text-independent speaker verification. The task of automatic speaker identification is one of the most difficult tasks in the field of language processing. The methods used in current speaker identification systems are not ideal, that imposes some restrictions on such systems. Some systems work well in good acoustic conditions, with minimal noise, but noticeably lose recognition accuracy in low signal-to-noise conditions. The requirements for the accuracy of speaker identification for such systems set a certain bar, which increases every year. Improving authentication accuracy extends the scope of such systems, including biometric multi-factor authentication systems, remote banking, access control systems, and more. Thus, to meet the needs of these systems consumers, the high requirements for the speaker recognition accuracy are required.**

Keywords — voice, recognition, model, coding, algorithms.

I. INTRODUCTION

The speaker recognition task involves two sub-tasks: identification and verification. Automatic speaker verification is the person confirmation by voice according to his / her ID (usually the speaker's name).

Several characteristics are used to assess the accuracy of speaker identification, one of which is the Equal Error Rate (EER), that is the most commonly used. An equal error of the first and second kind determines the error of the speaker recognition, provided that the likelihood of pretender missing and failure of the legitimate user. This characteristic is used to evaluate both text-dependent and text-independent speaker identification systems. The best speaker identification systems, tested on a fixed database containing phrases of several hundred speakers, show EER values of 3-5% [1, 2], tests are conducted at the National Institute of Standards and Technology (NIST). such accuracy is insufficient for current speaker identification systems.

On the one hand, the second kind of error can be considered as the most important, when the legal system user is taken as the impostor. It is possible to shift the system decision-making threshold towards reducing this error [3, 4, 5]. However, this will lead to an increase in first-class errors, that is, increase of the legal users denials frequency of a system access, which can cause frustration for users using the system. Thus, it is necessary to reduce these errors (the EER), which will reduce the likelihood confidential information loss when used in real banking and other systems.

The difference between automatic speaker identification is that, at first, an unknown speaker ID must determine who the speaker is - a legitimate user registered in the system or an intruder (in the case of an open identification task) [5, 6]. The system of automatic text-independent speaker verification presented in this paper solves the task of verifying a closed set of speakers by deciding whether or not the voice of the declared speaker is present on the audio. In this case, the existence of speakers that are not registered in the system is not taken into account.

In the case of automated systems, the functionality of the human-machine system access is possible with the help of a standard touch panel operator or modern contactless solutions, such as a speech signal. This is a natural, intuitive form of assignment and, as a rule, does not require long preparation. It can also be used for biometric purposes, for example, for speaker recognition [5, 7, 8]. This issue is important in automation systems because of the use of a particular machine by the operator, the high difficulty of circumventing security, especially in remote access systems, as well as the simplicity of assumptions (no passwords, physical markers, etc.) [9-12]. By using the speaker recognition, you can ensure that devices do not respond to random commands from outsiders, giving full control of the system.

An important factor affecting the effectiveness of speaker recognition (verification or identification) is the quality of transmission / recording of the speech signal. For public switched telephone networks and pulse code modulation, speaker recognition efficiency is up to 95% [2, 13]. However, transmitting a voice signal to a cellular network or the Internet through loss encoding algorithms reduces this efficiency, so methods that take into account lossy language encoding must be developed.

978-1-7281-3214-3/20/$31.00 ©2020 IEEE

II. VOICE IDENTIFICATION METHODS ANALYSIS

Typical voice recognition systems operate in two modes: training, which generates pronunciation models, and testing, during which the speech signal is compared with models from the database after appropriate people conversion. The decision is then made to identify the person being tested [14-16]. It is a common part of both modes to obtain special personality traits.

There are several methods of how to determine the features of speech.

At first, the person's laryngeal tone is calculated, and then when he passes the fundamental frequency verification, the speaker proceeds to the second recognition stage. Depending on the person characteristics, the larynx tone is in the range of 80 - 450 Hz.

Next the MFCC (Mel Frequency Cepstrum Coefficients) are extracted from the speech signal. The reason why the above coefficients were chosen is the need to reflect the human ear perception process, which perceives the frequency of the sound signal non-linearly-logarithmically.

Typically, the MFCC [17, 18] are mostly used. Briefly, the coefficients are determined by fragmenting the input into short segments (20-40 ms) overlapping the Hamming window and performing a Fourier transform. This produces a spectrum of signal power. This spectrum is passed through a comb of filters arranged uniformly on a chalk scale (13-25 MFCC) and perform a discrete cosine transform. The coefficients obtained give some characteristic of the frame. In the training mode, a model is generated from these coefficients, the most commonly used vector quantization algorithm (VQ, vector quantization) [19, 20] or the Gaussian mixture models [21], on which the GMM-UBM (Gaussian mixture models – universal background models) algorithm [22, 23]. The Gaussian mixture model assumes that our data is a mixture of multidimensional Gaussian distributions with defined parameters. Note, that in addition to the spectrum itself, the individuality of the voice is formed by speed and acceleration, simultaneously determine the first and second derivatives of MFCC (1):

$$\Delta c_m(q) = c_{m+\tau}(q) - c_{m-\tau}(q);$$
$$\Delta^2 c_m(q) = \Delta c_{m+\tau}(q) - \Delta c_{m-\tau}(q), \quad (1)$$

where the $q = \overline{0, Q-1}$, Q - the total number of language fragments.

This process increased the parameter vector. Delta and deltadelta MFCC coefficients represent their dynamic changes that identify individual speech units well. It should be clear that these two parameters are calculated for time intervals equal to two time moments. Too small intervals may not capture clear changes in MFCC coefficients, while large ones may determine the difference between too different states.

The similarity of the features set for each model is determined by the Euclidean distance or log-likelihood ratio.

A diagram of a typical voice recognition system is shown in Figure 1.

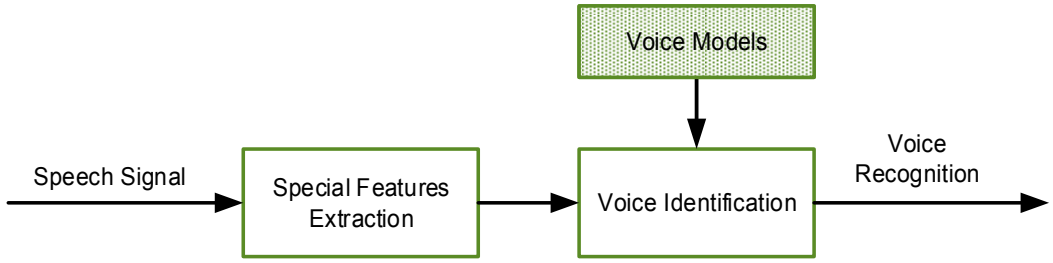


Fig. 1. Scheme of a typical voice recognition system

However, the systems based on the diagram below have some major drawbacks. The recognition result depends on the length of the signal, which should be quite long, at least 20 seconds (according to the literature). This value is not available for short-message control systems. Another typical recognition system disadvantage is the high sensitivity to the speech signal quality. The factor that reduces this quality in modern remote-control systems is the lossy encoding associated with transmitting voice signals over the GSM or Internet cellular network. An analysis of the literature on voice control systems shows that this problem is not yet sufficiently analyzed, and there are no studies that take into account currently used GSM encoders. Existing studies concern speech recognition only, they show the effect of encoding without problem solving and offer solutions that are physically impossible to implement due to lack of signal access before decoding, or show complex solutions difficult to implement in real time [23-26].

Voice recognition efficiency in embedded autonomous control systems is diminished by the use of fixed-point arithmetic in controllers.

They make it possible to reduce the power consumption of the embedded system and reduce its cost, but unlike systems using floating point arithmetic, they create more calculation errors. Hence, it is necessary to develop additional techniques to reduce these errors in order to avoid a decrease in voice recognition performance.

III. IMPROVEMENT OF AUTOMATIC VOICE RECOGNITION SYSTEMS

The first problem with voice recognition systems is the impact of speech duration on recognition performance. To increase the speed of correct identification, it is proposed to use algorithms for detecting signal activity, thereby maximizing the content of information in the analyzed signal. Algorithms based on signal energy, amplitude, higher order difference, and number of zero-crossing points were tested.

The second major issue is the impact of coding on recognition performance. It is advisable to divide this step into two: detection of signal encoding and the selected model influence on the efficiency of the decoded language recognition. A diagram containing proposed system enhancements to a typical solution is shown in Figure 2: VAD (Voice Activity Detection) – voice features determining based on MFCC frequencies, a speech encoder detector and its selection and voice recognition.

The third stage of the study was about implementation aspects related to improving voice recognition efficiency using fixed point arithmetic on the AVR controller (Fig.2).

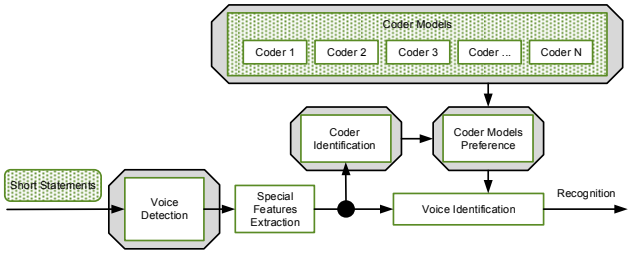


Fig. 2. Structural diagram of an advanced automatic voice recognition system

A. Voice bases and modeling algorithms that used

The authors used the TIMIT Voice Database [27], developed by Texas Instruments (TI) and the Massachusetts Institute of Technology (MIT), is a database of 630 people who speak 10 different short phrases in English.

During the study, the authors used three algorithms for speech signal modeling – vector quantization (VQ), a Gaussian mixture model (GMM) and a mixture of Gaussian distributions based on a common model of speech of the whole population (GMM-UBM).

To implement the identification algorithm, a GMM-UBM system is used. GMM (Gaussian Mixture Model) is a model of Gaussian mixtures, which will be a speaker model. With this approach, the initial data are presented in the form of clusters described by Gaussians.

The model of Gaussian mixtures is determined by the mathematical expectation vectors μ, the covariance matrix Σ, the weight vectors p and the number of components of the mixture M. To determine the first three values, training using the k-means and EM (Expectation Maximization) algorithms is used by the maximum likelihood method.

UBM (Universal Background Model) is a GMM trained on a relatively large amount of voice data. With the GMM-UBM approach, individual speaker models are trained using the MAP adaptation (Maximum A-Posteriori Adaptation). With this approach, mathematical expectations are shifted towards new data. The advantage of UBM is the quick adaptation of new speakers and the demands of a small amount of data for this.

To identify the speaker, you first have to find the model closest to the test recording (2)

$$S = \arg\max_{1<k<N} \sum_{t=1}^{M} \log P\left(\overline{x}_t \mid \lambda_k\right)$$

$$P\left(\overline{x} \mid \lambda\right) = \sum_{i=1}^{M} p_i b_i\left(\overline{x}\right)$$

$$p\left(i \mid \overline{x}_t, \lambda\right) = \frac{p_i b_i\left(\overline{x}_t\right)}{\sum_{k=1}^{M} p_k b_k\left(\overline{x}_t\right)}$$

$$b_i\left(\overline{x}\right) = \frac{1}{\left(2\pi\right)^{D/2}\left|\sum_i\right|^{1/2}} \exp\left[-\frac{1}{2}\left(\overline{x}-\overline{\mu}_i\right)^T \sum_i^{-1}\left(\overline{x}-\overline{\mu}_i\right)\right] \quad (2)$$

The $p_i, \overline{\mu}_i, \Sigma_i$ are the values of weights, mathematical expectations and covariance matrices of the model accordingly, and as the λ model is designated, is the feature vector. M is the number of components of the Gaussian mixture, D is the dimension of the feature vector.

After finding the closest model, you can assign the recording to a registered or unregistered speaker. To do this, we calculate the indicator (3):

$$\Lambda\left(X\right) = \log P\left(X, \lambda_{trg}\right) - \log P\left(X, \lambda_{UBM}\right) \quad (3)$$

Based on a comparison of this value with the threshold, a decision is made on the speaker under test.

The recognition efficiency and the time taken to generate the model of the speaker and to compare the model with the test statement were compared. The highest recognition efficiency was obtained for the GMM and GMM-UBM algorithms, in which the GMM algorithm was used in real time, due to the lower complexity and ability to quickly modify the speaker model database created.

IV. IMPACT OF VAD ON THE SHORT MESSAGE RECOGNITION EFFECTIVENESS

Voice Activity Detection (VAD) is the detection of voice activity in an input acoustic signal to separate active broadcasting from background noise or silence. Voice interpreted as noise can generate chipping, so the VAD system maximizes the amount of information in the signal [27].

This important processing step is used to minimize the degradation of voice message recognition from short control statements. Removal from silent or other signal fragments that have too low an energy or low-level noise character has a positive effect on the selection of only those specific speech signals that determine correct recognition at the testing stage. All four VAD algorithms gave approximately the same results (Fig.3).

The recognition efficiency has been investigated accordingly for four VAD algorithms: the signal energy algorithm, the Jang algorithm based on the signal amplitude, the Jang HOD (high-order differences) algorithm and the Jang ZCR (zero-crossing rate) [28].

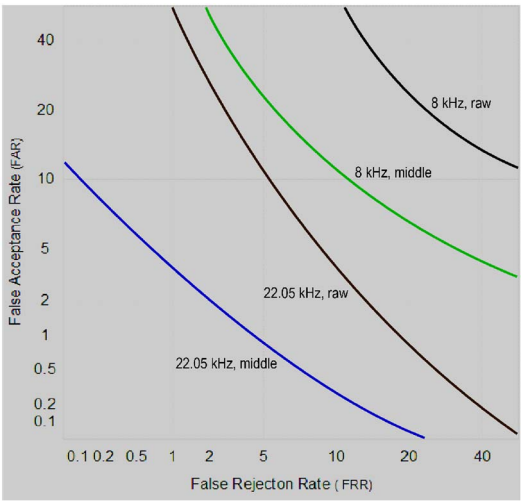


Fig. 3. FAR/FRR diagram – the recognition algorithm impact on the performance

CONCLUSIONS

These algorithms are tested in two ways – finding silence only at the beginning and end of the statement and in the full statement. The time of the speech signal activity detection for each algorithm for use in the real-time embedded system was estimated. The performance improvement achieved, measured by the Equal Error Rate (EER) quantitative measure, is up to 19%. The results in the form of graphs of the ratio of false non-recognition to the probability of false recognition FAR / FRR (False Acceptance Rate / False Rejection Rate) as a parameter of the encoding rate and detection of signal activity are presented in Figure 3. The word "middle" in the name refers to the middle of the expression. The greatest increase in efficiency was obtained for the transmission rate of 8 kbps, which testifies to the possibility of using voice quality recognition system (GSM).

ACKNOWLEDGMENT

The authors are appreciative to colleagues for their support and appropriate suggestions, discussing the results, reading the manuscript and criticism, which allowed to improve the materials of the article.

REFERENCES

[1] Jr.J.P. Campbell, "Speaker recognition: a tutorial", proc. of the IEEE, vol. 85, no. 9, 1997, pp. 1437–1462.

[2] M. Sahidullah and G. Saha, "A novel windowing technique for efficient computation of MFCC for speaker recognition", IEEE signal processing letters, vol. 20, no. 2, 2013, pp. 149–152.

[3] P. Motlicek et al., "Employment of subspace gaussian mixture models in speaker recognition", IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 4445–4449, 2015

[4] 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, 2015, pp. 439-441.

[5] C. S. Greenberg et al., "The NIST 2014 speaker recognition i-vector machine learning challenge", Odyssey: The Speaker and Language Recognition Workshop, 2014, pp. 224–230.

[6] Y. Lei et al., "A novel scheme for speaker recognition using a phonetically-aware deep neural network", IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2014, pp. 1695–1699.

[7] S.H. Shum et al., "Unsupervised clustering approaches for domain adaptation in speaker recognition systems", Odyssey: The Speaker and Language Recognition Workshop, 2014, pp. 265–272.

[8] T. Stafylakis et al., "Compensation for phonetic nuisance variability in speaker recognition using DNNs", Odyssey: The Speaker and Language Recognition Workshop, 2016, pp. 340–345.

[9] P. Kenny et al.. "Deep neural networks for extracting baum-welch statistics for speaker recognition", proc. Odyssey, pp. 293–298, 2014.

[10] D.A. van Leeuwen and R. Saeidi, "Knowing the non-target speakers: The effect of the ivector population for PLDA training in speaker recognition", IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2013, pp. 6778–6782.

[11] L. Xu et al., "Rapid Computation of I-vector" Odyssey: The Speaker and Language Recognition Workshop, 2016, pp. 47–52.

[12] K.S. Ahmad et al., "A unique approach in text independent speaker recognition using MFCC feature sets and probabilistic neural network", Advances in Pattern Recognition (ICAPR), 2015, pp. 1–6.

[13] M. McLaren, L. Ferrer and A. Lawson, "Exploring the role of phonetic bottleneck features for speaker and language recognition", IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2016, pp. 5575–5579.

[14] F. Richardson, D. Reynolds and N. Dehak, "Deep neural network approaches to speaker and language recognition", IEEE Signal Processing Letters, vol. 22, no. 10, 2015, pp. 1671–1675.

[15] N. Dehak et al., "Front-end factor analysis for speaker verification", IEEE Transactions on Audio, Speech, and Language Processing, vol. 19, no. 4, 2011, pp. 788–798.

[16] M.Nazarkevych, N. Lotoshynska, I. Klyujnyk, Y. Voznyi, S. Forostyna and I. Maslanych, "Complexity Evaluation of the Ateb-Gabor Filtration Algorithm in Biometric Security Systems", in 2019 IEEE 2nd Ukraine Conference on Electrical and Computer Engineering (UKRCON), pp. 961-964, July 2019.

[17] E. Variani et al., "Deep neural networks for small footprint text-dependent speaker verification", IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 4052–4056, 2014.

[18] S.B. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences", IEEE Transactions on Acoustics, Speech and Signal Processing, vol. 28, no. 4, 1980, pp. 357–366.

[19] B.S. Atal, "Automatic recognition of speakers from their voices", Proceedings of the IEEE, vol. 64, no. 4, pp. 460–475, 1976.

[20] D.Jurafsky and J.H. Martin, "Speech and Language Processing: second edition", Pearson Education, New Jersey, 2009.

[21] F. Eyben, F. Weninger, F. Gross and B. Schuller, "Recent developments in opensmile, the munich open-source multimedia feature extractor", proceedings of the 21st ACM international conference on Multimedia, pp. 835–838, 2013.

[22] D.A. Reynolds, "Gaussian mixture models", Encyclopedia of biometric recognition. Springer. Heidelberg, vol. 10, 2008, pp. 19–41.

[23] D.A. Reynolds and R.C. Rose, "Robust text-independent speaker identification using Gaussian mixture speaker models", IEEE Transactions on Speech and Audio Processing, vol. 3, no. 1, 1995, pp. 72–83.

[24] D.A. Reynolds, T.F. Quatieri and R.B. Dunn, "Speaker verification using adapted Gaussian mixture models", Digital Signal Processing, vol. 10, no. 1, 2000, pp. 19–41.

[25] S.O. Sadjadi, M. Slaney and L. Heck, "MSR identity toolbox v1.0: A MATLAB toolbox for speaker-recognition research", Speech and Language Processing Technical Committee Newsletter, 2013.

[26] Z. Fanfeng, "Application Research of Voice Control in Reading Assistive Device for Visually Impaired Persons", International Conference on Multimedia Information Networking and Security, Nanjing, Jiangsu, pp. 14-16, 2010.

[27] S. Park, Y. Park, A. Nasridinov and J. Lee, "A Person Identification Method in CUG Using Voice Pitch Analysis" IEEE Fourth International Conference on Big Data and Cloud Computing, Sydney, NSW, pp. 765-766, 2014.

[28] A. Paul, M. Panja, M. Bagchi, N. Das, R. M. Mazumder and S. Ghosh, "Voice recognition based wireless room automation system", International Conference on Intelligent Control Power and Instrumentation (ICICPI), Kolkata, pp. 84-88, 2016.

IEEE Third International Conference on Data Stream Mining & Processing
August 21-25, 2020, Lviv, Ukraine

Elastic Scale-Free Networks Model Based on the Mediaton-Driven Attachment Rule

Vadim Shergin
Artificial Intelligence dept.
Kharkiv National University of Radio Electronics
Kharkiv, Ukraine
vadim.shergin@nure.ua
ORCID 0000-0002-4388-8180

Larysa Chala
Artificial Intelligence dept.
Kharkiv National University of Radio Electronics
Kharkiv, Ukraine
larysa.chala@nure.ua
ORCID 0000-0002-9890-4790

Serhii Udovenko
Informatics and Computer Engineering dept
Simon Kuznets Kharkiv National University of Economics
Kharkiv, Ukraine
serhiy.udovenko@hneu.net
ORCID 0000-0001-5945-8647

Mariya Pogurskaya
Artificial Intelligence dept.
Kharkiv National University of Radio Electronics
Kharkiv, Ukraine
mariia.yerovenko@nure.ua

Abstract — The problem of modeling scale-free networks is considered. Mediation-driven attachment rule with non-constant number of added links is proposed. Applying the copy-factor as the control parameter instead of traditionally used number of incoming links makes it possible to generate networks model, which combines the key benefits of mediation-driven attachment, such as indirect using the nodes degree, and elasticity, such as different relative growth rates for links and nodes. The properties of proposed networks model were studied analytically and confirmed by simulation results.

Keywords — scale-free network; elasticity; mediation-driven attachment; copy factor; Yule-Simon distribution

I. INTRODUCTION

The scale-free network's theory is a relatively new science area that studies networks whose degree distribution of nodes follows to a power law [1]. Many of real-world networks are scale-free, for instance social networks, collaboration networks, citations of scientific papers, World Wide Web, neural and protein networks, some transportation networks, etc [2]. On the other hand, power law distribution is a characteristic feature of fractal properties of the system and its abilities to self-organization. This is another reason for the theoretical interest to the scale-free networks.

At the heart of most of today's scale-free network models is the model proposed by Barabási and Albert in 1999 (BA-model) [3]. It is based on two fundamental concepts: concept of growth and concept of preferential attachment: the probability for a new node to link with some existing one is proportional to its degree. Although at present there exists a lot of specifications and generalizations for the generative mechanism used to create network models [4-12], their common features ascends to BA-model.

The concept of implicit use of node statistics, implied in the mediation-driven attachment (MDA) model, recently proposed by Hassan et al [13], and the concept of elasticity, i.e. different relative growth rates of links and nodes [14-16], are the subject of interest. Applying the elasticity concept to the MDA model forms the object of current paper.

II. FORMAL PROBLEM STATEMENT

According to the concept of growth, the seed network comprises n_0 nodes and $L_0/2$ edges, and then at each time step, a new node is added to the network. Thus, network size n (i.e. number of nodes) generally is used as time measure. In addition, the node number (i) is considered as the time of its birth (i.e. $n-i$ is the node's age).

Aside from the time of birth, each node is characterized by its degree (d_i) and other properties P_i (fitness, etc). Network growth dynamics is determined by the number of links $m(n)$ that connect the incoming node with existing ones, or by the average degree of nodes $\bar{k}(n)$. The network may also have other control parameters N (such as additional attractiveness, elasticity etc).

Network properties are substantially determined by the rule of attachment, i.e. the probability π_i that a new edge points to node i, as a function of the above parameters:

$$\pi_i = f(i, k_i, n_0, L_0, P_i, N). \quad (1)$$

A key property of the scale-free networks is the distribution of the nodes by degree $p(k)$ that should, at least asymptotically, follow the Yule-Simon law (which is a discrete analogue of the power law):

$$p(k) = C\frac{\Gamma(k)}{\Gamma(k+\gamma)} \propto k^{-\gamma}, \quad k \geq k_0 . \quad (2)$$

The parameter $\gamma \geq 2$ is called the scale factor.

The problem is to find the attachment rule (1), in which k_i degree is used implicitly and which allows us to generate a scale-free (2) network under condition of a power law of growth of an average degree of nodes.

III. SCALE-FREE NETWORK MODELS OVERVIEW

The simplest and most known model for scale-free networks is the BA model [3]. It is based on two

fundamental concepts: concept of growth and concept of preferential attachment, according to which each incoming node connects to $m = const$ existing ones. The probability π_i that a new edge points to some node i is proportional to its degree k_i:

$$\pi_i = k_i / \sum_i k_i . \quad (3)$$

Thus, BA attachment rule is the simplest case of a general form (1). This model leads to the Yule-Simon distribution for nodes degree (2) with scaling factor $\gamma = 3$.

According to the fitness model [4], each node is associated with its fitness value η_i. Therefore, rule (3) is extented into $\pi_i = f(k_i, \eta_i) = k_i \eta_i / \sum k_i \eta_i$. There also exists more complicated variants of weighting [5-7]. This approach allows to reproduce the degree correlation among network nodes.

Another extensions of BA model are based on aging [6], i.e. using the attachment rule (1) in form $\pi_i(n) = f(i, k_i(n))$.

In the Price model [2] the network is supplied by global additional attractiveness parameter a, and attachment rule has the form $\pi_i = (k_i + a) / \sum (k_i + a)$. This approach was used for modelling citation networks, which are considered as directed. Therefore, in the attachment rule k_i denotes only in-links. Clearly, if $a = 0$ (i.e. if the BA rule (3) is used), the incoming paper having any reference yet, cannot be cited.

At now, there exists many other specifications and generalizations for the generative mechanism (1) used to create network models (based on nonlinear preferential attachment [8-9], rewiring [10], second level of neighborhood data [11] etc).

Not all of the above models produce scale-free networks. In fact [9], the only case in which the topology of the network is scale free is that in which the preferential attachment is asymptotically linear under k_i, i.e. $\pi_i \sim a_\infty k_i$. On the other hand, despite the wide occurrence, scale-free networks are not the only class of the real-world networks [12].

Beeing the core of scale-free network models, the BA-model bears some original sins inherited by most of descending models that limit its application area:

- In real networks scaling factor γ not a constant, but assumes a spectrum of values between $2 < \gamma \le 3$. Processes and phenomenon's having $\gamma \le 2$ also exists [1-2];
- In real-word networks, an average degree of nodes tends to grow with network size, while it is constant in BA-model. Thus, this model generates very sparse networks;
- BA model produce networks which are neutral in assortativity [10, 14], while social networks are strongly assortative and biological and technical ones are disassortative;
- Information about nodes degree is used in a direct way;
- Finally, it does seem strange that model of scale-free network is based on parameter m which is scale-dependent.

To overcome the above restrictions it was proposed to expand the list of basic concepts by a new one: concept of elasticity [15-16], i.e. different relative growth rates of links and nodes. According to this approach, elasticity factor λ as the ratio of the relative growth rates of links and vertices is used:

$$\lambda = \frac{\delta L(t)}{\delta n(t)} = \frac{\Delta L(t)}{L(t)} \Big/ \frac{\Delta n(t)}{n(t)} = n \frac{L(n+1) - L(n)}{L(n)} . \quad (4)$$

If $1 \le \lambda < 2$ is fixed, the total number of links grows asymptotically by power law with degree λ:

$$L(n) = 2 \frac{\Gamma(n + \lambda - 1)}{\Gamma(n-1)\Gamma(\lambda + 1)} \propto n^{\lambda} . \quad (5)$$

Also the average degree of nodes is not a constant m, but

$$\bar{k}(n) = L(n) / n \propto n^{\lambda - 1} . \quad (6)$$

If $\lambda > 1$, then each new node brings more links than the current average (6). For example, the more scientific papers have been written, the more references should review an author of a new paper.

Therefore, using the concept of elasticity generalizes the original BA model and makes it possible to generate dense networks and networks with scaling factor $\gamma = 1 + 2/(2 - \lambda)$, which is not less, than 3.

The explicit use the information about nodes degree forms another essential restriction of both BA model and inherited models. In real-world networks, the incoming node cannot assess the information for the whole of existing nodes because of the large size of the network, confidentiality reasons etc. Really, for a first choose a site or scientific paper, we are not interested how many references to other sites or papers it has. Similarly, a newcoming businessman cannot reach the information about who has how many trade links in order to choose his trading partner.

The mediation-driven attachment (MDA) model was proposed by Hassan et al. [13] to overcome this restriction. It can be regarded as a version of the earlier copy-model [17].

According to MDA rule, each incoming node chooses some existing one uniformly at random. This node is called mediator. Then incoming node links at random to m neighbors of this mediator. Returning to the above example, after visiting a mediator-website, or reading a mediator-paper, one may want to follow some available links.

The probability π_i for node i having degree k_i to attach with incoming one can be found by the following way: let the neighbors of node i are labelled as $1, 2, ..., k_i$. The probabilities for each of them to be chosen as mediator are

equal to $1/n$. The probability for node i to be reached from mediator j is $1/k_j$. Thus,

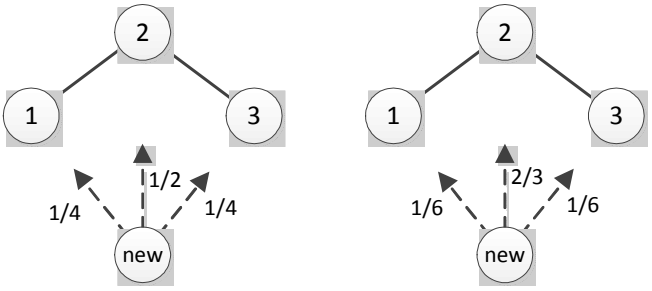
$$\pi_i = \frac{1}{n}\sum_{j=1}^{k_i}\frac{1}{k_j} = \frac{k_i}{n}\left(\frac{1}{k_i}\sum_{j=1}^{k_i}\frac{1}{k_j}\right). \quad (7)$$

As per (7), the probability for incoming node to connect with some existing ones also depends only on their degrees but in contrary to BA models this dependence is implicit.

According to results of numerical simulation from the original paper [13], the distribution for nodes degree asymptotically also follows to a power law (2) with scaling factor

$$\gamma \approx 3(1-0.43531e^{-0.11\cdot m}). \quad (8)$$

Comparison among BA and MDA rules is illustrated on Fig. 1. As per (1), in BA model the incoming node knows that $k_1 = k_3 = 1$, $k_2 = 2$. Thus, $\pi_i = (\frac{1}{4}, \frac{1}{2}, \frac{1}{4})$ respectively. According to MDA rule, the incoming node links with node 1 only if node 2 is picked as mediator (that happens with probability 1/3) and node 1 is chosen among the neighbors of node 2 (probability is 1/2). Thus, $\pi_1 = \frac{1}{6}$. Also $\pi_3 = \frac{1}{6}$. Node 2 is pointed if node 1 or 3 are picked as mediator, so $\pi_2 = \frac{2}{3}$. As it was mentioned in [13], for small m the MDA rule leads to super-preferential attachment, i.e. winners-take-all effect occurs. However, for large m this effect is replaced by winners-take-some effect that gives rise to simple hubs only.



BA attachment rule MDA attachment rule

Fig. 1. Comparison among BA and MDA attachment rules for m=1

For most real-world networks, the MDA rule looks much more natural than BA, but it also has some essential drawbacks. At the first and foremost, parameter m is considered as constant, that is obviously contradicts with the properties of real-world networks. Secondly, mediator node never connects by itself with incoming node. In other words, mediator is too latent that is not very common for real-world networks.

Summarizing, MDA rule is the most promising for scale-free network modelling, but to overcome restrictions arising with constancy of the number of added links, we propose to apply the elasticity concept to the MDA rule.

IV. Modification of the MDA rule based on the use of elasticity

According to the above analysis, we propose to modify the MDA rule by using copy-factor $0 \le q \le 1$ – fraction of mediator's neighbors that are linked with incoming node – as the control parameter instead of m. Mediator is always linked, so it is not latent, but explicit (Fig. 2). Thus, m is treated not as a constant, but as

$$m = 1 + q \cdot k_{med}. \quad (9)$$

where k_{med} is mediator's degree.

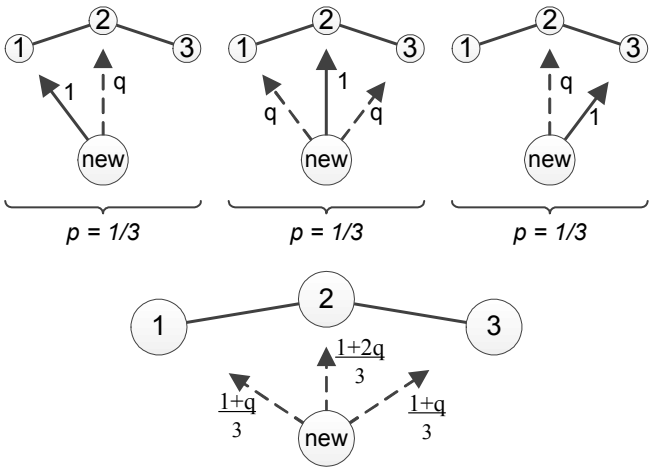


Fig. 2. Probabilities for new node to be attached according to proposed elastic MDA rule

According to (9), the expected value of the number of incoming edges at time n is

$$E\{m(n)\} = 1 + q \cdot E\{k_{med}\} = 1 + q \cdot \bar{k}(n) = 1 + q\frac{L(n)}{n}. \quad (10)$$

Thus, the dynamics equation of total link growth has the form

$$L(n+1) - L(n) = 2(1 + q \cdot L(n)/n). \quad (11)$$

Taking into account the initial conditions $L(1) = 0$, one can obtain the dependence of total number of links $L(n)$ and the average nodes degree $\bar{k}(n)$ on network size n:

$$\begin{aligned} L(n) &= \frac{2}{2q-1}\left(\frac{\Gamma(n+2q)}{\Gamma(n)\Gamma(1+2q)} - n\right), \\ \bar{k}(n) &= \frac{2}{2q-1}\left(\frac{\Gamma(n+2q)}{\Gamma(n+1)\Gamma(1+2q)} - 1\right), \end{aligned} \quad (12)$$

where $\Gamma(x)$ is the Euler's Gamma function.

As per (5)-(6) and (12), the modified MDA rule-based network is elastic with elasticity factor $\lambda = 2q$. The average degree of nodes is growing up to limit $\bar{k}_{\lim} = 2/(1-2q)$ for $0 < q < 0.5$, or by asymptotically power law $\bar{k}(n) \propto n^{2q-1}$ for $0.5 < q < 1$. Special case of $q = 0.5$ will be analyzed later.

The next problem is the distribution of nodes by degrees the proposed rule will lead to?

The incoming node $(n+1)$ attaches to existing one $1 \le i \le n$ with probability 1 given that i is mediator, and with probability q if mediator j is connected with i:

$$\Pr(i, n+1) = 1 \cdot \Pr(med = i) + \sum_{j \sim i} q \cdot \Pr(med = j) = \frac{1 + q \cdot k_i(n)}{n}. \quad (13)$$

Therefore, the dynamics equation for the degree of the node i has the form

$$k_i(n+1) - k_i(n) = \frac{1 + q \cdot k_i(n)}{n}. \quad (14)$$

It has the general solution

$$k_i(n) = C_i \frac{\Gamma(n+q)}{\Gamma(n)} - \frac{1}{q}. \quad (15)$$

Paramerers C_i can be found under condition that expected value for the degree of node i at time i is equal to the expected number of edges added at time $i-1$ (10): $k_i(i) = E\{m(i-1)\}$. Therefore,

$$C_i = \frac{1}{2q-1}\left(\frac{\Gamma(i-1+2q)}{\Gamma(i+q)\Gamma(2q)} - \frac{\Gamma(i)}{\Gamma(i+q)} \frac{1-q}{q} \right). \quad (16)$$

According to (16), the distribution of nodes by degrees is formed by the value of q. Given that $i \gg 1$, parameter C_i asymptotically grows as

$$C_i \propto \begin{cases} \frac{1}{(2q-1)\Gamma(2q)} i^{q-1}, & 0.5 < q < 1 \\ \frac{q-1}{(2q-1)q} \cdot i^{-q}, & 0 < q < 0.5 \end{cases} \quad (17)$$

In fact, the number of node (i) is treated as a rank by degree. So, as per (17), rank distribution is asymtotically follows to a power law with scaling factor $\beta = \min\{q, 1-q\}$. It corresponds to an asymptotically power distribution of nodes by degree (2) with scaling factor

$$\gamma \propto 1 + \frac{1}{\beta} = 1 + \frac{1}{\min\{q, 1-q\}}. \quad (18)$$

According to (18), the degree of "young" nodes (i.e. nodes with a large number) follows to power distribution with scaling factor $\gamma > 3$.

V. Special Cases of the Modified MDA Rule

As it follows from above analysis, properties of the proposed model of scale-free networks based on modified MDA rule are determined by the control parameter q that is a probability for mediator's neighbors to be linked with newcoming node. There are three special values of this parameter, which give rise to special models.

- In the case $q = 0$, the proposal rule leads to Callaway growing network [10, 18]. Each time step one node is added to the network and attached to some existing at random. Thus, $m = 1$ and $L(n) = 2(n-1)$. The range distribution of nodes by degree follows not to the power law (15)-(16), but to the logarithmic distribution:

$$k_i(n) = 1 + H_{n-1} - H_{i-1} \approx 1 + \log(n/i), \quad (19)$$

where H_n is n-th harmonic number.

Therefore, the distribution for nodes degree $p(k)$ is not a power law (2), but exponential, so this model is not scale-free.

- In the case $q = 1$, each incoming node connects to all of the existing one, forming a full graph structure.
- Finally, the most interesting case is $q = 1/2$. The dependences of total number of links $L(n)$ and the average nodes degree $\bar{k}(n)$ on network size has the form not (12), but

$$\begin{aligned} L(n) &= 2n(H_n - 1) \propto 2n \cdot \log(n), \\ \bar{k}(n) &= 2(H_n - 1) \propto 2 \cdot \log(n). \end{aligned} \quad (20)$$

The range distribution of nodes by degree follows to the subpower law (15) with

$$C_i = \frac{\Gamma(i) \cdot (H_{i-1} + 2)}{\Gamma(i + 1/2)} \propto \frac{\log(i)}{\sqrt{i}}. \quad (21)$$

A detailed study of the properties of this model forms one of the subjects of interest for future research.

VI. Numeric Simulation

The numerical experiments were provided. For the first step, we simulate networks based on the proposed elastic MDA rule to check their properties of growth. Network starts from two nodes, connected by edge. Each step one nodes were added and linked to the network according to proposed rule by some links (9). The dependence of average nodes degree under network size n and copy-factor q were studied. The simulated results and the corresponding theoretical dependences (12) are shown on Fig. 3.

For the next step, we checked the dependence for nodes of its degree on age. $M = 20$ networks of $n = 4096$ nodes were simulated and results (nodes degree, k_i) were averaged. The obtained numerical dependences along with the theoretical ones (15)-(16) are shown on Fig.4. It is easy to see that they are fully consistent to each other.

Rank distributions of nodes degree are shown on Fig. 5. As one can see, for about a quarter of the total number of nodes, the rank distributions fit to the predicted Yule-Simon laws, but the last about a quarter of nodes forms a tail, having much less links, than predicted. Asymptotical nature of the above theoretical formulas may be a cause. A more accurate analysis of the proposed attachment rule is needed for future research.

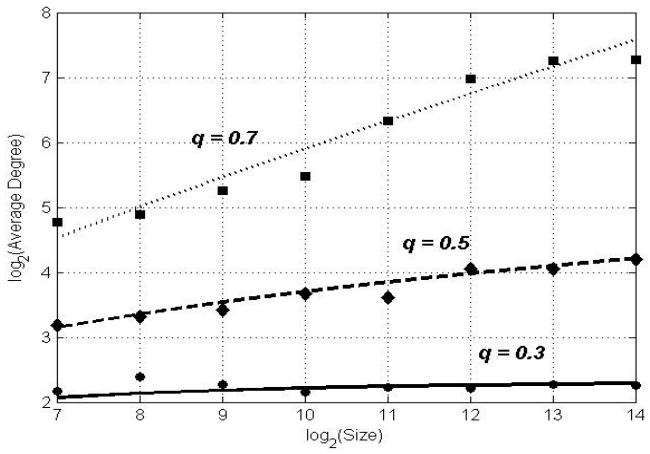


Fig. 3. Dependence of average nodes degree on network size for q=0.3, q=0.5 and q=0.7

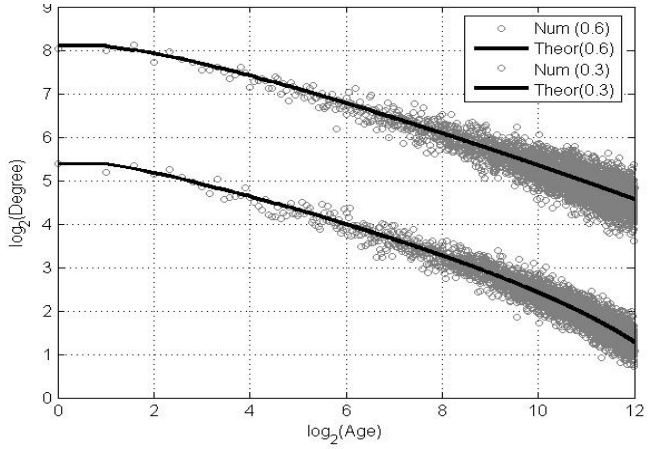


Fig. 4. Dependences of nodes degree on its age for q=0.6 and q=0.3

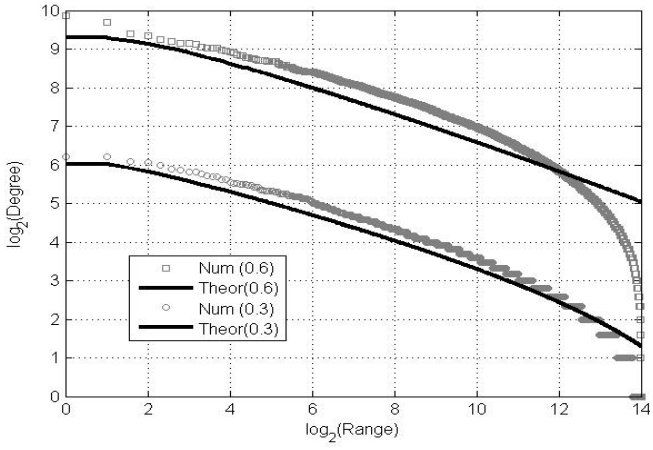


Fig. 5. Rank distribution of nodes degree for q=0.6 and q=0.3

CONCLUSION

The problem of modeling scale-free networks is considered. The properties of generated network are strongly depends on the attachment rule. An actual models and corresponding attachment rules were analyzed.

The mediation-driven attachment rule has an undoubted advantage due to indirect using of nodes statistics. However, using a constant number of incoming links as a contol parameter limits the scope of such model. Applying the copy-factor as the control parameter makes it possible to generate networks model, which combines the key benefits of mediation-driven attachment model, and elastic ones, such as different relative growth rates for links and nodes.

Using differentiation in relative growth rates of links and nodes make it possible to generate scale-free models for dense networks. In addition, copy-factor is internal character of the network, while an elasticity factor is external.

The dependence of average nodes degree on copy-factor and network size was obtained analytically. In addition, the dependence of nodes degree under its age was found. Numerical simulations were provided. The obtained results are in consistency with the theoretical ones.

A detailed study of the properties of proposed model at some special cases of copy-factor, as well as analysis of its assortativity properties, forms the subject of interest for future research.

REFERENCES

[1] M.E.J. Newman, “Power laws, Pareto distributions and Zipf's law,” Contemporary Physics, 46(5), 2005, pp. 323-351.

[2] S. N. Dorogovtsev and J.F.F. Mendes, “Evolution of Networks: From Biological Networks to the Internet and WWW”, Oxford, USA: Oxford University Press, 2003.

[3] R. Albert and A.-L. Barabasi, “Statistical mechanics of complex networks”, Rev. Mod. Phys, vol. 74, p. 42-97.

[4] G. Bianconi and A.-L. Barabási, “Competition and multiscaling in evolving networks”, Europhysics Letters, 54 (4), pp. 436–442. doi:10.1209/epl/i2001-00260-6

[5] S.N. Dorogovtsev, J.F.F. Mendes and A.N. Samukhin, “Structure of growing networks with preferential linking”, Phys. Rev. Lett., 85, 4633.

[6] G. Caldarelli and M. Catanzaro, “Networks: A Very Short Introduction”, Oxford University Press. p. 78.

[7] K. Choromański, M. Matuszak and J. MięKisz, “Scale-Free Graph with Preferential Attachment and Evolving Internal Vertex Structure”, Journal of Statistical Physics, 151 (6), pp. 1175–1183.

[8] J. Kunegis, M. Blattner and C. Moser, “Preferential Attachment in Online Networks”, Measurement and Explanations. WebSci13 ACM Web Science Conference, 2013. doi: 10.1145/2464464.2464514.

[9] P. L. Krapivsky, S. Redner and F. Leyvraz, “Connectivity of Growing Random Networks”, Phys. Rev. Lett, no. 85, 2000, pp. 4629–4632. doi: 10.1103/physrevlett.85.4629.

[10] R. Noldus and P. Van Mieghem, “Assortativity in complex networks”, J. Complex Networks, vol. 3, 2015, pp. 507-542.

[11] Ch. Dangalchev, “Generation models for scale-free networks”, Physica A 338, 2004, p. 659.

[12] A. Broido and C. Aaron, “Scale-free networks are rare”, Nature Communications, 10(1), 2019, p. 1017. doi:10.1038/s41467-019-08746-5

[13] M.K. Hassan and L. Islam, “Arefinul Haque, Syed Degree distribution, rank-size distribution, and leadership persistence in mediation-driven attachment networks”, Physica A. 469, 2017, pp. 23–30. doi:10.1016/j.physa.2016.11.001.

[14] V. Shergin, S. Udovenko and L. Chala “Assortativity Properties of Barabási-Albert Networks”, in Data-Centric Business and Application, Springer, 2020.

[15] V.L. Shergin, L.E. Chala and S.G. Udovenko, “Fractal dimension of infinitely growing discrete sets”, Advanced Trends in Radio Elecrtronics Telecommunications and Computer Engineering (TCSET), no. 348, 2018.

[16] V.L. Shergin and L.E. Chala, “The concept of elasticity of scale-free networks”, proc. Problems of Infocommunications. Science and Technology (PIC S&T), no. 62, 2017.

[17] R. Kumar and P. Raghavan, “Stochastic Models for the Web Graph (PDF)”, Foundations of Computer Science, 41st Annual Symposium on, pp. 57–65, 2000. doi:10.1109/SFCS.2000.892065

[18] D.S. Callaway, M.E.J. Newman, S.H. Strogatz and D.J. Watts, “Network Robustness and Fragility: Percolation on Random Graphs”, Physical Review Letters. 85 (25), 2000, pp. 5468–71. doi:10.1103/PhysRevLett.85.5468.

IEEE Third International Conference on Data Stream Mining & Processing
August 21-25, 2020, Lviv, Ukraine

Spark Structured Streaming: Customizing Kafka Stream Processing

Yuriy Drohobytskiy
Ternopil Volodymyr Hnatyuk National Pedagogical University
Ternopil, Ukraine
Dataengi UAB
Vilnius, Lituania
orcid.org/0000-0002-3333-1573

Vitaly Brevus
Dataengi UAB
Vilnius, Lituania
Ternopil Ivan Puluj National Technical University
Ternopil, Ukraine
orcid.org/0000-0002-7055-9905

Yuriy Skorenkyy
Ternopil Ivan Puluj National Technical University
Ternopil, Ukraine
orcid.org/0000-0002-4809-9025

Abstract — The aim of the present paper is to develop an improvement of large-scale multi-party data exchange and stream processing solution. The method of choice uses Apache Kafka streams as well as HDFS file granulation, and is exemplified in a real project of data ingestion into the Hadoop ecosystem. The management and conditional stream controlling procedures are proposed. Various ways to manage Kafka offsets during stream processing are considered.

Keywords—Distributed systems; Stream processing; Kafka streams; Spark v 2.3.2; HDFS file granulation

I. INTRODUCTION

Heterogeneous interconnected systems which produce a wide variety of data are nowadays common [1-3] for a range of applications, from energy generation and precision agriculture grids to an increasing number of small-scale and dispersed portable smart devices acquiring and transmitting diverse types of data. To support data-driven business models these data flows are to be properly processed in real time regardless of their variety and fragmentation [4]. To ingest data from the outer sources into HDFS in a cost-effective way is a complicated task to be solved by dedicated services that collect data from various sources outside of the Hadoop cluster, retrieve data from many databases, transform [5, 6] and enrich data to finally push it into the Kafka topic. On the other end, Spark application interacts with the Kafka topic and puts data into HDFS. 'External' service runs periodically, with periodicity chosen appropriately to the specific features of the task being solved. To resume such a short outline of the system's concept we note that the system's efficacy depends critically on details discussed in subsequent sections.

Current paper develops a solution to large-scale multi-party data exchange and stream processing problem. The considered Spark app reads data from Kafka, processes and storesdata in HDFS files. The standard streaming app runs permanently and processes data on the fly, which is an ineffective use of resources.

Typically, the data load which needs processing which lasts minutes to hours so resources of the Hadoop cluster will not be used effectively for Spark containers. The solution may be in starting an application in the defined moment to process all records from Kafka distributed streaming platform, and put it down for idle time interval (see Fig.1).

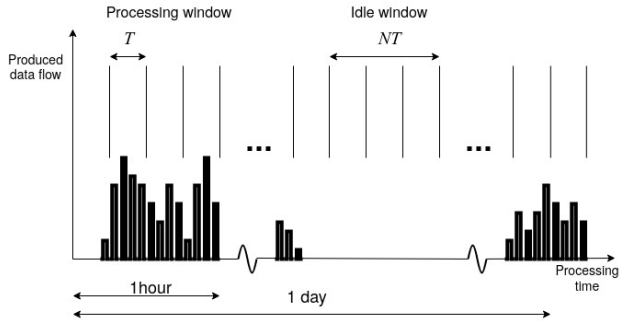


Fig. 1. Irregular stream server load.

A problem is to determine that the service has already completed the data processing. For this problem, a universal solution is absent as no alarm can be set for the last message arrival. A viable solution to this stream processing problem may be the enactment of a dedicated service application for stream processing.

II. STREAM PROCESSING CUSTOMIZATION

A. Solution idea

To optimally customize the stream processing for unpredictable flow of data from diverse producers one has to choose and wait a proper time interval and avoid wasting paid cloud computing services.

If during this interval there are no incoming messages in Kafka distributed streaming platform (see Fig. 2), then waiting phase is terminated, the consumed data are post-processed and application stopped untill the next predefined moment. This way, cluster resources are not wasted and all processing is timely performed in an appropriate way.

The practical realization of this simple idea is, however, not so straightforward as it may seem.

B. Implementation of the solution

To test the above idea we split the whole processing time into equal intervals T and monitor periodically if a new message arrives. To make sure that there is no incoming data in this application run we wait for N intervals and stop the application only if there were no messages in NT time. Let's examine a typical Apache Spark streaming application for big data processing that takes data from Kafka cluster and stores them into a Parquet binary file format in Hadoop Distributed File System (HDFS).

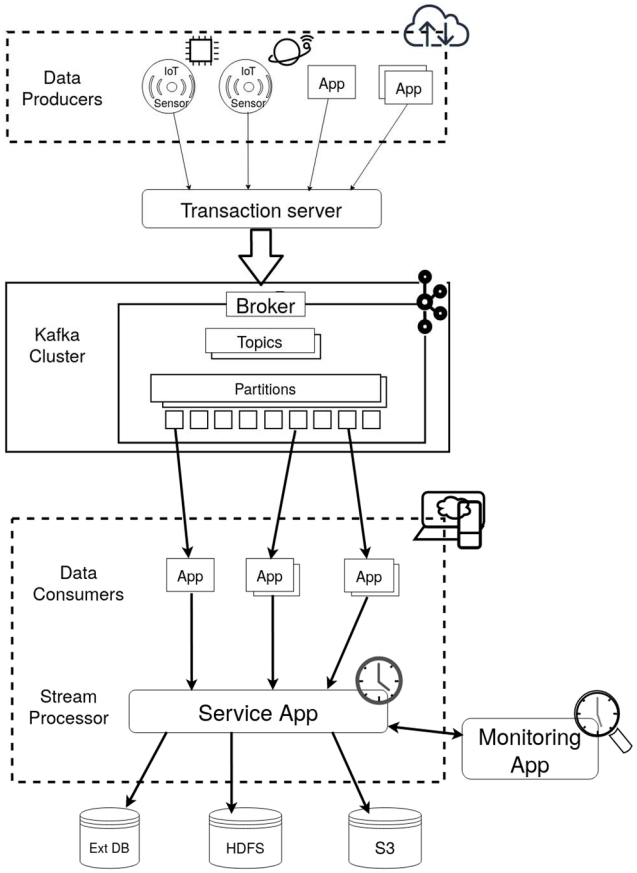


Fig.2 Architecture concept of the data processing platform.

More details should be taken into consideration [7-9] in real use cases, however these are irrelevant for the problem being solved. The presented code runs on Spark v 2.3.2, nevertheless it is fully compatible with the latest Spark 2.4.x version. The following code starts listening and consuming messages from Kafka topic:

```
val kafkaMessages: DataFrame = spark.readStream
 .format("kafka")
 .option("kafka.bootstrap.servers", bootstrapServers)
 .option("subscribe", topicIn)
 .option("startingoffsets", "latest")
 .load()
```

Here `bootstrapServers` specifies the address of Kafka brokers, `topicIn` is the name of the topic, `latest` ensures that only the new messages in the topic are listened to. The last option is unwise as it makes the application to start before the data-producing service outside the Hadoop. The solution to this problem, will be given later, too. For our purposes we require message information as string, and Kafka message information, namely partition and offset. Accordingly, we cast the message from Kafka to the following model:

```
case class KafkaMessage(
   partition: Int,
   offset: Long,
   value: String
)
```

with every field being self-explanatory.

```
val message = kafkaMessages.selectExpr("partition",
"offset", "CAST(value AS STRING)").as[KafkaMessage]
```

Now, message is the DataFrame consistsing of the `KafkaMessages`. To write messages into file we run the following stream:

```
val fileStream: StreamingQuery = message.writeStream
 .format("parquet")
 .outputMode("append")
 .trigger(Trigger.ProcessingTime(triggerInterval))
 .option("checkpointLocation", checkpointLocation)
 .option("path", outFilePath)
 .queryName(queryName)
 .start()
```

Here, `checkpointLocation` is the path for the Spark Streaming Checkpoint data to be stored in. This is necessary as Spark Streaming is fault-tolerant, and Spark needs to store its metadata into it. `queryName` is the arbitrary name of the streaming query, `outFilePath` is the path to the file on HDFS. `triggerInterval` is the period of time during which the Spark micro-batch is compiled and then processed by a Parquet writer, one at a time. So, at the moment, we have a stream that reads messages from Kafka and stores them into HDFS file. Also, we have temporal granularity of the stream.

According to the solution idea, we should be able to check number of messages ingested during each time interval, and if for N intervals we consumed no messages, the stream is stopped.

An interface to listen to the stream events (in `org.apache.spark.sql.streaming.StreamingQueryListener`) is:

```
abstract class StreamingQueryListener {

  import StreamingQueryListener._

  /**
   * Called when a query is started.
   * @note This is called synchronously with
   *
[[org.apache.spark.sql.streaming.DataStreamWriter
`DataStreamWriter.start()`]],
   *       that is, `onQueryStart` will be called on all
listeners before
   *       `DataStreamWriter.start()` returns the
corresponding [[StreamingQuery]]. Please
   *       don't block this method as it will block your
query.
   * @since 2.0.0
   */
  def onQueryStarted(event: QueryStartedEvent): Unit

  /**
   * Called when there is some status update (ingestion
rate updated, etc.)
   *
   * @note This method is asynchronous. The status in
[[StreamingQuery]] will always be
   *       latest no matter when this method is called.
Therefore, the status of [[StreamingQuery]]
   *       may be changed before/when you process the
event. E.g., you may find [[StreamingQuery]]
   *       is terminated when you are processing
`QueryProgressEvent`.
   * @since 2.0.0
   */
  def onQueryProgress(event: QueryProgressEvent): Unit

  /**
   * Called when a query is stopped, with or without
error.
   * @since 2.0.0
   */
  def onQueryTerminated(event: QueryTerminatedEvent):
Unit
}
```

To provide for the required functionality we create the listener as follows:

```
class StreamQueryListener(val query: StreamingQuery, val
maxEmptyTicks: Int = 3) extends StreamingQueryListener {

  private val queryId           = query.id
  private var currentEmptyCount = 0
  private var totalCount: Long  = 0
```

```
  override def onQueryStarted(event:
StreamingQueryListener.QueryStartedEvent): Unit = {
    if (event.id == queryId) {
      !s"Query started. (id = $queryId)"
    }
  }

  override def onQueryProgress(event:
StreamingQueryListener.QueryProgressEvent): Unit = {
    if (event.progress.id == queryId) {
      !s"Query progress. (id = $queryId)\n\tNumber of
input rows = ${event.progress.numInputRows},
currentEmptyCount = $currentEmptyCount (total count =
${totalCount + event.progress.numInputRows})"
      event.progress.numInputRows match {
        case 0 =>
          currentEmptyCount += 1
          checkCounterLimit()
        case x =>
          currentEmptyCount = 0
          totalCount += x
      }
    }
  }
  private def checkCounterLimit(): Unit = {
    if (currentEmptyCount >= maxEmptyTicks) {
      !s"Query will be STOPPED! (id = $queryId)"
      query.stop()
    }
  }
  override def onQueryTerminated(event:
StreamingQueryListener.QueryTerminatedEvent): Unit = {
    if (event.id == queryId) {
      !s"Query terminated. (id = $queryId)\n\tTotal rows
processed= $totalCount"
    }
  }
}
```

We add this listener as follows:

```
spark.streams.addListener(new
StreamQueryListener(fileStream, maxRetrives))
```

here `maxRetrives` is the number of retrieves with no messages to wait until the stream stops.

The main logic is in `onQueryProgress` method. We look at `event.progress.numInputRows` value which equals the number of rows obtained during the batch time window (set by `.trigger(Trigger.ProcessingTime(triggerInterval))`). If there are no messages in the stream we increment `currentEmptyCount` counter. When it reaches maximum allowed value then we can gracefully stop the stream by `query.stop()`. If we've got any messages during the time window then we clear the counter and start monitoring from the beginning. We also count the total number of processed messages here. (`!"string"` interpolator puts the string into logs.)

To complete the application workflow we wait for the stream to terminate:

```
fileStream.awaitTermination()
```

That's all for our task. We've got the streaming application which reads all data from Kafka topics and stops when the topic becomes 'empty'. So it does the job for our scheduled task.

III. Manual Kafka Offsets Management

Having in mind thet our goal is to optimally schedule the streaming application runs, it is unwise to re-read the topic from the beginning each time. Instead, one should start reading from the point it stopped last time. Using `.option("startingoffsets", "earliest")` for the `KafkaMessages` we will always read topic messages from the beginning. If the starting offsets is specified as `"latest"`, then reading sequence begins from the end and our goal will not mbe met as there could be new (unread) messages in Kafka *before* the application starts. To solve this problem, let's consider how Spark manages offsets for Kafka stream consumer (Fig. 3).

There are several options here:

1) Offset information could be stored by Kafka (usually Kafka uses `Zookeeper` for this). For this purpose each consumer should specify its own `group.id` and offsets are stored "per group". (This could be done automatically (autoCommit option), or manually). This option is completely useless with a structured streaming API, as there is no possibility to specify `group.id` option (see. documentation). Spark will assign a different `group.id` for the consumer each time the application starts.

2) The second option is what Spark proposes to do by default: offsets and the information about the output file writes are stored in the directory called `checkpoint`.

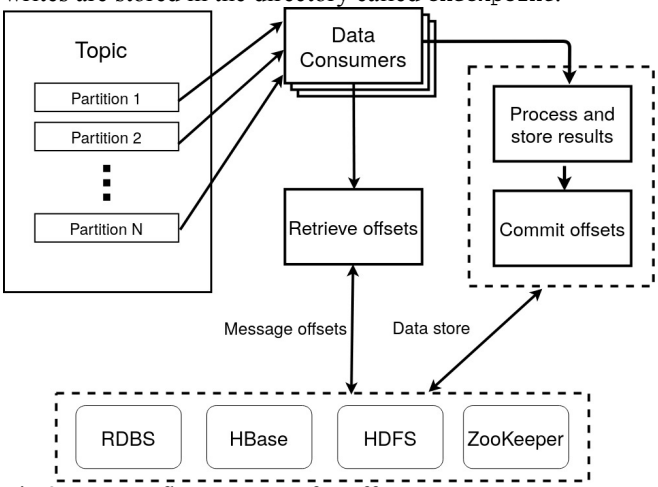


Fig.3. Process flow segment for offset management.

Checkpoints store intermediate information to ensure fault tolerance. If any sort of an exception occurs, i.e. JVM error, container fault or any other error takes place, then the application recovers from that point automatically. This powerful mechanism is to be used for critical applications. However, there are pitfalls in this approach. Firstly, the offsets are stored there too, so this folder can not be removed without the critical loss of the offset information. Secondly, the output files can not be removed, if this is done, the next application run will end with an error, as the information in the checkpoint will not match the output files. In our case, we'd like to remove the output files between subsequent application runs. The file will physically consist of the many parts (each of those parts is the data obtained during one processing time window), so after the application reads all data from the topic we want to aggregate all files into one big file and delete the intermediate files. To this end we are to remove the checkpoint directory as well. For this reason we consider other options to store Kafka offsets.

3) Finally, offsets can be manually stored and specified when creating a stream. This requires more effort but may represent the most flexible solution.

IV. IMPLEMENTATION

Multiple solutions may be attempted. We defined above the case class `KafkaMessage` for the received messages. It contains `partition` and `offset` information. Therefore, after we saved all messages in the files and stopped the stream, one can post-process the messages. The goal is to aggregate useful information into one large file and also store Kafka offsets for further usage. We split the data into Offsets information and useful data in the following way:

```
val offsets: DataFrame =
      fragmentedMsgs
        .select($"partition", $"offset")
        .groupBy($"partition")
        .agg(
          max($"offset").alias("offset")
        )
```

for the offsets information. And:

```
    val entities: Dataset[DataLakeEntity] =
fragmentedMsgs
      .select(from_json(col("value"),
DataLakeEntitySchemas.dataLakeEntitySchema).as("json_str"
))
      .select($"json_str.*")
      .as[DataLakeEntity]
```

for our `DataLakeEntity` information. (We use `Json` for the messages in Kafka topic, this could be different for the other application, i.e. `protobuf` or other formats could be used). As for offsets, the work is almost done:

```
   val offsetsList =
offsets.as[PartitionOffset].collect().toList

   if (offsetsList.nonEmpty) {
      // Store offsets somewhere, i.e.:
      offsetStore.insertOrUpdateOffsets(topicIn, offList)
   }
```

Now we need some storage for offsets. We can write them into Spark table, or in `HBase` DB or `PostgreSQL` DB etc. When starting the stream we should read offsets information back and pass it as option for the stream, like `.option("startingoffsets", "latest")`, but instead of the `"latest"` we should use special form, like:

```
{"topicA":{"0":23,"1":-1},"topicB":{"0":-1}}
```

Further information may be easily found in the documentation.

V. CONCLUSIONS

On the basis of the analysis of practically applicable solutions of large-scale multi-party data exchange and stream processing problems [10, 11] we suggest the mechanism for periodic monitoring and reading data from Kafka stream. The proposed method allows us to monitor data stream and manipulate it effectively under defined conditions. Multiple approaches to operate Kafka offsets in stream processing may be adopted depending on the peculiarities of the system and the processed data. While there are still options for offsets management which have not been discussed in detail in the present paper, this particular question can be addressed separately if a use case is specified.

For platforms driven by huge volumes of data coming live from multiple origins (IoT sensors, mobile devices etc), the ability to handle irregular data streams with unpredictable temporal distribution provides an important competitive edge, therefore the proposed architecture improvement of stream processing can lead to a dramatic cost reduction and performance increase for cloud services.

[1] J. Gama and P. P. Rodrigues, “An Overview on Mining Data Streams,” in *Foundations of Computational, IntelligenceVolume 6: Data Mining*, A. Abraham, A.-E. Hassanien, A. P. de Leon F. de Carvalho, and V. Snášel, Eds. Berlin, Heidelberg: Springer, 2009, pp. 29–45.

[2] J. Gama and P. P. Rodrigues, “Data Stream Processing,” in *Learning from Data Streams: Processing Techniques in Sensor Networks*, J. Gama and M. M. Gaber, Eds. Berlin, Heidelberg: Springer, 2007, pp. 25–39.

[3] G. Pal, G. Li, and K. Atkinson, “Big Data Real Time Ingestion and Machine Learning - IEEE Conference Publication,” presented at the 2016 IEEE First International Conference on Data Stream Mining Processing (DSMP), 2018, pp. 25–31.

[4] A. Batyuk and V. Voityshyn, “Apache storm based on topology for real-time processing of streaming data from social networks,” in *2016 IEEE First International Conference on Data Stream Mining Processing (DSMP)*, 2016, pp. 345–349

[5] P. J. Haas, “Data-Stream Sampling: Basic Techniques and Results,” in *Data Stream Management: Processing High-Speed Data Streams*, M. Garofalakis, J. Gehrke, and R. Rastogi, Eds. Berlin, Heidelberg: Springer, 2016, pp. 13–44.

[6] A. Batyuk, V. Voityshyn, and V. Verhun, “Software Architecture Design of the Real- Time Processes Monitoring Platform,” in *2018 IEEE Second International Conference on Data Stream Mining Processing (DSMP)*, 2018, pp. 98–101

[7] D. Vohra, “Apache Parquet,” in *Practical Hadoop Ecosystem: A Definitive Guide to Hadoop-Related Frameworks and Tools*, D. Vohra, Ed. Berkeley, CA: Apress, 2016, pp. 325–335.

[8] C. C. Aggarwal, “An Introduction to Data Streams,” in *Data Streams: Models and Algorithms*, C. C. Aggarwal, Ed. Boston, MA: Springer US, 2007, pp. 1–8.

[9] M. M. Gaber, A. Zaslavsky, and S. Krishnaswamy, “Data Stream Mining,” in *Data Mining and Knowledge Discovery Handbook*, O. Maimon and L. Rokach, Eds. Boston, MA: Springer US, 2010, pp. 759–787.

[10] T. Dunning and E. Friedman, *Streaming Architecture: New Designs Using Apache Kafka and MapR Streams*. O’Reilly Media, Inc., 2016.

[11] M. Armbrust *et al.*, “Structured Streaming: A Declarative API for Real-Time Applications in Apache Spark,” in *Proceedings of the 2018 International Conference on Management of Data*, Houston, TX, USA, 2018, pp. 601–613,.

IEEE Third International Conference on Data Stream Mining & Processing
August 21-25, 2020, Lviv, Ukraine

Modeling of Decision-Making Processes in Project Planning Based on Predictive Analytic Method

Nataliia Yehorchenkova
Project Management Department
Kyiv National University of Construction and Architecture
Kyiv, Ukraine
realnata@ukr.net

Oleksii Yehorchenkov
Department of Geoinformatics
Taras Shevchenko National University of Kyiv
Kyiv, Ukraine
alexee@ukr.net

***Abstract*— The objective of the proposed research is to develop a methodology for modeling and evaluation of decision-making processes in project planning, which potentially will reduce project manager wasting time for the development of a project schedule and a resources plan and its execution monitoring and control. The ultimate aim is to provide a project manager with a ready project plan which was developed based on project scope, assumptions and limitations by the digital project manager. The project manager can interact with the digital project manager, define assumptions and limitations, make changes in a ready project plan, etc. The digital project manager will define the project plan by choosing one from the project templates base of the company and change it according to requirements. The process of the template selecting has three stages: 1) defining project sphere, 2) choosing a group of templates to fit limitation, and 3) defining one template for the project plan. Exactly these stages are considered in this paper by using the modeling approach employs a stochastic production method.**

Keywords*— Mathematical model, Predictive analytic, Project management

I. INTRODUCTION

Mathematical modeling of project management processes is an important task of a lot of researchers in the world and as a result it is used in different practical approaches and applications. Despite this, modeling of project management processes is still a challenging problem because project management is a complex system consisted of a lot of persons, public and private organizations, technical applications that must interact with each other in the framework of time, budget and resources limits [1]. 49 processes of 10 areas of knowledge are defined in PMBoK and each of them has a major influence on project result, nonetheless the main is the decision-making process as such that plays a key role in project management. The importance of modeling this process presented in papers of Matthew J. Liberatore and Bruce Pollack-Johnson, Debu Mukerji [2-3], T. Gidel, R. Gautier & R. Duchamp [4] and others. The Decision-making process carried out by a project manager constantly and the outcome of the project fundamentally depends on the quality of these decisions. Nowadays there is a trend of project information growth that has to be collected, processed and analyzed by the project manager. More information leads to the more stress the project manager experiences and as a result the probability of making mistakes grows. So, for supporting project manager in decision-making process there is a necessity in using of special information technologies (ITPM), that will allow to receive in a short period necessary information in a convenient form and help to make the right decision. Examples of such information technologies are MS Project, Oracle Primavera, Clarizen, etc [5-7]. The use of such technologies is popular in project management, so development and implementation of ITPM is a quite fast-growing industry [8-9]. For today ITPM has gone from desktop applications to complex cloud platforms, and now there are researches in the field of development AI technologies for project management.

There was analyzed some of scientific works dedicated to development and using AI in project management. These researches were provided from ‘80 – ‘90 years of 20 Century [10-12]. Also, the research of the item continues today. Some of the modern researches are provided by Majid Shakhsi-Niaei (2015), María Martínez-Rojasa Nicolás (2016), Chaohong Gao (2015), Houda Hammouch (2015), M.Lachhab (2017), Benjamin Schreck and other [13-18].

The researches include describing of: 1) an intelligent system for retrieving and structuring data that allow to easily acquire and manage integrated formation for supporting decision making in the construction project management; 2) discusses several applications of intelligent systems in project management practice; 3) intelligent management platform design for engineering construction enterprise project management; 4) use the advantages of agent technology to make implementation and use PMBoK processes more efficient and have insights on the progress of projects and anticipate possible problems that can lead the project failure; 5) defining a common information model enabling the federation of all the points of view of the different actors with regards to Systems Engineering, Project Time Management, Project Cost Management, and Project Risk Management and 6) creating a machine learning model that uses the time-varying data to identify overarching patterns and predict critical problems ahead of time, these predictions would enable a project manager to selectively focus on a subset of projects, anticipate the occurrence of critical problems, determine the nature of the problems, pinpoint the areas that would be impacted, and take remedial action.

In the paper it is proposed to exploit the recent progress in the field of predictive analytic for the modeling decision-making process in project management for selecting a schedule template for developing a project plan. The developed model will be used in the algorithms of the AI project management system – a digital project manager [19].

II. HYPOTHESIS, RESEARCH QUESTION AND OBJECTIVES

Hypothesis:

One of the elements of project management digital transformation is the creation a such an intelligent computer

system (digital project manager) that provides a unified environment that included a combination of information systems for organization and will be able to manage processes of project management by itself without project management team participation.

The main research question for the paper:

How to teach a digital project manager to choose the right project template?

The creation of an intelligent computer system (digital project manager) is a pretty complex task. So, for looking for the answer it is important to make decomposition this question on simpler questions.

Authors propose to consider the next questions:

– How to separate projects by different classes due to the project area?

– How to offer a management model for every project class that is a template of a project plan formation?

– How to develop a method of project template selection?

Objective:

To define the algorithm of selecting a project schedule template by modeling the decision-making process for project planning using methods of predictive analytic.

III. The Primary Research Material

Digital project manager (further – d-MP) - software that implements typical project management processes in digital project management.

Digital project management (further – d-PM) – is a form of project management in which typical processes of management are realized in the digital environment.

The essence of d-PM is to select typical processes (processes that have identical signs for different projects, like task and resource planning, control of task executing) and transfer it in the digital environment of computers. At the same time creative processes (processes that need intelligence and creative approaches for project realization like team management, risk management, etc.) remain for the project team.

Transferring of typical processes to intelligence computer technology will relieve the project manager and project participant from routine work and give them more time to perform unique tasks that require a non-trivial approach [19].

Digital project manager [20] should be characterized by the next parameters:

- **Intelligence management** – the digital project manager should create and control the project plan and also realize resource management with minimal participation of the project team [21].
- **Virtual management** – the digital project manager should provide access for project participants to project from any part of the world.
- **Management "in cloud"** – all project information should be stored "in cloud" for providing access for project participants at any time from any kind of gadgets (laptop, smartphone, etc.).

Enlarged algorithm of process of planning in d-MP is shown on figure 1.

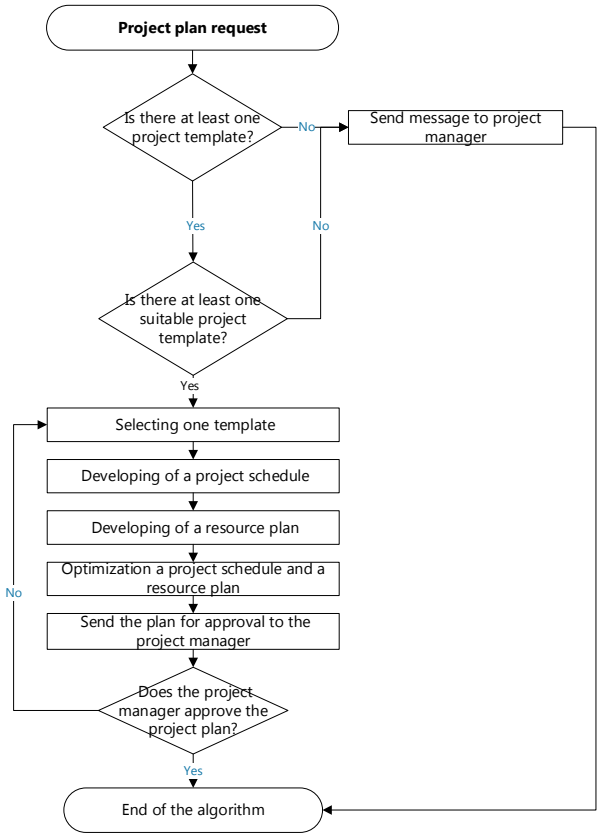


Fig. 1. Enlarged algorithms of processes of planning in d-MP

For defining the optimal project schedule template by d-MP the next tasks should be solved:

1. **Separating project templates by a project area (IT, construction, production, etc).**

A project area defines two factors of a schedule: a life cycle model and scope of a project. A project life cycle model affects on sequences of project tasks, for example, a construction project will have a waterfall model but IT project – V-model or RUP. Project scope defines the technology of project implementation. A construction project will have such tasks as architecture design or bricklaying, and IT project in turn will consist of stages like testing or programming. Under any other criteria like cost, duration, project start or finish date, or something else, exactly the project area defines the coordinate system in which the required project template will be located. That is why for creating a project plan it is important to make decomposition of project templates of a company by the different areas of project implementation.

For the development of a model of separating project templates by a project area could be used the methods of machine learning (Random forest, NLP, Neural networks and others):

2. **Highlighting templates that meet the requirements of the project manager**

Project results determine requirements for a project plan, so every template has to be determined following the values of the features by a project manager. Such features can include both general information like project duration, cost or evaluation of project success, etc., and specific information

according to the project area, for example number of floors for construction projects or GUI model for IT project. The development of a model of a template selection from the archive of a project-oriented enterprise by d-PM has to base on machine learning or predictive analytics methods.

3. Selecting an optimal project template.

In a field of highlighted templates can get some suitable examples and based on this situation, there may be several scenarios for d-PM:

1. Show to the project manager all suitable templates and leave the decision to him.

2. Select only one optimal template for the project manager. It is should be decided which method to use for this scenario: genetic algorithm, random selection, neural networks, etc.

Solving these tasks is an important part of the development of the digital project manager because namely answers to them will be a core of the decision-making process of this AI system.

For visualizing the main idea the authors propose the next example. There are one hundred project schedules (templates) in the archive of a project-oriented company. Every template is characterized by three criteria: project duration (fig.2), project cost (fig.3) and area (IT, Health, Production, Construction).

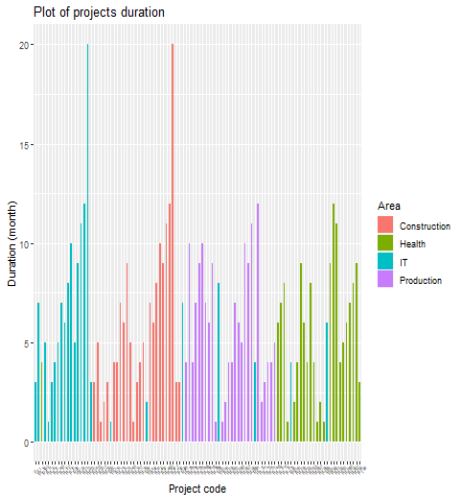


Fig. 2. Plot of projects duration

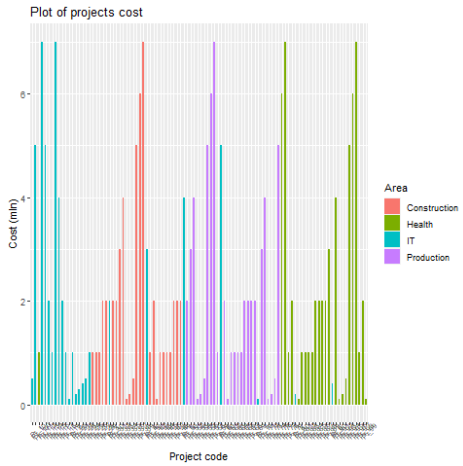


Fig. 3. Plot of projects cost

The decomposition of project templates by the project area is shown in figure 4. In the project area IT there was found through suitable templates (blue points) and one that is optimal (the red point) for the project manager.

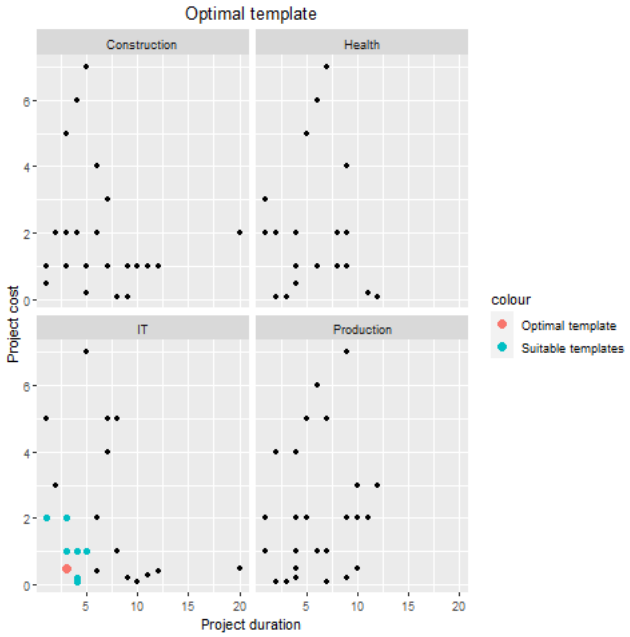


Fig. 4. Plot of projects duration

For the development of an algorithm of selecting a project schedule template each of the above tasks should be solved.

For this it is necessary to use scientific approaches and check the effectiveness of each possible model for finding solutions for these tasks.

In this paper authors propose to consider predictive analytic methods for development model for highlighting suitable templates.

IV. Stochastic-Production Method To Determine The Project Template

It is proposed to use a stochastic-production method to determine the project template.

For this method, there is necessary to pass from the system of deterministic products to the system of stochastic products. To do this, divide each product that contains a set of formal parameters into products, each of which corresponds to a separate parameter, or a combination of parameters related to each other. The kernels of production will be represented by a stochastic model:

if A, then with probability p_1 *–* B_1*, with probability* p_2 *–* B_2*, …, with probability* p_n *–* B_n*.*

Then formally, such a system of production can be defined as follows:

1. *A product name* – its unique name in the system.

2. *Area* - project area.

3. *Condition of use products* – the existence of at least one template in the enterprise archive for the realization of the kernel of production.

4. *The kernel of production* - parameters a_1 has a value within a_{min} - a_{max}, then choose a template with probability p_1 - B_1, with probability p_2 - B_2,…, with probability p_n - B_n.

5. *The post-condition* - to estimate the common conditional probability of a template choice for all selected products (for all products of the given area of application for which the condition is fulfilled).

A system of forecasting products is specified to identify the project template. Products form the knowledge base for choosing a template and determine the probability that a particular management model will be selected, with a given require parameter Π^x.

Products are formed from project and company statistics, or experts. The problem is that the combinations of parameters for each request/influence may be different. For this purpose the following method of definition of a template is offered:

1 For a given area of projects to determine the scope and activate (execute) the products for which the condition of application is fulfilled - the availability of appropriate templates in the enterprise archive.

2. If the set of activated products is empty, it is complete.

3. If the set consists of one product - determine the template most likely:

$$\mathrm{M}^x = \mathrm{B}_{\mathrm{ki}} \mid \mathrm{p}_{\mathrm{ki}} \geq \mathrm{p}_{\mathrm{kj}},\ j = \overline{1,n}, \qquad (1)$$

where B_{ki} – template in the activated product with the name (k);

p_{ki} – the probability of selecting a template B_{ki} in name-enhanced products (k);

p_{kj} – the probability of selecting a template B_{kj} in name-enhanced products (k);

M^x – selected template to request Π^x;

n – the number of templates.

Completion.

4. If multiple products have one that has a choice of the template with probability 1, select that template:

$$\mathrm{M}^x = \mathrm{B}_{\mathrm{k1}} \mid \mathrm{p}_{\mathrm{k1}} = 1. \qquad (2)$$

Completion.

5. If there are several products in a set of products that have a choice of different control models with probability 1, then this means that the product system is contradictory. Completion.

6. Estimate the likelihood of selecting a template based on the probabilities specified in the applicable product system. For this purpose it is necessary to solve the following problem:

$$\forall \mathrm{M}_i \in \mathrm{M}, \exists \mathrm{p}_{\mathrm{ki}}, \mathrm{k} = \overline{1,L}, \qquad (3)$$

де L – number of products in the system;

p_{ki} – the probability of selecting a template M_i in name-enhanced products (k);

M_i – template;

M – set of templates.

It is necessary to find an estimate of the multiple probabilities of selecting each of the templates and select the highest-rated template:

$$\forall M_j \in M, \exists \Omega(p_{1j}, p_{2j}, \dots, p_{kj}, \dots, p_{Lj}) \Longrightarrow \exists M_i \in M, \Omega(p_{1i}, p_{2i}, \dots, p_{ki}, \dots, p_{Li},) \geq \Omega(p_{1j}, p_{2j}, \dots, p_{kj}, \dots, p_{Lj}) \qquad (4)$$

де $\Omega(\mathrm{p}_{1\mathrm{j}}, \mathrm{p}_{2\mathrm{j}}, \dots, \mathrm{p}_{\mathrm{kj}}, \dots, \mathrm{p}_{\mathrm{Lj}})$ – estimate the probability of selecting a template $\mathrm{M_j}$;

$\Omega(\mathrm{p}_{1\mathrm{i}}, \mathrm{p}_{2\mathrm{i}}, \dots, \mathrm{p}_{\mathrm{ki}}, \dots, \mathrm{p}_{\mathrm{Li}})$ – estimate the probability of selecting a template $\mathrm{M_i}$;

p_{ki} – the probability of selecting a template $\mathrm{M_i}$ in products (k);

p_{kj} – the probability of selecting a template $\mathrm{M_j}$ in products (k).

For this estimate, we assume the product of the probabilities of selecting a template:

$$\Omega(p_{1j}, p_{2j,} \dots, p_{kj,} \dots, p_{Lj}) = \prod_{k=1}^{L} p_{kj}\,; \qquad (5)$$

$$\Omega(p_{1i}, p_{2i,} \dots, p_{ki,} \dots, p_{Li}) = \prod_{k=1}^{L} p_{ki}. \qquad (6)$$

7. Selecting a template with the highest probability estimate

$$\mathrm{M}^x = \mathrm{M}_{\mathrm{i}} \mid \Omega(\mathrm{p}_{1\mathrm{i}}, \mathrm{p}_{2\mathrm{i}}, \dots, \mathrm{p}_{\mathrm{ki}}, \dots, \mathrm{p}_{\mathrm{Li}}) \geq \Omega(\mathrm{p}_{1\mathrm{j}}, \mathrm{p}_{2\mathrm{j}}, \dots, \mathrm{p}_{\mathrm{kj}}, \dots, \mathrm{p}_{\mathrm{Lj}}),\ j = \overline{1,n}. \qquad (7)$$

8. Completion.

This method can only be applied if the project schedule is characterized by a clear and stable set of parameters that is repeated from project to project. If the set of parameters is unclear or not repeated, then the number of products should be very large and a large number of experts should be employed for a long time to describe them. Which reduces the effectiveness of digital project management. Moreover, it is almost impossible to describe all possible variants of parameters and their corresponding models. So there is could be concluded that using a predictive analytic method namely method of production is not effective for the modeling decision-making process for selecting a project template.

V. CONCLUSION

Modeling of the decision-making process for the development of project management information systems is an important part of scientific activities. Modern research background allows us to conduct investigations in this direction to find the most effective solutions.

Using a predictive analytic method for the modeling decision-making process of project scheduling is considered in the paper. This research is an important stage for the development of AI system d-MP that increases chances for the successful completion of the project by helping the project manager in carrying out his typical tasks making its execution faster and more accurate. One such typical task is selecting a project template that will be used for creating an optimal project schedule is met all requirements and limitations. Exactly it was for this kind of task that the application was considered stochastic-production method. As a result, it was decided that this method not suitable for the modeling process of selecting the project schedule as not universal and it requires the development of complex algorithms consisting of a lot of amount of formalized production.

The perspective for further research is first of all review and test experiments of several methods like neuro network or clustering and selecting the optimal for intelligence selecting project template. And the second is the development of algorithm and implementation in a software application.

REFERENCES

[1] A Guide to the Project Management Body of Knowledge (PMBOK® Guide)–Sixth Edition. Project Management Institute, 2016, 792p.

[2] M. Liberatore, B.Pollack-Johnson "Improving Project Management Decision Making by Modeling Quality, Time, and Cost Continuously", Engineering Management IEEE, vol.60, 2013, pp.518-528.

[3] D.Mukerji "A study to improve decision making process in planning for reduction of project failures: a thesis research note", European Journal of Management, vol. 13, 2013, pp. 85-102.

[4] T. Gidel, R. Gautier, R. Duchamp "Decision-making framework methodology: an original approach to project risk management in new product design", Journal of Engineering Design, vol.16 iss. 1, 2005, pp.1-23

[5] "Oracle" [Online]. Available: https://www.clarizen.com/

[6] "Primavera P6 Enterprise Project Portfolio Management" [Online]. Available:https://www.oracle.com/ru/applications/primavera/products/project-portfolio-management/

[7] "Microsoft Project" [Online]. Available: https://ru.wikipedia.org/wiki/Microsoft_Project

[8] "Top Trends in the Gartner Hype Cycle for Emerging Technologies, 2017" [Online]. Available: https://www.gartner.com/smarterwithgartner/top-trends-in-the-gartner-hype-cycle-for-emerging-technologies-2017/

[9] "Gartner Reveals the 2017 Hype Cycle for Data Management" [Online]. Available: https://www.gartner.com/newsroom/id/3809163

[10] An Intelligent Project Management System [Online]. Available: https://pdfs.semanticscholar.org/7a9/6/de987efb21716f46be9c7b6478ba5020a0ff.pdf

[11] F J Jüngen; W Kowalczyk "An intelligent interactive project management system" [Online]. Available: https://www.worldcat.org/title/intelligent-interactive-project-management-system/oclc/67523821&referer=brief_results

[12] S. Smith "Towards an intelligent planning system. International Journal of Project Management", vol. 10, iss. 4, 1992, pp. 213-218

[13] M.Shakhsi-Niaei, S.HIranmanesh "Intelligent Systems in Project Planning". In: Kahraman C., Çevik Onar S. (eds) Intelligent Techniques in Engineering Management. Intelligent Systems Reference Library, vol 87, 2015

[14] M. Martínez-Rojasa, N.Marín, M. Amparo "Vila Miranda An intelligent system for the acquisition and management of information from bill of quantities in building projects", Expert Systems with Applications, vol.63, 2016, pp. 284-294

[15] G.Chaohong "Intelligent Management Platform Design for Engineering Construction Project" [Online]. Available: https://ieeexplore.ieee.org/document/7473344?reload=true

[16] H.Hammouch, H.Medromi, A.Sayouti "Toward an intelligent system for project management based on the multi agents systems" [Online]. Available: https://ieeexplore.ieee.org/document/7358412

[17] M.Lachhab, C.Béler, E.L.Solano-Charris, T.Coudert "Towards Integration of Systems Engineering and Project Management Processes for a Decision Aiding Purpose" [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2405896317319213

[18] B.Schreck, S.Mallapur, S.Damle and et.al. "The AI Project Manager" [Online]. Available: https://www.featuretools.com/wp-content/uploads/2018/03/AIPM.pdf

[19] N.Yehorchenkova, O.Yehorchenkov "Conceptual Groundwork of Digital Transformation of Project Management", XIV International Scientific and Technical Conference «Computer Sciences and Information Technologies» , September 17 – 20, 2019

[20] "An Overview of Digital Project Management" [Online]. Available: https://www.villanovau.com/resources/project-management/an-overview-of-digital-project-management/#.Wj1YCd9l_Dd

[21] "Agent-based project team collaboration behavior simulation" [Online]. Available: http://ieeexplore.ieee.org/abstract/document/7539939

IEEE Third International Conference on Data Stream Mining & Processing
August 21-25, 2020, Lviv, Ukraine

Using Bayesian Regression for Stacking Time Series Predictive Models

Bohdan M. Pavlyshenko
Ivan Franko National University of Lviv
Lviv, Ukraine
email: b.pavlyshenko@gmail.com

***Abstract*—The paper describes the use of Bayesian inference for stacking regression of different predictive models for time series. The models ARIMA, Neural Network, Random Forest, Extra Tree were used for the prediction on the first level of model ensemble. On the second level, time series predictions of these models on the validation set were used for stacking by Bayesian regression. This approach gives distributions for regression coefficients of these models. It makes it possible to estimate the uncertainty contributed by each model to stacking result. The information about these distributions allows us to select an optimal set of stacking models, taking into account the domain knowledge. A probabilistic approach for stacking predictive models allows us to make risk assessment for the predictions that is important in a decision-making process.**

***Index Terms*—Keywords: time series, Bayesian regression, machine learning, stacking, forecasting, sales**

I. Introduction

Time series analytics is an important part of modern data science. The examples of different time series approaches are in [1]–[7]. In [8]–[13], a range of ensemble-based methods for classification problems is regarded. In [14], the authors studied a lagged variable selection, hyperparameter optimization, as well as compared classical and machine learning based algorithms for time series. Sales prediction is an important part of modern business intelligence [15]–[17]. Sales can be regarded as a time series. However, time series approaches have some limitations for sales forecasting. For instance, to study seasonality, historical data for large time period are required. But it frequently happens that there are no historical data for a target variable, e.g. when a new product is launched. Furthermore, we must take into consideration many exogenous factors influencing sales. We can consider time series forecasting as regression problem in many cases, especially for sales time series. In [18], we considered linear models, machine learning and probabilistic models for time series modeling. For probabilistic modeling, we studied the use of copulas and Bayesian inference approaches. In [19], we regarded the logistic regression with Bayesian inference for analysing manufacturing failures. In [20], we consider the use of machine learning models for sales predictive analytics. We researched the main approaches and case studies of implementing machine learning to sales forecasting. The effect of machine learning generalization has been studied. This effect can be used for predicting sales in case of a small number of historical data for specific sales time series in case when a new product or store is launched [20]. A stacking approach for building ensemble of single models using Lasso regression has also been studied. The obtained results show that using stacking techniques, we can improve the efficiency of predictive models for sales time series forecasting [20].

In this paper, we consider the use of Bayesian regression for stacking time series predictive models on the second level of the predictive model which is the ensemble of the models of the first level.

II. Bayesian Stacking Regression

Probabilistic regression models can be based on Bayesian theorem [21]–[23]. This approach allows us to receive a posterior distribution of model parameters using conditional likelihood and prior distribution. Probabilistic approach is more natural for stochastic variables such as sales time series. The difference between Bayesian approach and conventional Ordinary Least Squares (OLS) method is that in the Bayesian approach, uncertainty comes from parameters of model, as opposed to OLS method where the parameters are constant and uncertainty comes from data. In the Bayesian inference, we can use informative prior distributions which can be set up by an expert. So, the result can be considered as a compromise between historical data and expert opinion. It is important in the cases when we have a small number of historical data. In the Bayesian model, we can consider the target variable with non Gaussian distribution, e.g. Student's t-distribution. Probabilistic approach gives us an ability to receive probability density function for the target variable. Having such function, we can make risk assessment and calculate value at risk (VaR) which is 5% quantile. For solving Bayesian models, the numerical Monte-Carlo methods are used. Gibbs and Hamiltonian sampling are the popular methods of finding posterior distributions for the parameters of probabilistic model [21]–[23]. Predictive models can be combined into ensemble model using stacking approach [8]–[13]. In this approach, prediction results of predictive models on the validation set are treated as covariates for stacking regression. These predictive models are considered as a first level of predictive model ensemble. Stacking model forms the second level of model ensemble. Using Bayesian inference for stacking regression gives distributions for stacking regression coefficients. It enables to estimate the uncertainty of the first level predictive models. As predictive models for first

978-1-7281-3214-3/20/$31.00 ©2020 IEEE

level of ensemble we used the following models: ARIMA', 'ExtraTree', 'RandomForest', 'Lasso', 'NeuralNetowrk'. The use of these models for stacking by Lasso regression was described in [20].

For stacking, we have chosen the robust regression with Student's t-distribution for the target variable as

$$y \sim Student_t(\nu, \mu, \sigma) \tag{1}$$

where

$$\mu = \alpha + \sum_i \beta_i x_i, \tag{2}$$

ν is a distribution parameter, called as degrees of freedom, i is an index of the predictive model in the stacking regression, $i \in \{$ 'ARIMA', 'ExtraTree', 'RandomForest', 'Lasso', 'NeuralNetowrk' $\}$.

III. Numerical Modeling

The data for our analysis are based on store sales historical data from the "Rossmann Store Sales" Kaggle competition [24]. For Bayesian regression, we used Stan platform for statistical modeling [23]. The analysis was conducted in Jupyter Notebook environment using Python programming language and the following main Python packages *pandas, sklearn, pystan, numpy, scipy,statsmodels, keras, matplotlib, seaborn*. We trained different predictive models and made the predictions on the validation set. The model ARIMA was evaluated using *statsmodels* package, Neural Network was evaluated using *keras* package, Random Forest and Extra Tree was evaluated using *sklearn* package. In these calculations, we used the approaches described in [20]. Figure 1 shows the time series forecasts on the validation sets obtained using different models.

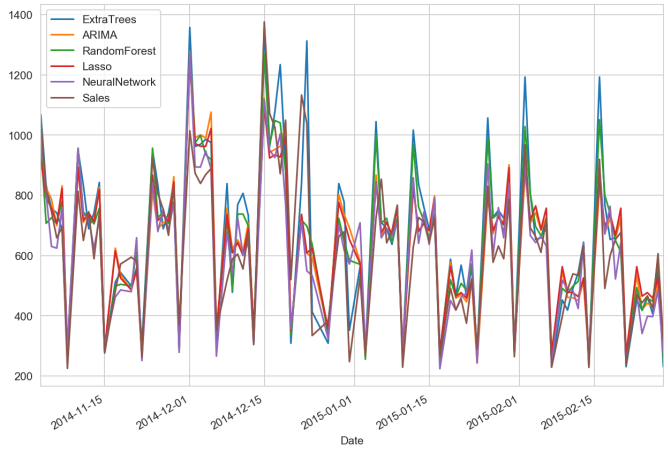


Fig. 1. Forecasting of different models on the validation set

The results of prediction of these models on the validation sets are considered as the covariates for the regression on the second stacking level of the models ensemble. For stacking predictive models we split the validation set on the training and testing sets. For stacking regression we normalized the covariates and target variable using z-scores:

$$z_i = \frac{x_i - \mu_i}{\sigma_i}, \tag{3}$$

where μ_i is the mean value, σ_i is the standard deviation. The prior distributions for parameters α, β, σ in the Bayesian regression model (1)-(2) are considered as Gaussian with mean values equal to 0, and standard deviation equal to 1. We split the validation set on the training and testing sets by time factor. The parameters for prior distributions can be adjusted using prediction scores on testing sets or using expert opinions in the case of small data amount. To estimate uncertainty of regression coefficients, we used the coefficient of variation which is defined as a ratio between the standard deviation and mean value for model coefficient distributions:

$$v_i = \frac{\sigma_i}{\mu_i}, \tag{4}$$

where v_i is coefficient of variation, σ_i is a standard deviation, μ_i is the mean value for the distribution of the regression coefficient of the model i. Taking into account that μ_i can be negative, we will analyze the absolute value of the coefficient of variation $|v_i|$. For the results evaluations, we used a relative mean absolute error (RMAE) and root mean square error (RMSE). Relative mean absolute error (RMAE) was considered as a ratio between the mean absolute error (MAE) and mean values of target variable:

$$RMAE = \frac{E(|y_{pred} - y|)}{E(y)} 100\% \tag{5}$$

Root mean square error (RMSE) was considered as:

$$RMSE = \sqrt{\frac{\sum_i^n (y_{pred} - y)^2}{n}} \tag{6}$$

The data with predictions of different models on the validation set were split on the training set (48 samples) and testing set (50 samples) by date. We used the robust regression with Student's t-distribution for the target variable. As a result of calculations, we received the following scores: RMAE(train)=12.4%, RMAE(test)=9.8%, RMSE(train)=113.7, RMSE(test)= 74.7. Figure 2 shows mean values time series for real and forecasted sales on the validation and testing sets. Vertical dotted line separates the training and testing sets. Figure 3 shows the probability density function (PDF) for the intersect parameter. One can observe positive bias of this (PDF). It is caused by the fact that we applied machine learning algorithms to non stationary time series. If a non stationary trend is small, it can be compensated on the validation set using stacking regression.

Figure 4 shows the box plots for the PDF of coefficients of models. Figure 5 shows the coefficient of variation for the PDF of regression coefficients of models.

We considered the case with the restraints to regression coefficient of models that they should be positive. We received the similar results: RMAE(train)=12.9%, RMAE(test)=9.7%, RMSE(train)=117.3, RMSE(test)=76.1. Figure 6 shows the

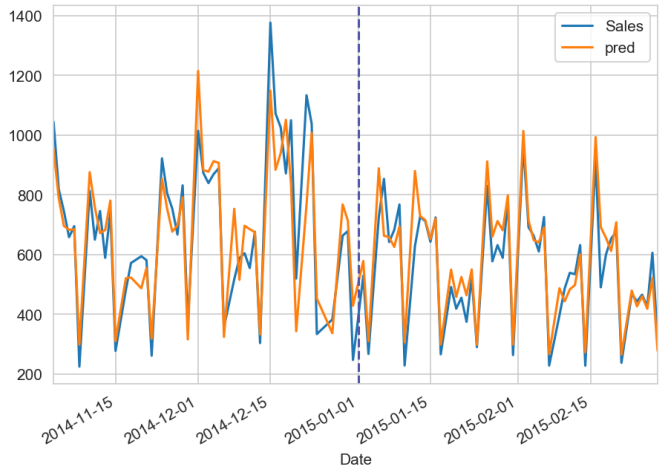


Fig. 2. Mean values time series for real and forecasted sales on validation and testing sets

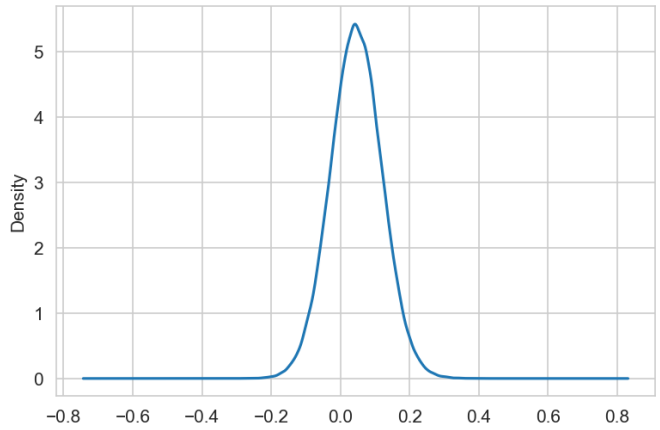


Fig. 3. The PDF for intersect parameter of stacking regression

box plots for the PDF of model regression coefficients for this case.

All models have a similar mean value and variation coefficients. We can observe that errors characteristics RMAE and RMSE on the validation set can be similar with respect to these errors on the training set. It tells us about the fact that Bayesian regression does not overfit on the training set comparing to the machine learning algorithms which can demonstrate essential overfitting on training sets, especially in the cases of small amount of training data. We have chosen the best stacking model ExtraTree and conducted Bayesian regression with this one model only. We received the following scores: RMAE(train)=12.9%, RMAE(test)=11.1%, RMSE(train)=117.1, RMSE(test)=84.7. We also tried to exclude the best model ExtraTree from the stacking regression and conducted Bayesian regression with the rest of models without ExtraTree. In this case we received the following scores: RMAE(train)=14.1%, RMAE(test)=10.2%, RMSE(train)=139.1, RMSE(test)=75.3. Figure 7 shows the box plots for the PDF of model regression coefficients, figure

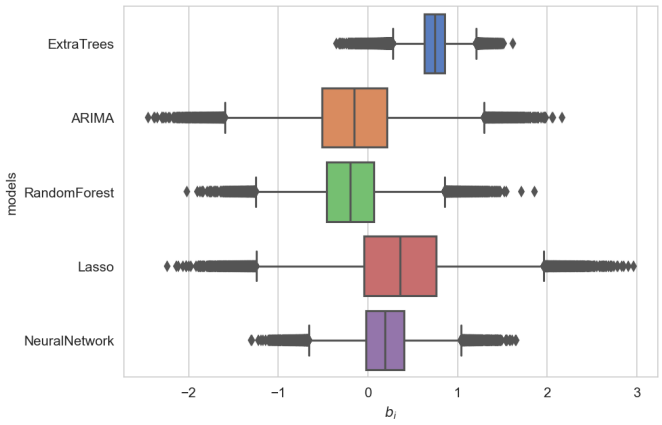


Fig. 4. Box plots for the PDF of regression coefficients of models

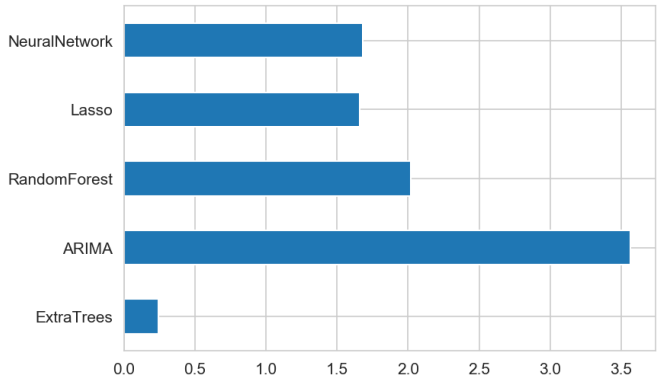


Fig. 5. Absolute values of the coefficient of variation for the PDF of regression coefficients of models

8 shows the coefficient of variation for the PDF of regression coefficients of models for this case study. We received worse results on the testing set. At the same time these models have the similar influence and thus they can potentially provide more stable results in the future due to possible changing of

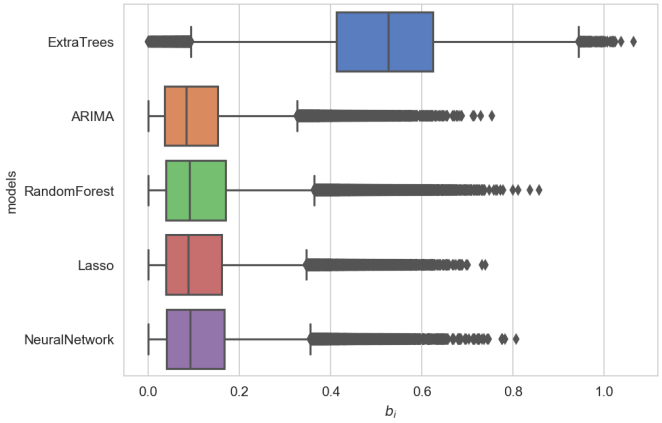


Fig. 6. Box plots for the PDF of regression coefficients of models

the quality of features. Noisy models can decrease accuracy on large training data sets, at the same time they contribute to sufficient results in the case of small data sets. We considered

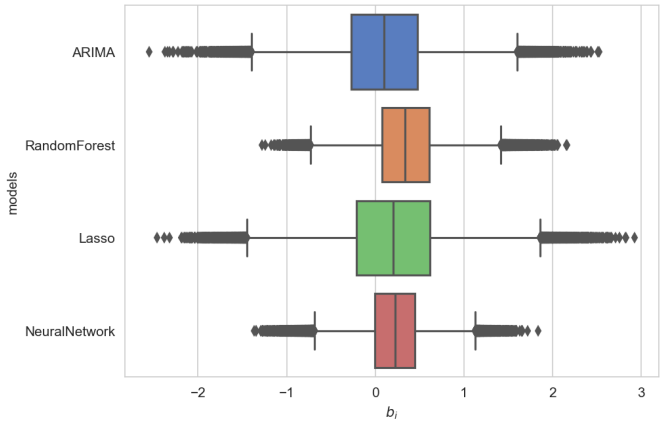


Fig. 7. Box plots for the PDF of regression coefficients of models

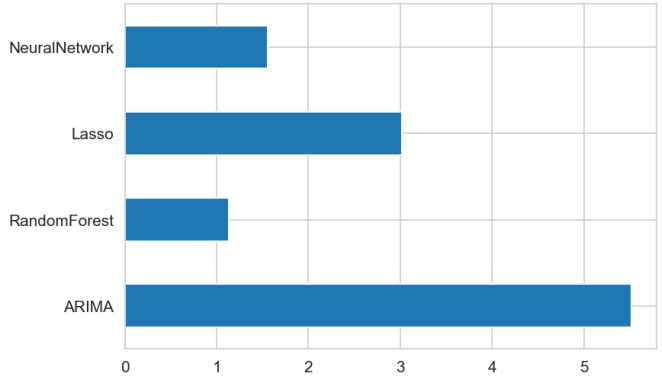


Fig. 8. Absolute values of the coefficient of variation for the PDF of models regression coefficients

the case with a small number of training data, 12 samples. To get stable results, we fixed the ν parameter of Student's t-distribution in Bayesian regression model (1)-(2) equal to 10. We received the following scores: RMAE(train)=5.0%, RMAE(test)=14.2%, RMSE(train)=37.5, RMSE(test)=121.3. Figure 9 shows mean values time series for real and forecasted sales on the validation and testing sets. Figure 10 shows the box plots for the PDF of models coefficients. Figure 11 shows the coefficient of variation for the PDF of models regression coefficients. In this case, we can see that an other model starts playing an important role comparing with the previous cases and ExtraTree model does not dominate. The obtained results show that optimizing informative prior distributions of stacking model parameters can improve the scores of prediction results on the testing set.

IV. Conclusion

In the work, we considered a two-level ensemble of the predictive models for time series. The models ARIMA, Neural Network, Random Forest, Extra Tree were used for the prediction on the first level of ensemble of models. On the second stacking level, time series predictions of these models on the

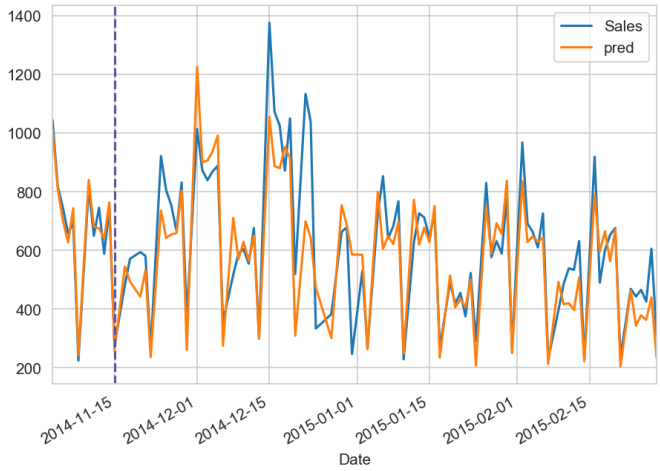


Fig. 9. Mean values time series for real and forecasted sales on the validation and testing sets in the case of small training set

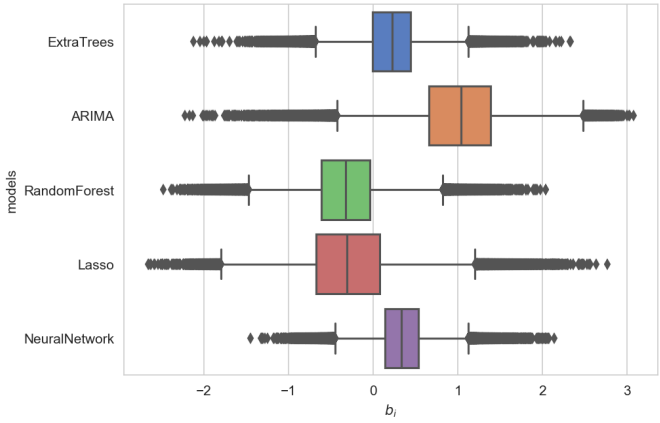


Fig. 10. Box plots for the PDF of regression coefficients of models in the case of small training set

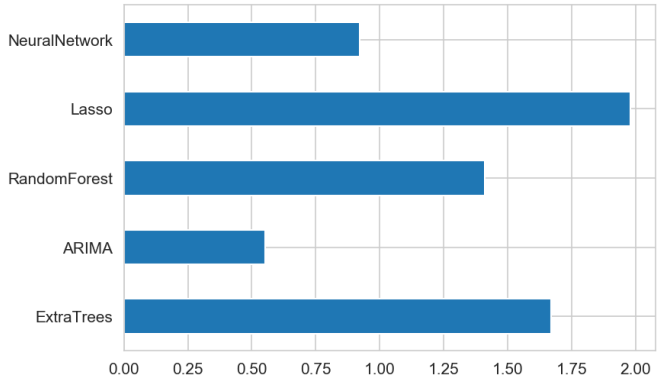


Fig. 11. Absolute values of the coefficient of variation for the PDF of regression coefficients of models in the case of small training set

validation set were conducted by Bayesian regression. Such an approach gives distributions for regression coefficients of these models. It makes it possible to estimate the uncertainty contributed by each model to the stacking result. The information about these distributions allows us to select optimal set of stacking models, taking into account domain knowledge. Probabilistic approach for stacking predictive models allows us to make risk assessment for the predictions that is important in a decision-making process. Noisy models can decrease accuracy on large training data sets, at the same time they contribute to sufficient results in the case of small data sets. Using Bayesian inference for stacking regression can be useful in cases of small datasets and help experts to select a set of models for stacking as well as make assessments of different kinds of risks. Choosing the final models for stacking is up to an expert who takes into account different factors such as uncertainty of each model on the stacking regression level, amount of training and testing data, the stability of models. In Bayesian regression, we can receive a quantitative measure for the uncertainty that can be a very useful information for experts in model selection and stacking. An expert can also set up informative prior distributions for stacking regression coefficients of models, taking into account the domain knowledge information. So, Bayesian approach for stacking regression can give us the information about uncertainty of predictive models. Using this information and domain knowledge, an expert can select models to get stable stacking ensemble of predictive models.

REFERENCES

[1] C. Chatfield, *Time-series forecasting*. Chapman and Hall/CRC, 2000.

[2] P. J. Brockwell, R. A. Davis, and M. V. Calder, *Introduction to time series and forecasting*, vol. 2. Springer, 2002.

[3] G. E. Box, G. M. Jenkins, G. C. Reinsel, and G. M. Ljung, *Time series analysis: forecasting and control*. John Wiley & Sons, 2015.

[4] R. J. Hyndman and G. Athanasopoulos, *Forecasting: principles and practice*. OTexts, 2018.

[5] R. S. Tsay, *Analysis of financial time series*, vol. 543. John Wiley & Sons, 2005.

[6] W. W. Wei, "Time series analysis," in *The Oxford Handbook of Quantitative Methods in Psychology: Vol. 2*, 2006.

[7] R. J. Hyndman, Y. Khandakar, *et al.*, *Automatic time series for forecasting: the forecast package for R*. No. 6/07, Monash University, Department of Econometrics and Business Statistics, 2007.

[8] D. H. Wolpert, "Stacked generalization," *Neural networks*, vol. 5, no. 2, pp. 241–259, 1992.

[9] L. Rokach, "Ensemble-based classifiers," *Artificial Intelligence Review*, vol. 33, no. 1-2, pp. 1–39, 2010.

[10] O. Sagi and L. Rokach, "Ensemble learning: A survey," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 8, no. 4, p. e1249, 2018.

[11] H. M. Gomes, J. P. Barddal, F. Enembreck, and A. Bifet, "A survey on ensemble learning for data stream classification," *ACM Computing Surveys (CSUR)*, vol. 50, no. 2, p. 23, 2017.

[12] T. G. Dietterich, "Ensemble methods in machine learning," in *International workshop on multiple classifier systems*, pp. 1–15, Springer, 2000.

[13] L. Rokach, "Ensemble methods for classifiers," in *Data mining and knowledge discovery handbook*, pp. 957–980, Springer, 2005.

[14] G. Papacharalampous, H. Tyralis, and D. Koutsoyiannis, "Univariate time series forecasting of temperature and precipitation with a focus on machine learning algorithms: A multiple-case study from greece," *Water Resources Management*, pp. 1–33, 2018.

[15] J. T. Mentzer and M. A. Moon, *Sales forecasting management: a demand management approach*. Sage, 2004.

[16] T. Efendigil, S. Önüt, and C. Kahraman, "A decision support system for demand forecasting with artificial neural networks and neuro-fuzzy models: A comparative analysis," *Expert Systems with Applications*, vol. 36, no. 3, pp. 6697–6707, 2009.

[17] G. P. Zhang, *Neural networks in business forecasting*. IGI Global, 2004.

[18] B. M. Pavlyshenko, "Linear, machine learning and probabilistic approaches for time series analysis," in *Data Stream Mining & Processing (DSMP), IEEE First International Conference on*, pp. 377–381, IEEE, 2016.

[19] B. Pavlyshenko, "Machine learning, linear and bayesian models for logistic regression in failure detection problems," in *Big Data (Big Data), 2016 IEEE International Conference on*, pp. 2046–2050, IEEE, 2016.

[20] B. M. Pavlyshenko, "Machine-learning models for sales time series forecasting," *Data*, vol. 4, no. 1, p. 15, 2019.

[21] J. Kruschke, *Doing Bayesian data analysis: A tutorial with R, JAGS, and Stan*. Academic Press, 2014.

[22] A. Gelman, J. B. Carlin, H. S. Stern, D. B. Dunson, A. Vehtari, and D. B. Rubin, *Bayesian data analysis*. Chapman and Hall/CRC, 2013.

[23] B. Carpenter, A. Gelman, M. D. Hoffman, D. Lee, B. Goodrich, M. Betancourt, M. Brubaker, J. Guo, P. Li, and A. Riddell, "Stan: A probabilistic programming language," *Journal of statistical software*, vol. 76, no. 1, 2017.

[24] "Rossmann Store Sales. Forecast sales using store, promotion, and competitor data. Kaggle.Com." Available online: http://www.kaggle.com/c/rossmann-store-sales.

IEEE Third International Conference on Data Stream Mining & Processing
August 21-25, 2020, Lviv, Ukraine

Heteroskedasticity Models for Financial Processes Modelling and Forecasting

Nataliia Kuznietsova
Institute for Applied System Analysis
National Technical University of Ukraine "Igor Sikorsky KPI"
Kyiv, Ukraine
natalia-kpi@ukr.net

ORCID 0000-0002-1662-1974

Petro Bidyuk
Institute for Applied System Analysis
National Technical University of Ukraine "Igor Sikorsky KPI"
Kyiv, Ukraine
pbidyuke_00@ukr.net

ORCID 0000-0002-7421-3565

***Abstract* — This study is directed towards modelling financial processes which are widely available on financial markets and forecasting the prices of stock markets. It was approved many times that the financial market is really dynamically changeable, and variety of heteroskedastic models where chosen for studying for this reason. Such models as Autoregressive Moving Average (ARMA), Autoregressive Conditional Heteroskedastic Model (ARCH, and Generalized ARCH model (GARCH)), Exponential Generalized Model of Conditional Heteroscedastic Autoregression (EGARCH), Fractionally Integrated Generalized Model of Conditional Heteroskedastic Autoregression (FIGARCH) were discussed and investigated. An example of real data analysis from stock market illustrates the possibility and reasonability of application the heteroskedasticity models for predicting dynamics of variance.**

Keywords — stock market, Google stock prices, heteroscedasticity, variance, EGARCH, FIGARCH, GARCH, ARCH models

I. Introduction

Modern financial processes are characterized by high dynamics, non-stationarity and nonlinearity, large and time-varying volatility, the presence of deterministic and random components in time series data [1, 2]. Financial processes are affected by the multiplicity of random perturbations of different nature (noise components), which add significant uncertainties to the data analysis. The changes in financial processes are characterized by volatility estimates, but it cannot be described by homogeneous models due to the sharp changes in the factors influencing the financial processes. Therefore, to describe modern financial processes in the markets, exchange rates, processes in insurance companies such models should be used to take into account changes in both financial parameters and the process itself, compared to its previous states, as well as the processes and characteristics that affect it. In this research we consider the basic types of heteroskedastic models for volatility estimation of the financial processes and losses.

A lot of modern scientific studies are devoted to the volatility modelling and forecasting, in particular [3–6]. The peculiarities of financial processes require further improvement of mathematical models and methods for their parameters evaluation. Thus, the problem of obtaining the high-quality forecasts of this important statistical decision parameter requires substantial further research.

II. Models Review of Financial Processes Volatility Estimation

Volatility is a key parameter used to characterize pricing, portfolio optimization, VaR analysis and risk assessment as well as management techniques. Therefore, the "adequate" modelling and more accurate assessment of financial processes volatility is very important task in the financial sphere. In this article the modern mathematical models of heteroskedastic processes, i.e. processes with variable variance, in particular models of regression type are analysed. Formally, heteroskedastic process can be defined as follows:

$$\mathrm{var}[\varepsilon(k)] = \sigma_\varepsilon^2 \neq const. \qquad (1)$$

Heteroskedasticity means that the process variance decreases/increases in time, or is a more complex function of time, which could be found during constructing a model of the process under study.

A. Models with Variable Volatility

The time series of the financial process is a model developed for describing income volatility y_t as follows:

$$y_t = \mu_t + \sigma_t \varepsilon_t, \qquad \varepsilon_t \sim N(0,1), \quad t = 1,2,...,T,$$

$$\mu_t = \alpha + \sum_{i=1}^{k} b_i x_{i,t},$$

where μ_t, is the mean value depending on the constant α and the regression coefficients $b_1,...,b_k$; σ_t is the value of variable process volatility in the time. Variables $x_{1,t},...,x_{k,t}$ may also include non-system variables with time. The random process is assumed to have perturbations $\{\varepsilon_t\}$, of a normal distribution with zero mean and unity variance that is $\{\varepsilon_t\} \sim N\{0, 1\}$.

It is generally accepted that models with variable volatility can be divided into two classes: observable and parametric [7].

Both classes of models can be generally presented using the following structure: $y_t \mid z_t \sim N\left(\mu_t, \sigma_t^2\right)$.

The simplest examples of the models belonging to the first class are the Autoregressive Conditional Heteroskedasticity (ARCH) model [4] and the Generalized Autoregressive Conditional Heteroskedasticity model (GARCH) [8, 9].

In the class of nonparametric models z_t is a function of unobservable or hidden components. The log-normal stochastic volatility model proposed by Taylor in 1986 is the simplest and most famous example of a volatility model [10]:

$$\begin{aligned} & y_t \mid h_t \sim N\left(0, \exp\left(h_t\right)\right), \\ & h_t = \alpha + \beta\, h_{t-1} + \eta_t, \qquad \eta_t \sim NID\left(0, \sigma^2\right), \end{aligned} \tag{2}$$

where h_t is the log volatility that is unobservable but can be estimated using observable data; NID abbreviation means normally identically distributed values.

B. Autoregressive Conditionally Heteroscedastic Model (ARCH)

If the variance of sequence $\{\varepsilon(k)\}$ is non-constant then the trend for this parameter changing can be described using the autoregressive moving average (ARMA) model. For example, we can denote by $\{\hat{\varepsilon}(k)\}$ the first order model residuals (errors): $y(k) = a_0 + a_1 y(k-1) + \varepsilon(k)$. In this case the conditional variance of the principal variable is defined as

$$\begin{aligned} & \mathrm{var}[y(k+1) \mid y(k)] = E_k\{[y(k+1) - a_0 - a_1 y(k)]^2\} = \\ & = E_k[\varepsilon^2(k+1)]. \end{aligned}$$

It was still accepted here that $E_k[\varepsilon^2(k+1] = \sigma^2$ is a constant. Now let assume that the conditional variance is also a variable. One simple approach to formally describing such a variable is to apply a model of type AR (q) to the squares of residual estimates; for example:

$$\begin{aligned} & \hat{\varepsilon}^2(k) = \alpha_0 + \alpha_1 \hat{\varepsilon}^2(k-1) + \alpha_2 \hat{\varepsilon}^2(k-2) + \ldots + \\ & + \alpha_q \hat{\varepsilon}^2(k-q) + v(k), \end{aligned} \tag{3}$$

where $v(k)$ is the white noise process.

If all the coefficients $\alpha_1, \alpha_2, \ldots, \alpha_q$ are equal statistically zero, then the variance estimate is simply a constant. Otherwise, the conditional variance for the process $y(k)$ is described by equation (3). It is possible to use this equation to predict conditional variance by one step ahead as follows:

$$\begin{aligned} & E_k[\hat{\varepsilon}^2(k+1)] = \alpha_0 + \alpha_1 \hat{\varepsilon}^2(k-1) + \alpha_2 \hat{\varepsilon}^2(k-2) + \ldots + \\ & + \alpha_q \hat{\varepsilon}^2(k+1-q). \end{aligned}$$

While the residuals $\varepsilon(k)$ used in equation (3) can be obtained from regression or autoregression with a moving average equations, a lot of potential practical and theoretical applications can be found for the ARCH model [11, 12].

In addition to the type (3) equation, more complex forms for description of variance behaviour can be selected. For example, the process perturbations can be incorporated in a multiplicative form:

$$\varepsilon^2(k) = v^2(k)[\alpha_0 + \alpha_1 \varepsilon^2(k-1)], \tag{4}$$

where $v(k)$ is multiplicative perturbation in the form of white noise, moreover, $\{v(k)\} \sim (0,1)$ that is, it has zero mean and unity variance; variables $\varepsilon(k-1)$ and $v(k)$ have statistically independent (uncorrelated) values.

The model of the heteroskedastic process (4) can be extended to any model order and written as:

$$\varepsilon(k) = v(k)\left(\alpha_0 + \sum_{i=1}^{q} \alpha_1\, \varepsilon^2(k-i)\right)^{1/2}. \tag{5}$$

In this equation the random variable $\varepsilon(k)$ is influenced by all the values from $\varepsilon(k-1)$ to $\varepsilon(k-q)$, and therefore the conditional variance can be regarded as an autoregressive process of order q.

C. Generalized Autoregressive Model with Conditional Heteroskedasticity (GARCH)

A further extension of the ARCH model is description of conditional variance as ARMA process. Let the model errors be described by the equation $\varepsilon(k) = v(k)\ [h(k)]^{1/2}$, where $\sigma_v{}^2 = 1$, and

$$h(k) = \alpha_0 + \sum_{i=1}^{q} \alpha_i \varepsilon^2(k-i) + \sum_{i=1}^{p} \beta_i h(k-i). \tag{6}$$

Since the process $\{v(k)\}$ is defined as white noise, which is not correlated with the values of $\varepsilon(k-i)$, the conditional and unconditional mean for $\varepsilon(k)$ are equal to zero. Obviously, an unconditional expectation $E[\varepsilon(k)] = E[v(k)\ (h(k))^{1\backslash 2}] = 0$. Conditional variance of the variable $\varepsilon(k)$ is defined as $E_{k-1}[\varepsilon^2(k)] = h(k)$.

The generalized ARCH model, called GARCH (p, q), consists of two components: autoregression and a moving average for the variance of the heteroscedastic process. The first order process is obtained at, $p = 0$, $q = 1$, and it can be formally defined as the GARCH (0,1). If all the coefficients, $\beta_i = 0,$, then the model GARCH (p, q) is equivalent to the model ARCH(p). In order to ensure that the conditional variance is finite, the roots of the characteristic equation written for (6) must lie within a circle of unit radius.

The main feature of the GARCH model is that the perturbation acting on the process $\{y(k)\}$, is the ARMA process. Therefore, it can be expected that the residuals (errors) of the ARMA model (the process model) will be in accordance with the characteristics of the heteroscedastic process. This statement can be explained as follows. Let $\{y(k)\}$ is the

ARMA process. If the ARMA model is adequate for the process, then the autocorrelation function (ACF) and partial autocorrelation function (PACF) of the residuals should indicate that this is a white noise process. On the other hand, ACF residual squares can be used to predefine the order of the GARCH process. While, $E_{k-1}[\varepsilon(k)] = (h(k))^{1/2}$, so equation (6) can be written in the following form:

$$E_{k-1}[\varepsilon^2(k)] = \alpha_0 + \sum_{i=1}^{q} \alpha_i \varepsilon^2(k-i) + \sum_{i=1}^{p} \beta_i h(k-i). \quad (7)$$

D. Exponential Generalized Model of Conditional Heteroscedastic Autoregression (EGARCH)

The Exponential GARCH Model (EGARCH) is an ARCH model that takes into account the asymmetric momentum and volatility of good and bad news regarding the results of operations at the exchange. The EGARCH model with a special variable that distinguishes between the volatility of good and bad news is presented in [6]; EGARCH (p, q) model:

$$y_t = \sigma_t \varepsilon_t, \quad t = 1, 2, \ldots, T$$

$$\log\left(\sigma_t^2\right) = \alpha_0 + \left[1 - \sum_{j=1}^{q} \beta_j B^j\right]^{-1} \left[1 + \sum_{i=1}^{p} \alpha_i B^i\right] g\left(\varepsilon_{t-1}\right), \quad (8)$$

where all variables and parameters are defined in the same way as in ARCH model; in addition, B^j та B^i are linear shift operators, and $g(\varepsilon_t) = \lambda_1 \varepsilon_1 + \lambda_2 \left(|\varepsilon_t| - E\left(|\varepsilon_t|\right)\right)$, where λ_1, and λ_2, are extra parameters. In EGARCH (p, q) models with λ_1, the bad news will have a greater impact on the volatility than the good news impulses of the corresponding variable.

E. Fractionally Integrated Generalized Autoregression with Conditional Heteroscedasticity (FIGARCH)

Many financial time series are characterized by high invariance of volatility at selected time intervals. Here important is existence of autocorrelation for various volatility measures. This property is called long-term volatility memory [12, 13, 14]. Consider the long-memory model, named the Fractionally Integrated GARCH (FIGARCH (p, q)) model, proposed by Bailey in 1996 [12]:

$$y_t = \sigma_t \varepsilon_t, \quad t = 1, 2, \ldots, T,$$

$$\sigma_t^2 = \alpha_0 \left[1 - \sum_{j=1}^{q} \beta_j B^j\right]^{-1} +$$

$$+ \left\{1 - \left[1 - \sum_{j=1}^{q} \beta_j B^j\right]^{-1} \sum_{i=1}^{p} \alpha_i B^i (1-B)^d\right\} y_t^2,$$

where all variables and parameters are defined as in ARCH models, except that B^j and B^i are operators for which: $x_t B^j = x_{t-j}$, and d is the fractional difference parameter.

The FIGARCH (p, q) model describes "long-memory" of financial volatility due to parameter, d. If $0 < d < 1$, then the conditional volatility σ_t^2 will slowly dampen with hyperbolic rate, what is a typical sign of "good memory" [12].

III. An Example of Forecasting the Volatility of Financial Time Series

Google's stock prices for the international currency market from January 2015 to February 2020 were selected as input for the simulation. There are 1276 records in total and 5 attributes for each record, namely: <OPEN>, <HIGH>, <LOW>, <CLOSE>, <VOL>. The target variable was to predict the stock price at the time of closing on a current day. Since the data is presented as time series, let's use the models described above to predict the price and volatility of the stock (Fig. 1).

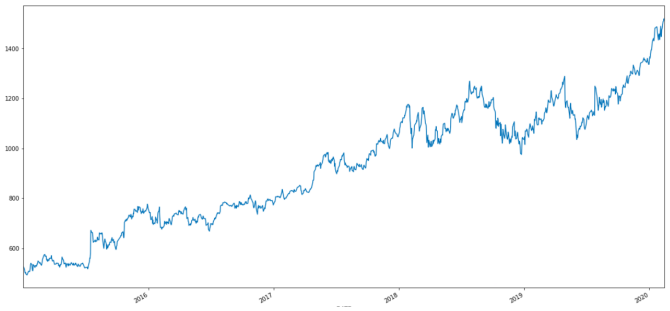


Fig. 1. A view of the time series represented by Google stock prices

To build the process model, one must identify whether the series described are stationary or non-stationary. There are several stationarity tests, such as Dickey-Fuller test, its modifications, and KPSS test described in [15–17].

The results of the Dickey-Fuller test showed that adf = 0.15589, and p-value = 0.96967. The results obtained indicate that the assumption of the series stationarity is not confirmed and the time series is non-stationary. The non-stationarity of a series may be caused by time variable expectation or variance, or both at once. In order to check for a trend, one has to do some manipulations with the series, including trying to break it into separate components and highlight the trend. We will perform the multiplicative decomposition of the series and thus make a study of trend, seasonality and emissions (Fig.2). As can be seen from the figures, the series clearly contains linear trend, but the seasonal effects and scattering of the noise components are negligible, the average value of these components is 1, so their effect on the series is rather weak.

To reduce the effects of random perturbations, we transform the series to stationary form. We take the logarithm of the data and calculate the first difference (since it was suggested that there is a first-order trend), which will remove the trend. The time series after the transformations is presented in Fig. 3. Next, we re-check the series for stationarity by calculating the Dickie-Fuller statistics after removing the trend. The values obtained are: adf = -12.96252, and p-value: 3.20315e-24, indicating that the series became stationary after its trend was removed. So let's fix this the fact

and build the following models for the initial series, given the first-order trend.

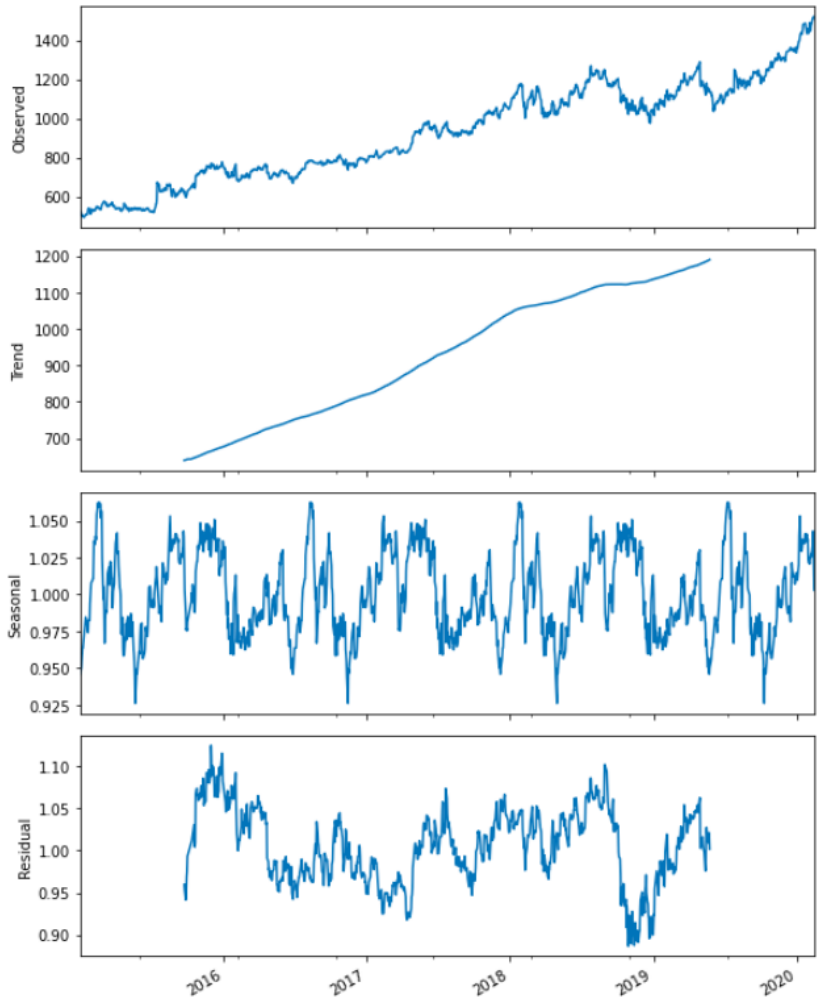


Fig. 2. Checking the time series for trends, seasonal effects and anomalous emissions

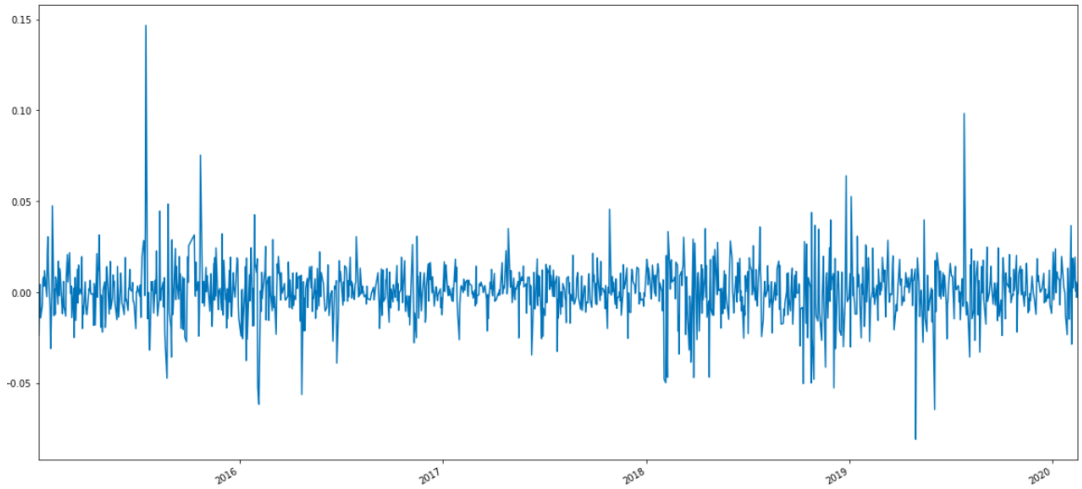


Fig. 3. Time series after deleting the first order trend

Since after the described transformations the new series becomes stationary, the models for stationary series can be applied. Let's construct an autoregressive model with a moving average, ARMA, and calculate the values of the coefficients p and q. To do this, compute ACF and PACF for 20 lags, choose the order of the model for which the PACF is greater than the threshold. The order of the model p = 8 (the coefficient most different from 0 in PACF) and q = 4 for the moving average part were determined. For this model, AIC = -7117.504, and BIC = -7090.413.

While ARMA model was built for the input series after removing the trend component, it is necessary to return to the input initial row to solve the problem of stock price forecasting (without logarithm and remove the first difference). The prediction error value was obtained by RMSE 2.3693, which is quite small relative to the scale of the series and indicates the high accuracy of the model [17, 18].

Since the first step revealed the first-order trend, it makes sense to construct an auto-regression model with integrated moving average (ARIMA) to predict the values of a series. We construct an ARIMA model with parameter d = 2 (Fig. 4). Similarly, select the parameters of the autoregressive component and the moving average and construct ARIMA (1, 2, 1). Statistical quality characteristics: AIC = -7090.791, BIC = -7070.182.

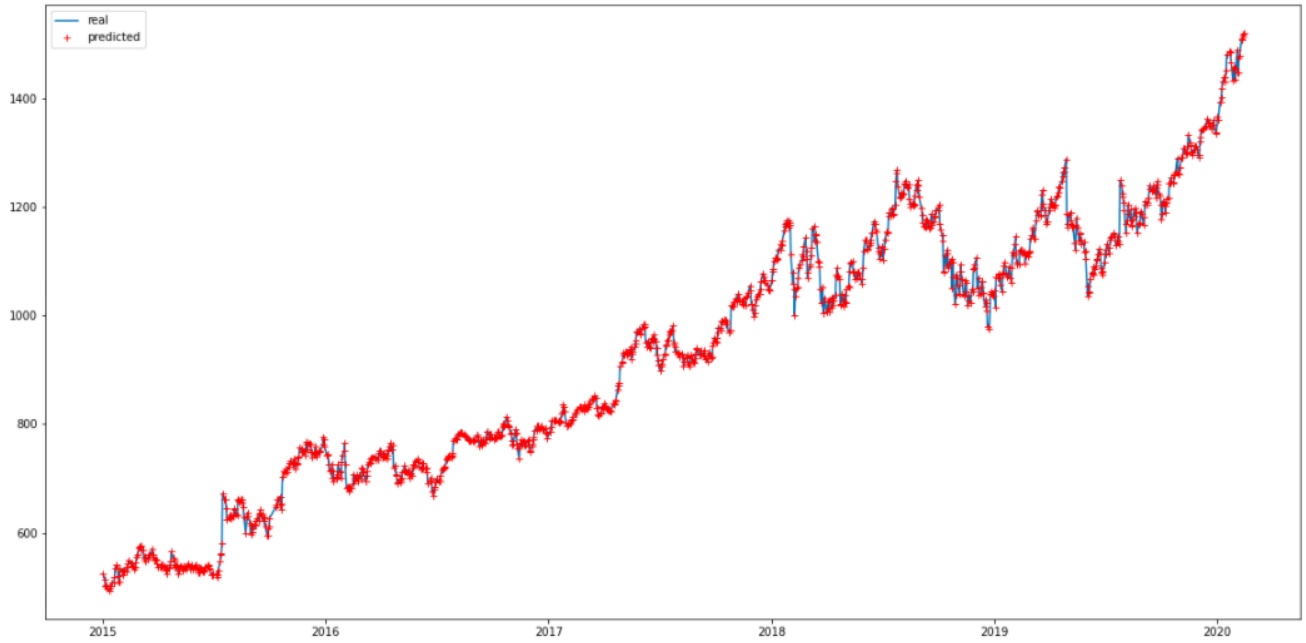


Fig. 4. Modeling results based on the ARIMA model (1, 2, 1)

The quality of the ARIMA model (1, 2, 1) was higher, in particular the value of the stock price prediction error was as follows: RMSE = 0.0211.

TABLE I. COMPARISON OF THE MODEL PERFORMANCE FOR GOOGLE STOCK PRICE FORECASTING

Model	R^2	Sum squared residuals	Akaike	DW
ARMA(8, 4)	0.9999	7179.9811	-7117.504	1.996
ARIMA(1, 2, 1)	1.0000	0.5545	-7090.791	1.995

The residuals for heteroskedasticity were checked using the Breusch-Pagan Test, and it was found that the F-Test with p-value is very small and so the residuals are heteroskedastic. Therefore, the next step was to carry out forecasting of financial process variance by constructing heteroscedastic models for forecasting the variance with the following characteristics: GARCH (11, 2), EGARCH (2, 1, 1), FIGARCH (1, 0), the quality of estimation for which is given in Table II.

TABLE II. QUALITY OF MODELS FOR ESTIMATING FINANCIAL SERIES VARIANCE

Model	AIC	BIC	DW
ARCH(2)	-7229.68	-7209.07	1.9958
GARCH(1, 1)	-7237.21	-7216.60	1.9956
EGARCH(2, 1, 1)	-7278.46	-7247.54	1.9960
FIGARCH(1, 0)	-7260.01	-7239.40	1.9500

We apply the obtained model parameter estimates to predict the future value of financial market volatility using the developed program. By performing static forecasting consistently at each step, it is possible to compare the model-based forecast and the actual observed price values (Table III and Table IV).

TABLE III. COMPASON OF THE FORECASTS QUALITY FOR THE ABSOLUTE VALUE OF GOOGLE STOCK PRICE

Model	RMSE	RSS
ARMA(8, 4)	0.01535	0.0031
ARIMA(1, 2, 1)	0.01568	0.0050

TABLE IV. THE RESULTS OF GOOGLE'S STOCK PRICE VOLATILITY FORECAST FOR ONE-STEP AHEAD

Model	Forecast	Absolute difference
ARCH(2)	0.0001505	0.00003636
GARCH(1, 1)	0.0001416	0.00002746
EGARCH(2, 1, 1)	0.0001064	0.00000774
FIGARCH(1, 0)	0.0001377	0.00002356

The best result of one-step ahead forecasting of Google's stock price volatility (financial series variance) was achieved with the exponential autoregressive model with conditional heteroscedasticity EGARCH (2, 1, 1), and the prediction accuracy is up to the fifth decimal point, indicating solving the prediction problem with acceptable accuracy.

IV. CONCLUSIONS

In the study the modern models for evaluating financial processes that allow taking into account the rapid fluidity and volatility of financial processes, forecasting the value of financial market indicators and losses for the next moments of time were analysed. The use of such models is appropriate for estimating potential losses from fluctuations of the financial indicators, and involves assessing relative risks through the magnitude and probability of the losses [19].

The experimental studies performed have shown that heteroscedastic models is good for predicting financial process variance, and such models are useful in assessing financial market fluctuations. The results of computational experiments with actual data regarding volatility estimation and forecasting indicate the possibility of achieving high-quality results for short-term forecasting. The future studies will be directed towards refinement of existing volatility models structures and constructing specialized decision support system allowing for automation of the data analysis process as a whole, as well as forecast and financial risk estimation.

REFERENCES

[1] R.S. Tsay, Analysis of financial time series. New York: John Wiley & Sons, Inc., 2010, 715 p.

[2] P.I. Bidyuk, N.V. Kuznietsova, "Forecasting the volatility of financial processes with conditional variance models," Journal of Automation and Information Sciences, no. 46 (10), 2014, pp. 11 – 19.

[3] R.Y. Chou, "Volatility persistence and stock returns – some empirical evidence using GARCH," Journal of Applied Econometrics, n. 3, 1987, pp. 279 – 294.

[4] R. F. Engle "Autoregressive conditional heteroscedasticity with estimates of the variance of United Kingdom inflation," Econometrica, vol. 50, 1982, pp. 987–1007.

[5] R. Harris, R. Sollis, Applied Time Series Modelling and Forecasting. West Sussex: Jihn Wiley & Sons Ltd., 2005, 313 p.

[6] D. B. Nelson, "Conditional heteroscedasticity in asset returns: a new approach," Econometrica, vol. 59, no. 2, 1991, pp. 347–370.

[7] T. Bollerslev, R. F. Engle, D. Nelson, ARCH models. Handbook of Econometrics, Eds. R. Engle and D. McFadden. North-Holland, Amsterdam, vol. 4, 1993, pp. 2959–3038.

[8] J-P. Peters, "Estimating and forecasting volatility of stock indices using asymmetric GARCH models and (Skewed) Student-t densities," Ecole d'Administration des Affaires, University of Liege, Belgium, 2001. 20p. URL: https://pdfs.semanticscholar.org/f445/493c46b3b4d52e958f69d502ccf8f03d51c3.pdf.

[9] S. J. Taylor, "Financial returns modeled by the product of two stochastic processes – a study of the daily sugar prices: 1961 – 1975. Time Series Analysis: Theory and Practice," Anderson, O. D. (ed.), 1. – Amsterdam: North-Holland, 1982, pp. 203–226.

[10] S. J. Taylor, Modelling Financial Time Series. Chichester: John Wiley, 1986. 268 p.

[11] T. Bollerslev, "Generalized autoregressive conditional heteroskedasticity," Journal of Econometrics, vol. 31, 1986, pp. 307–327.

[12] R. T Baillie, T. Bollerslev, H. O. Mikkelsen, "Fractionally Integrated Generalized Conditional Heteroskedasticity," Journal of Econometrics, vol. 74, 1996, pp. 3–30.

[13] Z. Ding, C. W. J. Granger, R. F Engle, "A Long Memory Property of Stock Returns," Journal of Empirical Finance, vol. 1, 1993, pp. 83–106.

[14] T. Bollerslev, H. O. Mikkelsen, "Modeling and pricing long memory in stock market volatility," Journal of Econometrics, vol. 73, 1996, pp. 151–184.

[15] T. L. Dubrova, Statistical methods of forecasting. Moscow: UNITI-DANA, 2003, 205p.

[16] D. Kwiatkowski, P. Phillips, P. Schmidt., Y. Shin, "Testing the null hypothesis of stationarity against the alternative of a unit root," Journal of Economics, n. 54, 1992, pp. 159–178.

[17] N. V. Kuznietsova, M. Seebauer, S. Zabielin, "Some methods for estimating financial risks in banking," IEEE 1st Conf. on System Analysis and Intelligent Computing, SAIC 2018, 2018, pp. 271–274. DOI: https://ieeexplore.ieee.org/document/8516873.

[18] J. D. Hamilton, Time Series Analysis. Princeton, NJ: Princeton University Press, 1994, 815 p.

[19] N. V. Kuznietsova, P.I. Bidyuk, "Modeling of credit risks on the basis of the theory of survival," Journal of Automation and Information Sciences, n. 49(11), 2017, pp. 11-24

IEEE Third International Conference on Data Stream Mining & Processing
August 21-25, 2020, Lviv, Ukraine

Software and Algorithmic Support for Finite Element Analysis of Spatial Heat-and-Moisture Transfer in Anisotropic Capillary-Porous Materials

Yaroslav Sokolovskyy
Department of Information Technologies
Ukrainian National Forestry University
Lviv, Ukraine
sokolowskyyyar@yahoo.com

Andriy Nechepurenko
Department of Information Technologies
Ukrainian National Forestry University
Lviv, Ukraine
andriy.nechepurenko@gmail.com

Tetiana Samotii
Department of Information Technologies
Ukrainian National Forestry University
Lviv, Ukraine
t.samotiy@nltu.edu.ua

Svitlana Yatsyshyn
Department of Information Technologies
Ukrainian National Forestry University
Lviv, Ukraine
svitlana0981@gmail.com

Olha Mokrytska
Department of Information Technologies
Ukrainian National Forestry University
Lviv, Ukraine
mokrytska@nltu.edu.ua

Volodymyr Yarkun
Department of Information Technologies
Ukrainian National Forestry University
Lviv, Ukraine
yarkun@nltu.edu.ua

***Abstract* — On the basis of a three-dimensional mathematical model of non-isothermal heat-and-mass transfer in capillary-porous materials, taking into account the anisotropy of thermophysical properties, software was developed to perform finite element analysis of moisture transfer using CUDA technology. The program database is divided into two parts. One part is located on the local device where the application is installed and the other part is on the Azure server. This solution made it possible to use CUDA parallelization software from Azure Cloud add-ons or from the local device where the program is located. The application of parallel technologies using the cuBLAS library made it possible to transfer all large-scale matrix computations to the graphics processor, which reduced resource consumption with increasing grid density.**

Keywords* — *software, mathematical model, heat-and-moisture transfer, finite element method, CUDA, cuBLAS, Azure

I. INTRODUCTION

At the present time, we are constantly working on improving the technologies associated with the drying of wood. The research results and methods developed are currently undergoing testing at the enterprises to identifiy all the new advantages and disadvantages of the scientific approaches. Most of the methods are predominantly developed by conducting studies on problems of heat-and-moisture transfer in capillary-porous materials using the finite element method. Since most software resources are spent on discretization and finite element calculations when software is being run, it is important to find the optimal technology for parallelizing such processes. It should also be noted that the construction of a three-dimensional mathematical model and the study of spatial case of non-isothermal heat-and-mass transfer remain relevant. Therefore, the choice of an effective numerical method for implementing the 3D approach and the correct use of CUDA parallel computing technologies in conjunction with Azure cloud technologies can give impetus to the creation of an object-oriented software package. The software should provide for the implementation of a user-friendly interface with extensive graphical visualization of finite element partitioning; it is necessary to provide the results of simulations of moisture transfer process and to implement effective methods of working with data in cloud technologies.

II. MATHEMATICAL MODEL

The system of differential equations (1)-(2) in partial derivatives describes a three-dimensional mathematical model of heat-and-mass transfer during drying of capillary-porous materials, taking into account the anisotropy of thermophysical properties.

$$\rho_0 c\frac{\partial T(x,y,z,\tau)}{\partial \tau}=\frac{\partial}{\partial x}\left(\lambda_x\frac{\partial T(x,y,z,\tau)}{\partial x}\right)+\frac{\partial}{\partial y}\left(\lambda_y\frac{\partial T(x,y,z,\tau)}{\partial y}\right)+\frac{\partial}{\partial z}\left(\lambda_z\frac{\partial T(x,y,z,\tau)}{\partial z}\right)+\varepsilon\rho_0 r\left(\frac{\partial U(x,y,z,\tau)}{\partial \tau}\right), \quad (1)$$

$$\frac{\partial U(x,y,z,\tau)}{\partial \tau}=\frac{\partial}{\partial x}\left(a_{m_x}\frac{\partial U(x,y,z,\tau)}{\partial x}\right)+\frac{\partial}{\partial y}\left(a_{m_y}\frac{\partial U(x,y,z,\tau)}{\partial y}\right)+\frac{\partial}{\partial z}\left(a_{m_z}\frac{\partial U(x,y,z,\tau)}{\partial z}\right)+\frac{\partial}{\partial x}\left(a_{m_x}\sigma\frac{\partial T(x,y,z,\tau)}{\partial x}\right)+\frac{\partial}{\partial y}\left(a_{m_y}\sigma\frac{\partial T(x,y,z,\tau)}{\partial y}\right)+\frac{\partial}{\partial z}\left(a_{m_z}\sigma\frac{\partial T(x,y,z,\tau)}{\partial z}\right) \quad (2)$$

with boundary conditions:

$$\lambda_n \frac{\partial T}{\partial n}\Big|_{x=S} + \rho_0 (1-\varepsilon) r \beta_n \left(U|_{x=S} - U_{p_n}\right) = \alpha_n \left(T_c - T|_{x=S}\right);$$

$$\left(a_{m_n}\frac{\partial U}{\partial n} + a_{m_n}\sigma\frac{\partial T}{\partial n}\right)\Big|_{x=S} = \beta_n \left(U_{p_n} - U|_{x=S}\right); \quad (3)$$

$$\left(a_{m_n}\frac{\partial U}{\partial n} + a_{m_n}\sigma\frac{\partial T}{\partial n}\right)\Big|_{x=0} = 0; \ \frac{\partial T}{\partial n}\Big|_{x=0} = 0;$$

and initial conditions for irregular moisture removal:

$$T|_{\tau=0} = T_0(x,y,z); \ U|_{\tau=0} = U_0(x,y,z), \quad (4)$$

where $T(x,y,z,\tau)$, $U(x,y,z,\tau)$ – are the desired functions, respectively, of temperature and moisture content in the region

$$V = \left\{(x,y,z): x\epsilon[0,R_1], y\epsilon[0,R_2], z\epsilon[0,R_3], \tau\epsilon\left[0,\tau^*\right]\right\};$$

FEM is finite element method; R_1, R_2, R_3 are geometric dimensions of lumber; S is geometric surface of the region; τ^* is time of the process; c is heat capacity of wood; ρ_0 is heat capacity and base density of wood; $\lambda_1, \lambda_2, \lambda_3$ are coefficients of thermal conductivity in the directions of anisotropy; ε is phase transition criterion; τ is current time; r is specific heat of vaporization; α_n is heat exchange coefficient; a_{mx}, a_{my}, a_{mz} are coefficients of moisture conductivity in the directions of anisotropy; T_c is ambient environment; U_p – equilibrium humidity; β_n is moisture exchange coefficient; σ is thermo-gradient coefficient; n is normal vector to the surface.

For the regular mode, characterized by the corresponding values of the criterion Fourier numbers (F0) and Bio (Bi), the initial conditions are specified by spatial quadratic functions[16, 19]. Software was developed for the numerical implementation of the finite element algorithm and for the analysis of the mathematical model of heat-and-moisture transfer in capillary-porous materials [12,15].

III. Algorithmic Aspects

The variational formulation of the mathematical model of heat-and-mass transfer follows from the finite element method. In turn, the numerical implementation of the mathematical model (1) - (4) is also provided by the use of FEM. To set the problem, the assumption is made that the change in moisture content can be represented as the sum of the values due to the gradients of moisture content and temperature $\frac{\partial U}{\partial \tau} = \frac{\partial U_1}{\partial \tau} + \frac{\partial U_2}{\partial \tau}$. Then equation (2) can be written:

$$\frac{\partial U_1}{\partial \tau} = \frac{\partial}{\partial x}\left(a_{m_x}\frac{\partial U_1}{\partial x}\right) + \frac{\partial}{\partial y}\left(a_{m_y}\frac{\partial U_1}{\partial y}\right) + \frac{\partial}{\partial z}\left(a_{m_z}\frac{\partial U_1}{\partial z}\right), \quad (5)$$

$$\frac{\partial U_2}{\partial \tau} = \frac{\partial}{\partial x}\left(a_{m_x}\sigma\frac{\partial U_2}{\partial x}\right) + \frac{\partial}{\partial y}\left(a_{m_y}\sigma\frac{\partial U_2}{\partial y}\right) + \frac{\partial}{\partial z}\left(a_{m_z}\sigma\frac{\partial U_2}{\partial z}\right). \quad (6)$$

Taking into account the assumption, the implementation of the mathematical model is reduced to solving the equivalent variational problem based on the minimization of functionals [3].

To approximate continuous values of temperature T and moisture content in the wood U in the discrete model, the study region V is divided into a finite number of elements. Then the moisture content $\{U^{(e)}\}$ and the temperature $\{T^{(e)}\}$ are determined for each discretization element:

$$\{T^{(e)}\} = [N^{(e)}]\{T\}, \{U^{(e)}\} = [N^{(e)}]\{U\}. \quad (7)$$

We introduce the notation

$$\{g_T^{(e)}\}^T = \left\{\frac{\partial T}{\partial x} \ \frac{\partial T}{\partial y} \ \frac{\partial T}{\partial z}\right\}, \{g_U^{(e)}\}^U = \left\{\frac{\partial U}{\partial x} \ \frac{\partial U}{\partial y} \ \frac{\partial U}{\partial z}\right\}. \quad (8)$$

Then

$$\{g_T^{(e)}\}^T = [B^{(e)}]\{T\}, \{g_U^{(e)}\}^U = [B^{(e)}]\{U\} \quad (9)$$

where $\{U\}$ is the moisture content vector; $\{T\}$ is temperature vector; $[N^{(e)}]$ are element shape functions, $[B^{(e)}]$ is gradient matrix.

The first-order tetrahedron is a finite element [4]; therefore, taking into account the functions of shape and differentiation of functionals with respect to $\{U\}$ and $\{T\}$, we obtain their minimum. The resulting system of FEM matrix equations follows after grouping the integrals and equating them to zero:

$$\begin{aligned} [C]\frac{\partial\{U\}}{\partial\tau} + [K]\{U\} + \{F\} &= 0; \\ [\overline{C}]\frac{\partial\{T\}}{\partial\tau} + [\overline{K}]\{T\} + \{\overline{F}\} &= 0, \end{aligned} \quad (10)$$

where:

$$[C] = \int_V \rho_0 [N]^T [N] dV; \ [K] = \int_V [B]^T [D^*][B] dV +$$

$$+ \int_S \rho_0 \beta_w [N]^T [N] dS; \{F\} = \int_V [B]^T [H][B][T] dV - \int_S \rho_0 \beta U_p [N]^T dS$$

– according to the matrix of thermo-physical properties of material, damping and loading, {N} – matrix of form functions. Analogical matrixes are those $[\overline{C}], [\overline{K}], \{\overline{F}\}$, related with the coefficient of heat conductivity and heat exchange.

Finding the spatial functions of moisture content and temperature at any point in the time interval is reduced to finding the solution to the system of equations (10) using the algorithm predictor - corrector. For the values of change of temperature {T} and humidity {U} in time the Finite difference method is used[13,14]. Then numeral realization of mathematical model (7) – (8) is taken to the decision of the kind of system equalizations:

$$[A]\{U\}_{next} = \{R\}; \qquad [A_T]\{T\}_{next} = \{R_T\}.$$

As thermo-physical descriptions of wood depend on a temperature and humidity, and equalization of model(5) – (6) are related with each other, the iteration process of equalizations realization (10) is carried out on every sentinel step taking into account additional iteration procedure, which specifies influence of humidity on apportionment of temperature in material and vice versa. Completing of iterations for equalizations (10) foresees implementation of conditions: {Un} – {Un-1} ≤ 10-4 i {Tn} – {Tn-1} ≤ 10-4.

Discretization of the study area is one of the main steps in solving the problem of moisture transfer using FEM. The task is accomplished through the construction of a three-dimensional finite-element grid which in turn is created by building up the layers of triangulations constructed at the initial stage. The construction of tetrahedrons is possible due to the establishment of connections between triangular faces. The discretization uses the data structure Tetrahedrons-Triangles-Nodes [3,10]. The algorithm Direct construction of the Delaunay triangulation was chosen as the basis, namely, a step-by-step algorithm with a k-d search tree. The key point is the quality assessment of the constructed tetrahedron in the three-dimensional region. This is especially true for a group of algorithms of direct construction, because their feature is the construction of immediately regular tetrahedrons without further restructuring [3]. The coefficient for assessing the quality (performance index) of the constructed tetrahedron is calculated as the ratio of the volume of the tetrahedron to the sum of the areas of its faces: restructuring

$$\gamma = \frac{12\sqrt{3}\left(AB \times AC\right) \bullet AD}{\left(AB^2 + BC^2 + CA^2 + AD^2 + BD^2 + CD^2\right)^{3/2}} \qquad (11)$$

where *A, B. C. D* are apices of the tetrahedron.

The normalizing factor is 12√3. Accordingly, the maximum volume of an equilateral tetrahedron will be 1. If the coefficient of quality assessment of the tetrahedron is equal to 0, then it must be removed.

IV. Software Implementation

To develop software in the context of object-oriented programming, the main entities of the finite element method were identified. To carry out finite element calculation, software was used in the Microsoft Visual Studio environment in C # using CUDA parallel computing technology. The software application includes three modules: area discretization, system matrix formation, and equation system solution [6,7].

The three-level software architecture pattern is taken as the basis for the implementation of moisture transfer modeling (Fig. 1). It integrates three main components: the client, the application server, and the database server. The client is the user interface. The application server connected to the database server is located in Azure Cloud which is connected to the local database. Communication with the database can be performed simultaneously by several client terminals. The sequence of queries in such an architecture is shown in Fig. 1. For a better understanding of the architecture, a detailed description is provided below. After entering the input data by the user, the application sends a request to the environment, in which the interface is running, about the possibility of using CUDA. The software application receives a response about the capabilities of the graphics processor and performs the calculation of the task. Where calculations will take place depends on the availability of CUDA cores. If they are not available in the working environment, then the software sends data to Azure Cloud, where calculations are made using the cloud technology. Then the results are returned to the program. The results of the calculations are simultaneously stored in two databases: in the local database and in the database hosted in Azure. With an Internet connection, these databases are synchronized. Thus, the user can have access to new data even with limited access to the Internet.

The use of CUDA parallel computing technologies is determined by the resource-intensiveness of such things as high-dimensional matrix operations and the implementation of triangulation and its reflection. All matrix operations were implemented using the Nvidia cuBLAS library.

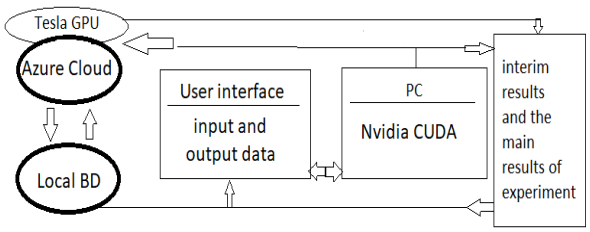


Fig. 1. Software architecture.

The use of the API cuBLAS with increasing volume of input data and an increase in the discretization of the grid made it possible to increase the speed of this type of software by 40%. This is primarily due to the ability of cuBLAS to accelerate matrix operations by performing computational operations in a single GPU or effectively zooming in and distributing work in the configurations of several GPUs (Graphics Proccesing Unit). The advantages that also affect performance are: full support for all 152 standard cuBLAS procedures; API and error log for debugging and tracking; CUDA threads support for simultaneous operations.

The software is configured in such a way that, if it is not possible to use CUDA technologies on the local server where the user application is installed, the request for processing input data is redirected to Azure Cloud and calculations are performed using the Nvidia Tesla graphics accelerators (Fig. 2).

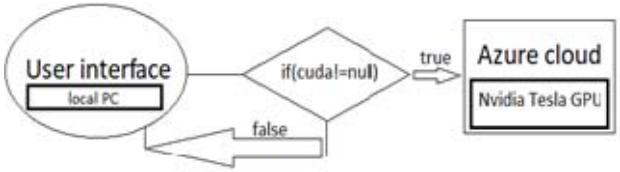


Fig. 2. The architecture of requests for carrying out parallel computing.

The application of the cuBLAS library takes precedence over operations with matrices and arrays in CUDA. Unlike CUDAfy, the matrix multiplication operation does not need making memory copies. API BLAS and LAPACK were created in such a way that when specifying the starting point, step length, size length, etc., it is possible not to do the extra work of copying CPU-GPU-CPU data. The following is a code snippet (cuBLAS) from the developed software [20].

```
int M = N * P; int K = N * P;
for (int i = 0; i < N; i++)
{
        double *d_In = d_A + i * P;
        double *d_Out = d_C + i * P;
        cublasSetStream(streams[i]);
```

```
    cublasDgemm(cublasHandle, CUBLAS_OP_N, CUBLAS_OP_N,
&alpha, d_In, M, d_B, M, &beta, d_Out, K);
    }
```

The advantage is the approach of partial use of CUDA parallelization in two parts of software. This approach is caused by loading into the central processor of discretization and matrix computing of the FEM. Thanks to the parallelization of the most resource-intensive parts of the program, GPU actually unloaded the CPU and made it a kind of software manager which shapes the sequence of tasks and monitors their execution.

Since nodes are the first element in the software data structure, it is necessary to combine them into a k-d tree in order to speed up the software operation [10,12]. The essence of building a tree lies in dividing the plane into two half-planes by a straight line that is parallel to one of the coordinate axes, for example, the OY axis. Later on, each of these half-planes can be divided repeatedly by a straight line parallel to another coordinate axis. These steps are repeated, changing the direction of the dividing line at each step. The choice of lines dividing the plane is conditioned by the principle of simple dichotomy, that is, the number of nodes on each side of the dividing line should be approximately equal.

The implementation of discretization uses the algorithm of direct construction which is quite laborious. But the decision to transfer multiple checks of the fulfilment of the Delaunay condition from the CPU to the GPU threads makes this method the most successful option for application in this area of research. The number of parallel threads used in the software implementation is equal to the number of operations required to check the Delaunay condition (calculation of the center and radius of the circle circumscribed around a triangle). Accordingly, each block thread is responsible for a certain operation necessary for calculating the Delaunay criterion [2]. In particular, block (0; 0) is responsible for calculating the Delaunay criterion for a triangle, etc.

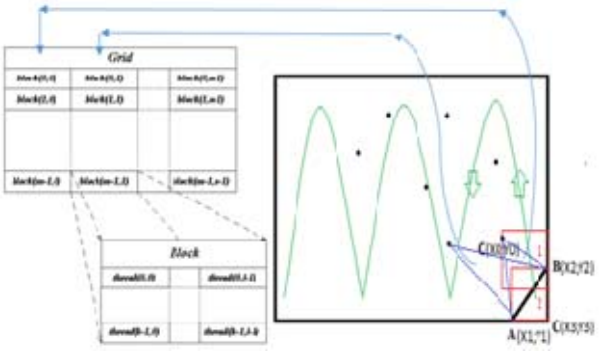


Fig. 3. The thread diagram of parallel execution of operations to verify the Delaunay condition.

V. Number Experiment

Having examined the mathematical model (1) - (4) and the developed software, we studied the dynamics of spatial heat-and-moisture transfer in wood during the drying process. For this purpose, the following environmental parameters were used: $T_s = 80°C$; relative humidity $\varphi = 60\%$ [1,18]; characteristics of pine wood: $a_{m_x} = 0.97\,cm^2/s$, $a_{m_y} = 1.3\,cm^2/s$, $a_{m_z} = 1.8\,cm^2/s$, $\rho_0 = 450\,kg/m^3$; coefficients of thermal conductivity and moisture conductivity: $\alpha = 23W/(m^2 \cdot K), \beta = 2 \cdot 10^{-6}\,m/s$ [2, 8]. Other characteristics and necessary empirical dependencies were obtained on the basis of experimental data [4,5,11]. Especially, on the basis of working of experimental data dependence of coefficient of wood hydraulic conductivity as functions from a temperature and humidity is used: am1(T,U) = amt(T) amu(U), am1 / am2 = 1,25. For determi-nation of coefficient of humidity exchange we use the dependence: $\alpha = 0{,}95$ (T/φε exp(-2σVp /rTR)) 10-9, where Vp, σ – molar volume and superficial strain of liquid, φ – relative humidity of environment. Value r = r(U) have been received on the basis of wood structure modeling by the system of inconstant capillaries of radius r, that depends on humidity [9, 17].

Figs. 4 - 5 show the dynamics of changes in moisture content $U^* = (U - U_0)/U_0$ for different values of the time of wood drying.

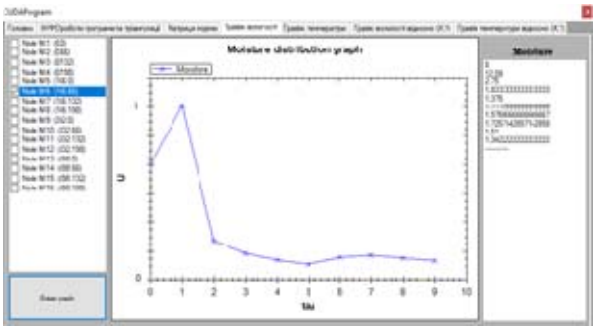


Fig. 4. Change in moisture content for $\tau = 10$ h $(R_1 = R_2; R_3 = 0)$

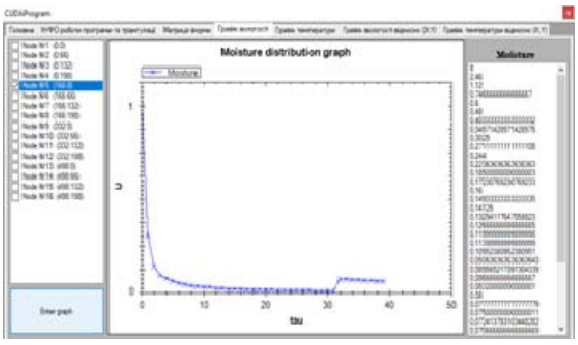


Fig. 5. Change in moisture content for $\tau = 40$ h $(R_1 = R_2; R_3 = 0)$

Figs. 6-7 also shows the dynamics of temperature change with the same parameters.

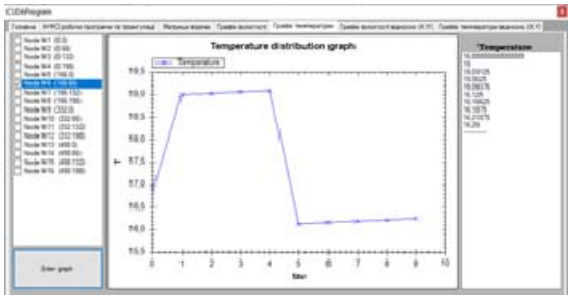


Fig. 6. Change in temperature at the node $\tau = 10$

The indicator values of using parallelization in conjunction with Azure Cloud in modeling the heat-and-mass transfer problem are shown in Fig.8. With an increase in load, the CPU becomes more resource intensive compared to the GPU. But there is not so much difference between the indicator values than expected. However, with an increase in

the density of the triangulation grid, the difference in performance becomes much more noticeable.The cause of this phenomenon is the structure of GPUs.

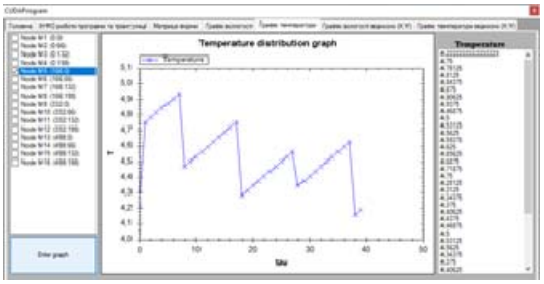


Fig. 7. Change in temperature at the node $\tau = 40$

When performing this task, information processing is performed in all cores of the GPU, regardless of the load of all threads. Accordingly, the GPU handles small calculations more slowly, which in turn causes programmer to use all possible threads on the GPGPU. That is, parallelization in the GPU is better to use for large calculations.

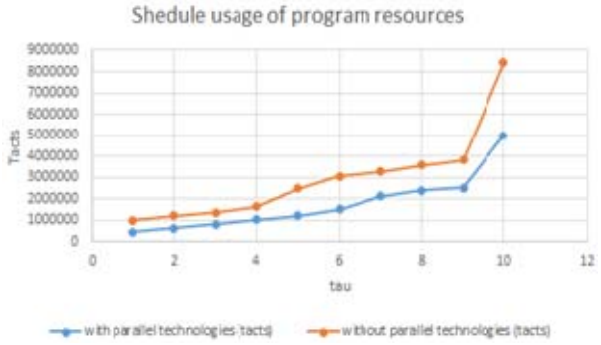


Fig. 8. Parallelization indices, tact / h

But CUDA technology has its own negative aspects. One of them is the design of graphics processors. That is, when performing any task information is being processed in all processor cores. This in turn makes the programmer use all possible flows on the GPGPU, which is not always optimal for the speed of the program [21].

CONCLUSIONS

A mathematical model of heat-and-mass transfer in anisotropic capillary-porous materials is presented as a system of differential equations with initial and boundary conditions of the third kind. To study moisture transfer processes, a software architecture was designed that combines Azure cloud technology, which contains the input and output data of the project, with a user interface. The developed software uses a test of the possibility of using CUDA parallelization technologies on the device where the software application is installed. If this is not possible, according to the software settings, all parallel computing is performed using the Azure service. The matrix operations of the project were implemented using the cuBLAS library, which made it possible to accelerate the obtaining of the result, despite the increase in the dimension of the matrices.

REFERENCES

[1] S.J. Kowalski, J. Banaszak, "Modeling and experimental identification of cracks in porous materials during drying", Drying Technology, vol. 31(12), 2013, pp. 1388-1399.

[2] J. Bodig, B.A. Jayne, "Mechanics of wood and wood composites", Krieger Publishing Company, p. 712, 1993.

[3] Daniel S.H. LO "Finit element mesh generation", CRC Press is an imprint of Taylor & Francis Group, 2015, pp. 239-326.

[4] R.S. Duško, A. D. Dedić, A. S. Mujumdav, and N. Lj Ćuprić, "Two-dimensional mathematical model for simulation of the drying process of thick layers of natural materials in a conveyor-belt dryer", Thermal Science, Vol. 21, no. 3, 2017, pp. 1369-1378.

[5] T. Schnabel, H. Huber, and A. Petutschnigg, "Modelling and simulation of deformation behaviour during drying using a concept of linear difference method", Wood Science and Technology, 2017, DOI 10.1007/s00226-017-0897-6

[6] https://www.nvidia.com.ua/object/cuda-parallel-computing-ru.html

[7] Hybrid DSP. CUDAfy user manual, USA, p. 45, 2015.

[8] John F. Sian, "Wood: influence of moisture on physical properties", Virginia, p. 227, 1995.

[9] I.F. Siau, "Wood influence of moisture on physical properties", USA, Virginia State University, p. 236 , 1995.

[10] A.V. Skvortsov, N.S.Mirza, "Construction algorithms and analysis of triangulation", Tomsk University, p 168, 2006. [in Russian]

[11] Y. Sokolovskyy, I. Boretska, I. Kroshnyy, and B. Gayvas, "Mathematical models and analysis of the heat-mass-transfer in anisotropic materials taking into account the boundaries of phase transition", Proceedings of the 2019 IEEE 15th International Conference on the Experience of Designing and Application of CAD Systems (CADSM), Polyana, 26 February – 2 March 2019, pp.28-33

[12] Y. Sokolovskyy, I. Kroshnyy, and V. Yarkun "Mathematical modeling of visco-elastic-plastic deformation in capillary-porous materials in the drying process", Proceedings of the 2015 IEEE 10th International Scientific and Technical Conference on Computer Sciences and Information Technologies (CSIT), Lviv, 14-17 September 2015, p. 52-56.

[13] Y. Sokolovskyy, M. Levkovych, O. Mokrytska, and V. Atamanvuk, "Mathematical Modeling of Two-Dimensional Deformation-Relaxation Processes in Environments with Fractal Structure", Proceedings of the 2018 IEEE 2nd International Conference on Data Stream Mining and Processing (DSMP), Lviv, 21-25 August 2018, p.375-380.

[14] Y. Sokolovskyy, V. Shymanskyi, and M. Levkovych, "Mathematical Modeling of Non-Isothermal Moisture Transfer and Visco-Elastic Deformation in the Materials with Fractal Structure". Proceedings of the 2016 IEEE 11th International Scientific and Technical Conference on Computer Sciences and Information Technologies (CSIT), Lviv, 6-10 September 2016, p. 91-95.

[15] Y. Sokolovskyy, O. Sinkevych, "Calculation of the drying agent in drying chambers", Proceedings of the 2017 14th International Conference The Experience of Designing and Application of CAD Systems in Microelectronics (CADSM), Lviv, 21-25 February 2017, p.27-31

[16] V. Shymanskyi, Yu. Protsyk, "Simulation of the Heat Conduction Process in the Claydite-Block Construction with Taking Into Account the Fractal Structure of the Material", XIII-TH International Scientific And Technical Conference; Computer Science And Information Technologies; CSIT-2018, 2018, pp.151-154

[17] P. Perré "Coupled heat and mass transfer in wood and wood-based products: macroscopic formulation, upscaling and multiscale modelling", CompWood, 2019, pp. 5-5.

[18] O. Pinchevska, A. Spirochkin, J. Sedliačik, and R. Oliynyk, "Quality assessment of lumber after low temperature drying from the view of stochastic process characteristics", Wood Research, Vol. 61 (6), 2016, pp. 871-883.

[19] Ya. Sokolowskyi, V. Shymanskyi, A. Nechepurenko, S. Xaborskyj, and M. Dmytrus, "Mathematical modeling of rheological behavior and distribution of temperature and moisture fields in the drying of lumber", 13-th International Conference "The Experience of Designing and Application of CAD Systems in Microelectronics", Svalyava, 2015. pp. 214-217.

[20] A.V. Boreskov, "Parallel GPU computing. CUDA Architecture and Software Model", Moscow University Publishing House, p. 336, 2012.

[21] Shane Cook, "CUDA Programming: A Developers Guide to Parallel Computing with GPUs", USA, Morgan Kaufmann Publishers Inc., 2013, p. 591.

IEEE Third International Conference on Data Stream Mining & Processing
August 21-25, 2020, Lviv, Ukraine

Innovative, Scientific and Technical Activities in Ukraine: Modern Trends and Forecasts

Liubov Halkiv
Institute of Economics and Management
Lviv Polytechnic National University
Lviv, Ukraine
ORCID: 0000-0001-5166-8674

Oleh Karyy
Institute of Economics and Management
Lviv Polytechnic National University
Lviv, Ukraine
ORCID: 0000-0002-1305-3043

Ihor Kulyniak
Institute of Economics and Management
Lviv Polytechnic National University
Lviv, Ukraine
ORCID: 0000-0002-8135-4614

Solomiya Ohinok
Institute of Economics and Management
Lviv Polytechnic National University
Lviv, Ukraine
ORCID: 0000-0001-5462-5362

***Abstract*—The research is based on the official materials of Ukraine's state institutions, for whose elaboration content analysis, methods of logic, comparison and analogies are used. The fractal theory is applied to test the hypothesis, in particular, parametric methods of R/S-analysis (the nature of dynamic changes in the indicators of innovative and scientific and technical activity of Ukraine are investigated and trends are identified) and the method of correlation-regression analysis (the nature, density, and direction of change of innovative and scientific and technical activities indicators are evaluated). The research proves public demand, evaluates current trends and makes forecasts of the indicators of innovative and scientific and technical activities of Ukraine.**

***Index Terms*—innovative activity, scientific and technical activity, indicators, trend, modeling, forecasting.**

I. Introduction

In today's conditions of development of knowledge-intensive economy, the fact that the basis of competitiveness of public systems of micro-, meso-, macro- and mega-levels, the drivers of their technological and technological advancement and activators of progressive transformations is formed by innovative and scientific-technical processes. The high dynamism of globalization shifts in the context of Ukraine's integration into the world community requires strengthening, diversification and increase of efficiency of activity of domestic producers and providers of innovations. As the practice of the advanced countries proves, in the limited potential conditions of factors of socio-economic progress is a need to adopt an innovative development paradigm, thus actualizing the problem of finding prudent solutions on the levers of ensuring the trajectory of intellectualization of social productions, based on their technological resources, and their technological development the embodiment of creative development. This requires a theoretical and methodological justification of approaches and the development of conceptual foundations of the practice of evaluating the processes that accompany the paradigm shifts of innovative orientation.

The transition of Ukraine to an innovative model of development requires considerable efforts to accumulate the innovative potential of domestic productive forces and to develop effective mechanisms for its use to maximize returns. Improving the efficiency of innovation and scientific and technological processes is necessary not only for improving the macroeconomic situation but also for strengthening Ukraine's position on international markets, for strengthening social, environmental and technological components of national security. Increasing the level of the domestic economy innovation, first of all, will ensure a proper level of domestic competitiveness enterprises and the economy as a whole, will allow bringing the economic development indicators of the country closer to the average world indicators and will guarantee a painless process of their integration into the European community. One of this task the components is the widespread introduction of innovation in all areas of activity. This will solve the problems related to raising the standard of citizens living, guaranteeing the economic security of the country, solving social and environmental problems, increasing the country's export potential, etc.

Given that the world has stopped on the threshold of the fourth industrial revolution that has already begun, innovative and scientific and technological activity is an important factor in the transition of enterprises to Industry 4.0 [11].

In this context, the activation of innovative and scientific and technical activities of domestic enterprises plays an important role. After all, enterprises are the link where new types of products, new technologies, new organizational and technical solutions are developed and introduced into production, and new production processes are mastered. Unfortunately, the level of innovation activity of domestic enterprises is rather low today, which searches for ways to improve the efficiency of innovation processes and intensify their innovation activity.

According to the estimates of scientists, independent experts, and international organizations, the national economy has sufficient scientific and innovative potential, but the level of realization of this potential does not meet the requirements

978-1-7281-3214-3/20/$31.00 ©2020 IEEE

of the dynamic development of the economy on innovative grounds. The purpose of creating interactive systems is due to the need to find ways to accelerate the materialization of existing developments in innovative solutions. This calls for the formation of a scientific approach to solving the problem of managing innovation processes through the formation of a mechanism for establishing cooperation: from optimizing the choice of the idea of a new product to providing the final consumer with after-sales service [1], [2]. The innovation development of the economic system at the present stage, in addition to stimulating the innovative activity of existing entities, deals with the formation of network structures. The researches [8], [9] address the indicators for networks modeling by developing and testing a series of points indicating how networks affect economic growth.

The results of the analysis of the received array of scientific published works make it possible to state that the topic of innovative and scientific and technical activity is permanently in the field of scientists' view, the interest in this topic does not disappear, and the most recent research published over the last five years focuses on the following four aspects:

1) Policy and legal basis of the country's scientific and technical activity.
2) Characteristic aspects of scientific and technical and innovative activity.
3) Commercialization of the results of scientific and technical and innovative activity.
4) Evaluation of the results of scientific and technical and innovative activity.

Despite a large amount of research on the evaluation and efficiency of innovation processes, there are still several unresolved issues that need to be thoroughly covered in scientific sources, which will help to develop the national innovation economy at all levels of government and help improve the level of state competitiveness, and individual businesses in national and international markets. This is what has made this research relevant.

Based on the above mentioned fact, the authors aim to carry out modeling and forecasting of the indicators of innovative and scientific and technical activity on the example of Ukraine (share of innovatively active enterprises in the total number of enterprises; volume of innovative activity financing; number of new technological processes implemented; number of the names of innovative products introduced; specific weight of the volume of sold innovative products in the total volume of products sold; number of new technologies purchased; number of producers of scientific and technical works; number of specialists performing scientific and technical works; amount of funding for research and development) and is determined by their effect.

II. Data and Methods of Research

The authors propose to use the indicators whose quantitative values are publicly available on the website of the State Statistics Service of Ukraine [10], for the construction of models of the dynamics of innovative and scientific and technical activity, namely:

1) indicators of the innovative activity of Ukraine:

x_{11} – share of innovatively active enterprises in the total number of the surveyed industrial enterprises, %;

x_{12} – total amount of innovative activity financing, thousand UAH;

x_{13} – number of new technological processes implemented at industrial enterprises, processes;

x_{14} – number of the names of introduced innovative types of products at industrial enterprises, units;

x_{15} – specific weight of the volume of realized innovative products in the total volume of industrial products sold, %;

x_{16} – number of new technologies (technical achievements) acquired, units;

2) indicators of the scientific and technical activity of Ukraine:

x_{21} – number of organizations performing scientific and technical works, units;

x_{22} – number of specialists who perform scientific and technical works (at the end of the year, people);

x_{23} – amount of financing of internal expenses for the implementation of research and development, thousand UAH.

Taking into account the fact that the data for 2018 are missing in Ukraine since the periodicity of the state statistical observation "Innovative Activity of the Industrial Enterprise" was changed by the state regulator from "annual" to "once every two years", starting with the report for 2015, the period from 2005 to 2017 was selected for analysis. The numerical values of the indicators until 2015 were collected based on the state statistical observation of Form 1-innovation (annual) "Inspection of Innovative Activity of the Industrial Enterprise", which covered the enterprises of economic activity of sections B, C, D, E by the Classification of Types of Economic Activity (CTEA). The information for 2016 is formed based on the data of the state statistical observation of form INN "Inspection of Innovative Activity of Enterprises for the Period 2014-2016" (by the international metrology) according to the following separate reporting units: enterprises by the types of economic activity of sections B, C, D, by CTEA.

It is known that undetermined phenomena in social systems, to which the indicators of scientific and innovative development of the economy belong, may be characterized by chaotic dynamics. The mathematical theory of fractals is used to investigate the nature of this kind of a dynamic change, identify trends and for further prediction. Different fractal structures in economic systems determine the fractal behavior of the economic indicators of such systems. This made it necessary to investigate the fractal behavior of indicators of innovation and scientific and technological activity of Ukraine with their further prediction.

In the 60s Benoît Mandelbrot was the first one to use the notion of the "fractal" to name objects with complex geometry that could not be characterized by an integral dimension [7]. A major characteristic of fractal objects is the fact that metrically measured properties such as the length or area are dependent

on the measurement scale. The fractal dimension is a measure of object complexity [6].

The basis of this theory is R/S analysis (the method of normalized span) – a set of statistical techniques and methods of the analysis of time series that allow determining some of their important characteristics, such as the availability and depth of long-term memory (non-periodic cycles). This method allows distinguishing a random series from a non-random one. The Hurst exponent (H) and fractal dimension (D) are used to determine the nature of changes in the levels of a series of dynamics and their forecasting in R/S-analysis. The value of the fractal dimension (D) is two units higher than the Hurst exponent H [3]–[5].

Indicator H takes the value from 0 to 1, characterizing the measure of a quantitative assessment of the level of the data "noisiness" presented in the form of a time series. The strength of the trend resistance (persistence) of the behavior of a dynamics series increases with the Hurst exponent approaching 1 (Table I).

TABLE I
CHARACTERISTIC OF THE TIME SERIES BY THE HURST EXPONENT

Hurst exponent (H)	Characteristic of the time series
H = 0	The series is "dead". Dynamic changes are either absent or cyclic or have large amplitude of oscillation.
$0 \leq H < 0.5$	The series is anti-persistent (unstable). The closer H is to zero, the more unstable the dynamics of the indicator is. In the future, the dynamic process will show a trend opposite to that which was characteristic of the previous period.
H = 0.5	The series is random. Events are uncorrelated, "randomly walking". The current indicator's state is in no way related to its future state.
$0.5 < H \leq 1$	The persistent (trend-resistant) series. The future values of series levels depend on the past ones. The trend, as evidenced by the time series, will be prolonged for a certain future period. The series is characterized by the presence of fractal properties.

Source: drawn up by the authors according to [3], [6]

III. RESULTS AND DISCUSSION

The theoretical representation of linear regression equations, the values of the Hurst exponents and fractal dimension are presented in Table II-III.

As it can be seen from the analysis (Table III), only indicator x_{13} ("Introduction of new technological processes at industrial enterprises") is characterized by unstable dynamics and the absence of a clear trend, which allows choosing a model of the random walk "one-dimensional Brownian motion". In this case, it is expedient to use exponential smoothing and moving average for forecasting. All the other indicators are characterized by a steady trend of their change, that is, trend dependence, which will be extended into the future.

TABLE II
EQUATIONS OF LINEAR REGRESSION OF THE INDICATORS OF THE INNOVATIVE AND SCIENTIFIC AND TECHNICAL ACTIVITIES OF UKRAINE

Indicator's name	Linear regression equation
x_{11}	ln (R/S) = 0.782 ln(n) – 0.008
x_{12}	ln (R/S) = 0.740 ln(n) – 0.574
x_{13}	ln (R/S) = 0.334 ln(n) + 0.509
x_{14}	ln (R/S) = 0.696 ln(n) + 0.149
x_{15}	ln (R/S) = 0.751 ln(n) – 0.132
x_{16}	ln (R/S) = 0.763 ln(n) – 0.817
x_{21}	ln (R/S) = 0.875 ln(n) – 0.391
x_{22}	ln (R/S) = 0.911 ln(n) – 0.599
x_{23}	ln (R/S) = 0.793 ln(n) – 0.275

Source: authors' calculations.

TABLE III
EQUATIONS OF THE VALUE OF THE HURST EXPONENT AND FRACTAL DIMENSION OF THE INDICATORS OF THE INNOVATIVE AND SCIENTIFIC AND TECHNICAL ACTIVITIES OF UKRAINE

Indicator's name	Hurst exponent (H)	Fractal dimension (D)	Characteristic of the time series
x_{11}	0.782	1.219	Persistent
x_{12}	0.740	1.259	Persistent
x_{13}	0.334	1.666	Anti-Persistent
x_{14}	0.696	1.304	Persistent
x_{15}	0.751	1.249	Persistent
x_{16}	0.763	1.237	Persistent
x_{21}	0.875	1.125	Persistent
x_{22}	0.911	1.089	Persistent
x_{23}	0.793	1.207	Persistent

Source: authors' calculations.

We use trend models based on the idea of extrapolation to predict the indicators of innovative and scientific and technical activity of Ukraine, which are characterized by stable dynamics. Extrapolation is generally understood to mean the spread of patterns, relationships, and relationships that exist in the study period beyond. Linear trend models can be used to forecast these indicators (Table IV-V).

The exponential smoothing method is a type of mathematical transformation that is used in predicting time series. It got its name because every subsequent iteration takes into account all the previous values of a series, but the degree of incorporation decreases by exponential law. Moving Average (Holt-Winters method) – is an advanced method of exponential smoothing of the time series. This method successfully copes with both medium-term and long-term forecasts, as it can detect micro-trends (short-term trends) at times immediately preceding the forecast and to extrapolate these trends into the future. This method can be used to forecast anti-persistent indicator x_{13} – "number of new technological processes implemented at industrial enterprises." (Table IV-V).

Despite some positive tendencies of indicators change of innovations introduction, forecast values show an insufficient level of innovative processes development in Ukraine and a high probability of similar tendencies continuation in the future. Therefore, it is necessary to change the basic principles of innovation policy in enterprises.

TABLE IV
LINEAR TREND MODELS OF THE INDICATORS OF INNOVATIVE AND SCIENTIFIC AND TECHNICAL ACTIVITIES OF UKRAINE FOR 2005-2017

Indicator's name	Trend line equation	Determination factor (R^2)
x_{11}	x_{11} = 0.5363 t + 11.308	0.76
x_{12}	x_{12} = 610,769 t + 6E+06	0.26
x_{13}	Forecasting by the exponential smoothing method	
x_{14}	x_{14} = 66.203 t + 2,515.4	0.21
x_{15}	x_{15} = -0.5354 t + 7.6288	0.96
x_{16}	x_{16} = 36.484 t + 424.54	0.24
x_{21}	x_{21} = -49.516 t + 1,570.1	0.97
x_{22}	x_{22} = -1,767.9 t + 100,759	0.32
x_{23}	x_{23} = 640,501 t + 5E+06	0.94

Source: authors' calculations.

TABLE V
FORECASTING THE INDICATORS OF INNOVATIVE AND SCIENTIFIC AND TECHNICAL ACTIVITIES OF UKRAINE FOR 2019-2021

Indicator's name	Indicators' predicted values			Expectation of the trend of indicator's change
	2019	2020	2021	
x_{11}	19.4	19.9	20.4	↑
x_{12}	15,161,535	15,772,304	16,383,073	↑
x_{13}	2,283	2,684	2,608	–
x_{14}	3,508	3,575	3,641	↑
x_{15}	0.1	0	0	↓
x_{16}	972	1,008	1,045	↑
x_{21}	827	778	728	↓
x_{22}	74,241	72,473	70,704	↓
x_{23}	14,607,515	15,248,016	15,888,517	↑

IV. CONCLUSION

It should be noted that all the indicators used by the authors as indicators of innovative and scientific and technical activity are logically considered to be indicators that stimulate scientific and technical progress and economic innovation. Nevertheless, three indicators (x_{15} – the specific weight of the volume of realized innovative products in the total volume of industrial products sold, x_{21} – number of organizations performing scientific and technical works, x_{22} – number of specialists performing scientific and technical works) show a downward trend. The differently vectored orientation of dynamic changes (the tendency of a decrease in the share of innovative products in their total volume against the background of an increase in the share of innovatively active enterprises, an increase in expenses for innovative activity, name extension and an increase in the number of innovations) cannot indicate the effective innovative development of Ukraine. Similarly, under conditions of a trendy decline in producers of scientific and technical developments, the outflow of human science capital, risks and threats for the research sphere progress is activated. This situation is partly explained by the existence of a discrepancy between the declared purpose and the actual implementation of the state scientific and technical and innovative policy.

The innovative and scientific and technical activity in Ukraine has not taken the place of the systemic core of social progress. Despite the persistence of upward dynamics of the share of innovatively active enterprises, the volume of scientific and innovative activity financing, the number of new technological processes implemented and new technologies acquired, the time series of the list of names of innovative products is anti-persistent. Instead, these important indicators of scientific and innovation progress, which characterize the level of innovative product sales, the number of organizations and human capital of the research field, show a downward trend. The asymmetries found do not level the role of science and innovation in the national economy's development.

The ways of solving the identified problems lie in the actualization of the content of scientific and technical and innovation policy, updating the existing legislative framework in the direction of the effectiveness of socially acceptable mechanisms of increasing the role of scientific and innovative potential in the national complex, the implementation of the world's best practices.

The results obtained can be the basis for managerial decision-making by state, regional and local authorities regarding the development of innovative activity growth strategies for enterprises, regions, and Ukraine as a whole.

REFERENCES

[1] N. Chukhrai and L. Lisovska, "Formation of innovation's consumer utility," Actual problems of economics, No. 11(149), pp. 27–34, 2013.

[2] N. Chukhray, N. Shakhovska, O. Mrykhina, M. Bublyk, and L. Lisovska, "Methodical approach to assessing the readiness level of technologies for the transfer," Selected papers from the International conference on computer science and information technologies, CSIT 2019, Vol. 1080: Advances in intelligent systems and computing, pp. 259–282, 2020,

[3] E. Feder, Fractals. M.: Mir, 1991.

[4] H.E. Hurst, "Long-term storage capacity of reservoirs," Trans. Am. Soc. Civ. Eng., 116, pp. 770–808, 1951.

[5] H.E. Hurst, R.P. Black, and Y.M. Simaika, Long-Term Storage: An Experimental Study. London: Constable, 1965.

[6] B. Mandelbrot, Fractals, Case and Finance (1959-1997). Izhevsk: RC Dynamics, 2004.

[7] B. Mandelbrot, "How long is the coast of Britain? Statistical self-similarity and fractal dimension," Science, No. 155, pp. 636–638, 1967.

[8] O. Prokopenko, O. Kudrina, and V. Omelyanenko, "ICT Support of Higher Education Institutions Participation in Innovation Networks," Proceedings of the 15th International Conference "ICT in Education, Research and Industrial Applications. Integration, Harmonization and Knowledge Transfer," pp. 482–487, 2019.

[9] O. Prokopenko, V. Omelyanenko, T. Ponomarenko, and O. Olshanska, "Innovation networks effects simulation models," Periodicals of Engineering and Natural Sciences, vol. 7, No. 2, pp. 752–762, August, 2019.

[10] Science, technology and innovation. State Statistics Service of Ukraine. Access of mode: http://www.ukrstat.gov.ua/druk/publicat (accessed: 11/15/2019).

[11] N. Shpak, M. Odrekhivskyi, K. Doroshkevych, and W. Sroka, "Simulation of Innovative Systems under Industry 4.0 Conditions," Soc. Sci., vol. 8, 202, 2019. DOI: 10.3390/socsci8070202.

IEEE Third International Conference on Data Stream Mining & Processing
August 21-25, 2020, Lviv, Ukraine

Analysis of the Distribution of COVID-19 in Italy Using Clustering Algorithms

Anastasiya Doroshenko
ACS Department
Lviv Polytechnic National University,
Lviv, Ukraine
anastasia.doroshenko@gmail.com

***Abstract*—This article describes the using of various clustering methods for in-depth analysis and further researches of the spread of the virus COVID-19 in Italy during February-April 2020.**

Keywords— clustering algorithms, hierarchical clustering, distance map, k-Means, data science, data mining, COVID-19.

I. INTRODUCTION

At the end of December 2019, the first cases of pneumonia of unknown origin in locals associated with the local Huanan animal and seafood market were discovered in the city of Wuhan, Hubei Province of central China. On December 31, 2019, Chinese authorities informed the World Health Organization (WHO) of an outbreak of unknown pneumonia. Since January 22, Wuhan has been quarantined; from January 24 - urban districts adjacent to Wuhan. However, despite such precautions, it was not possible to keep the virus in China. On January 31, two Chinese tourists with coronavirus who arrived in Rome were recorded in Rome. On 30 January, WHO recognized the outbreak of the new coronavirus as a public health emergency of international concern. Despite the fact that only on February 11, 2020 the disease was called the new coronavirus disease (COVID-2019), by this time the virus not only put all Wuhan residents in strict quarantine, but also spread rapidly throughout Italy. By mid-March, Italy came in second in the world in the number of cases of COVID-2019. During the pandemic, the efforts of scientists from various fields are aimed at combating COVID-2019 - medical scientists are looking for drugs, biologists are developing the genome and creating a vaccine, and specialists in the field of Data Science, using known methods and data analysis algorithms, studied the spread of the virus around the world, analyzed the data collected and used them to predict the future development of the pandemic.

II. STATEMENT OF THE PROBLEME

A. *Unsufficient amount of collected historical data*

To successfully use Data Mining methods, it is necessary to have a sufficient amount of collected historical data. Also, they need preprocessing - data cleaning (missing data, Noisy Data), Data Transformation (Normalization, Attribute Selection, Discretization) [2]. However, in a difficult time of pandemic, it was not possible to just wait until a sufficiently large amount of data was accumulated, because every day, the number of sick and dead from COVID-2019 is increasing. Accordingly, it was necessary to use methods that show high accuracy on small amounts of data [6].

B. *Protection of patient personal data in accordance with GDPR*

The second problem in studying the distribution of coronavirus was that information about patients, their age, gender, place of residence, concomitant diseases, are particularly sensitive data and cannot be disclosed according to the General Data Protection Regulation (GDPR). This problem can be solved by applying anonymization methods, etc. However, this is quite risky from a legal and ethical point of view, respectively - it is impossible to get this data quickly, in real-time, and accordingly - to get the most correct result. Accordingly, only those statistics that are public were available for analysis [1]. These are indicators such as the number of new cases per day, deaths per day, the total number of cases, the number of hospitalized people, etc.

Although the first cases were detected in the center of Italy - Rome, the epicenter of the disease was the north of the country - the province of Lombardy. Also, new cases were recorded in almost all provinces. However, in some provinces, the level of the disease remained low, while in others, after a very short time, they became “red zones” - with a high number of sick and dead. To analyze what features are common between the provinces with a high incidence rate, it was decided to cluster the provinces with the two most popular clustering methods - k-Means and hierarchical clustering.

III. COMPARISON OF CLUSTERING METHODS

Depending on the specific task, clustering methods can be used for various purposes:

- To analyze many objects and understand their structure, dividing them into groups of similar objects. As a rule, in this case, it is desirable to choose a small number of clusters.
- Reduce the amount of data stored in the case of an extra-large sample Xℓ, leaving one of the most typical representatives from each cluster. In this case, it is important not the number of clusters, but the provision of a high degree of similarity of objects within each cluster.
- Select atypical objects that do not fit into any of the clusters. This task is called a single-class classification, the discovery of atypicality or novelty (novelty detection). As a rule, in this case, individual objects that do not fit into any of the clusters are of the greatest interest.

978-1-7281-3214-3/20/$31.00 ©2020 IEEE

For all these cases, hierarchical clustering can be applied - a taxonomy, when large clusters are split up into smaller ones, which in turn are split up even smaller, etc. The result of taxonomy is not only the division of many objects into clusters but also a tree-like hierarchical structure. Instead of a cluster number, an object is characterized by listing all the clusters to which it belongs, from large to small.

We will compare the effectiveness of the algorithms chosen for clustering on a sample of 1113 entries describing the epidemiological situation in Italy from February 20, 2020, to April 16, 2020, according to 17 parameters: Date, Country, Region Code, Region Name, Latitude, Longitude, Hospitalized Patients, Intensive Care Patients, Total Hospitalized Patients, Home Confinement, Current Positive Cases, New Positive Cases, Recovered, Deaths, Total Positive Cases, Tests Performed.

Data for this research was collected by Sito del Dipartimento della Protezione Civile - Emergenza Coronavirus: la risposta nazionale and uploaded into github repositories https://github.com/pcm-dpc/COVID-19 [2].

A. k-Means

The k-Means clustering algorithm is a non-hierarchical, iterative clustering method. He gained great popularity due to its simplicity, clarity of implementation and fairly high quality of work.

The main idea of the k-means algorithm is that the data is randomly divided into clusters, after which the center of mass for each cluster obtained in the previous step is iteratively recalculated, then the vectors are again divided into clusters according to which of the new centers is closer in selected metric.

The purpose of the algorithm is to separate n observations $X = \{x_1, x_2, \dots, x_n\}, x_i \in R^d, i = 1, .., n$ to k clusters $K_1, K_2, \dots K_k$, $k \in N, k \leq n$ so that each observation belongs to exactly one cluster $K_1 \cap K_2 \neq \emptyset \bigcup_{i=1}^{k} K_i = X$, located at the smallest distance from the observation $\arg min_K \sum_{i=1}^{k} \sum_{x \in K_i} \rho(x, \mu_i)^2$,

$\mu_i, i = 1, \dots, k -$ are the centers of the clusters,

$\rho(x, \mu_i)$ – is the distance function between x and μ_i.

It should be noted that the algorithm does not apply to data for which the concept of "average" is not defined, for example, categorical data.

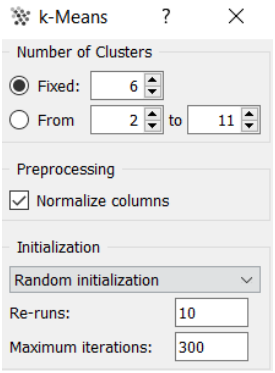


Fig. 1. Parametrs of k-Means algoritm

The advantages of the algorithm are:

- Relatively high efficiency with ease of implementation;
- High-quality clustering;
- Possibility of parallelization;
- The existence of many modifications.

The disadvantages of the algorithm are:

- The number of clusters is a parameter of the algorithm
- Sensitivity to initial conditions - initialization of cluster centers significantly affects the result of clustering.
- Sensitivity to emissions and noise - Emissions that are far from the centers of these clusters are still taken into account when calculating their centers.
- The possibility of convergence to a local optimum - an iterative approach does not guarantee convergence to an optimal solution.

For experiments, the Orange software product and installed additional widgets were used [5].

Pre-processed data is processed using the k-means algorithm with the parameters shown in Fig. 1

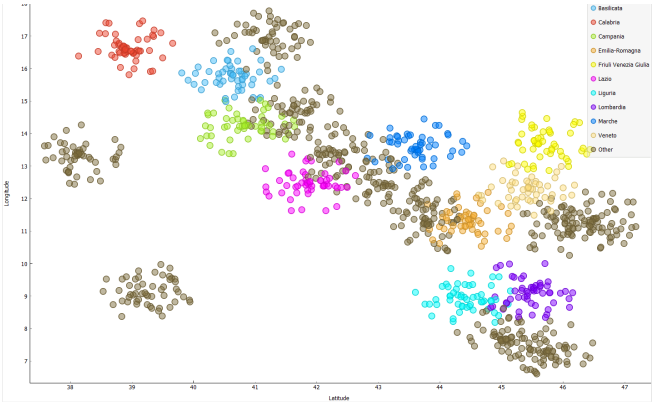


Fig. 2. The results of clustering of Italy`s regions in two-dimensional space (Latitude, Longitude).

Analyzing the obtained results, it should be noted that we obtained the best clustering results for the coordinate space (Latitude, Longitude) (fig.2).

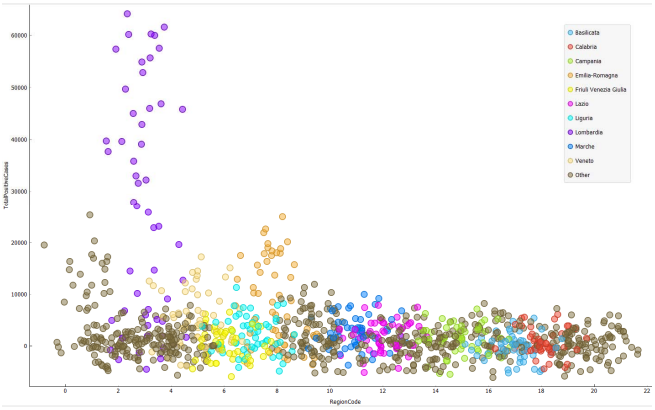


Fig. 3. The results of clustering of Italy`s regions in two-dimensional space (Region Code, Total Positive Cases).

For all other dependencies that interested us - such as the classification according to the number of sick, dead, etc., by region, only 1 cluster (Lombardy) could be distinguished, the remaining clusters overlapped and were inseparable (fig.3).

B. Hierarchical clustering

Hierarchical clustering algorithms build not one partition of the sample into disjoint classes, but a system of nested partitions. The result of taxonomy is usually represented as a taxonomic tree - a dendrogram. Among hierarchical clustering algorithms, two main types are distinguished. Divisible or top-down algorithms break down a sample into increasingly smaller clusters. Agglomerative or ascending algorithms are more common in which objects are combined into more and more large clusters.

First, each object is considered a separate cluster. For single-element clusters, the distance function $R(\{x\},\{x'\}) = \rho(x,x')$ is naturally determined.

Then the merge process starts. At each iteration, instead of the pair of the closest clusters U and V, a new cluster $W = U \cup V$. is formed. The distance from a new cluster W to any other cluster C is calculated from the distances $R(U,V)$, $R(U,C)$ and $R(V,C)$,, which by this moment should already be known: $R(U \cup V,C) = \alpha U R(U,C) + \alpha V \cdot R(V,C) + \beta R(U,V) + \gamma R(U,C) - R(V,C)$. αU, αV, β, γ - numerical parameters.

This universal equation summarizes almost all reasonable ways to determine the distance between clusters. It was proposed by Lance and Williams in 1967 [4].

In practice, the following methods are used to calculate the distances R (W, C) between the clusters W and C. For each of them, the correspondence to the Lance-Williams formula for certain combinations of parameters is proved:

- single linkage clustering or nearest neighbor distance

$$R^n(W,C) = \min_{w \in W, c \in C} \rho(w,c) \quad a_U = a_V = \frac{1}{2}, \beta = 0, \gamma = -\frac{1}{2}$$

- complete linkage clustering or Furthest neighbor distance

$$R^F(W,C) = \max_{w \in W, c \in C} \rho(w,c) \quad a_U = a_V = \frac{1}{2}, \beta = 0, \gamma = \frac{1}{2}$$

- centroid linkage clustering

$$R^C(W,C) = \frac{1}{|W||C|} \sum_{w \in W} \sum_{c \in C} \rho(w,c)$$

$$a_U = \frac{|U|}{|W|} a_V = \frac{|V|}{|W|}, \beta = \gamma = 0$$

The Orange Software also implements other algorithms for calculating the distances between clusters.

Distance Metric:

- Euclidean distance is a geometric distance in multidimensional space.
- Manhattan. According to this metric, the distance between two points is equal to the sum of the modules of the difference of their coordinates.

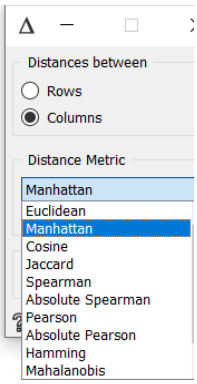


Fig. 4. The algorithms for calculating the distances between clusters in Orange.

- Cosine the similarity factor of two non-zero vectors in the pre-Hilbert space, which is calculated as the cosine of the angle between them.
- Jaccard. It's a binary measure of similarity. This method used to evaluate the similarity of finite sets, in computer science, to search for similar documents, plagiarism, etc.
- Spearman. According to this metric, the distance is linear correlation between the rank of the values, remapped as a distance in a [0, 1] interval)
- Pearson According to this metric, the distance is linear correlation between the values, remapped as a distance in a [0, 1] interval)
- Hamming. It's a binary measure of similarity. The number of features at which the corresponding values are different. [5].

The widget supports four ways of measuring distances between clusters:

- Single linkage computes the distance between the closest elements of the two clusters
- Average linkage computes the average distance between elements of the two clusters
- Weighted linkage uses the WPGMA method
- Complete linkage computes the distance between the clusters' most distant elements

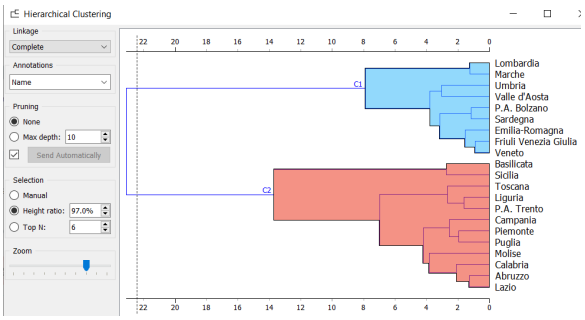


Fig. 5. Hierarchical clustering diagram with a height ratio 97 %

Huge dendrograms can be pruned in the Pruning box by selecting the maximum depth of the dendrogram. This only affects the display, not the actual clustering.

The widget offers three different selection methods:

- Manual
- Height ratio
- Top N

Applying hierarchical clustering for our test set, we obtain this results (fig.5, fig.6).

Comparing the selected clustering methods, we can note that both methods did a good job of dividing regions into clusters. However, in the k-Means algorithm, the resulting clusters describe a well-known division of regions according to their geographical location. For other criteria - similarity in the number of sick, hospitalized, dead - the accuracy of clustering is insufficient.

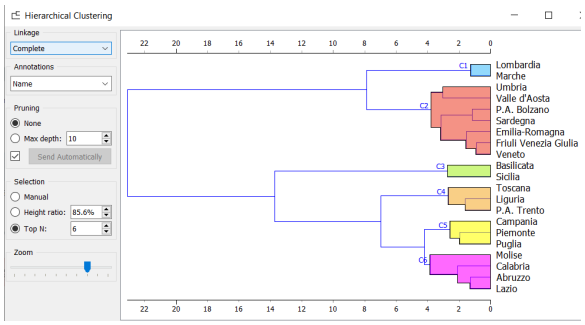


Fig. 6. Hierarchical clustering diagram for 6 clusters

At the same time, the use of hierarchical classification showed more interesting results. Given the height ratio of 97%, the two clusters formed showed the division of regions into those with high incidence (blue cluster, C1) and those with low incidence (red cluster, C2) (fig.5).

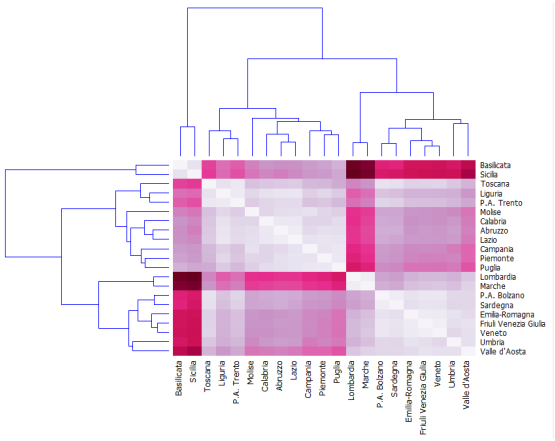


Fig. 7. Distance Map between clusters for Hierarchical clustering

Given a specified criterion — the obligatory selection of 6 clusters — we obtained clusters whose elements at first glance are not so unambiguously interconnected - for example, the number of cases in one cluster can be of the same order, but not the closest from all regions (fig.6).

CONCLUTIONS

After a detailed analysis of the resulting clusters and additional information about the industry of the Italian regions (fig.8), it was found that the clusters obtained combine regions with a similar level of industrial development and similar industries.

Thus, it became clear that the number of sick, hospitalized and deceased citizens of Italy depends on which of the industries in it is developed. Thus, the best epidemiological situation is observed in regions with poorly developed economies or in those where the key areas of activity are: agricultural, construction and trade. The worst is in regions with heavy industry and other industries.

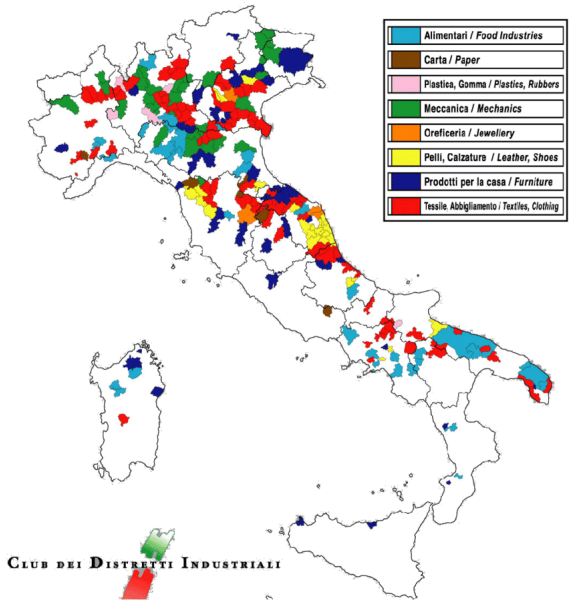


Fig. 8. Regional specialization in Italy in the macroeconomic sector

The subject of further research may be a more detailed analysis of the revealed dependence, however, it is obvious that the task of identifying new, hidden knowledge from the available data is accomplished through the use of the hierarchical classification method.

REFERENCES

[1] Sito del Dipartimento della Protezione Civile - Emergenza Coronavirus: la risposta nazionale and uploaded into github repositories https://github.com/pcm-dpc/COVID-19.

[2] O. Mishchuk, R. Tkachenko and I. Izonin, "Missing Data Imputation through SGTM Neural-like Structure for Environmental Monitoring Tasks" In: Hu, Z.B., Petoukhov, S., (eds) Advances in Computer Science for Engineering and Education II. ICCSEEA2019. Advances in Intelligent Systems and Computing. Springer, Cham, pp. 142-151, 2019.

[3] L. Rokach and O. Maimon. The data mining and knowledge discovery handbook, Springer US, 2005.

[4] G. N. Lance and W. T. Willams, "A general theory of classification sorting strategies. 1. hierarchical systems" Comp. J, no. 9, 1967, pp. 373–380.

[5] Hierarchical Clustering, Orange Widgets Catalog, Documentation https://orange.biolab.si/widgets/unsupervised/hierarchicalclustering

[6] A. Doroshenko, "Application of global optimization methods to increase the accuracy of classification in the data mining tasks", in: Luengo D., Subbotin S. (Eds.): Computer Modeling and Intelligent Systems. Proc. 2-nd Int. Conf. CMIS-2019, vol-2353: Main Conference Zaporizhzhia, Ukraine, April 15-19, 2019., pp.98-109, 2019.

[7] A. Markusen,"Sticky Places in Slippery Space: A Typology of Industrial Districts," Economic Geography, Taylor & Francis Journals, vol. 72(3), pages 293-313, July 1996.

B. Barone and C. Brasili, "La specializzazione e l'internazionalizzazione dell'industria manifatturiera delle regioni italiane". In book: Rapporto ICE 2010-2011 L'Italia nell'economia internazionale. Chapter: 7 - Il territorio Publisher: Istitito Commercio Estero (ICE), Ministero dello Sviluppo Economico, 2011.

IEEE Third International Conference on Data Stream Mining & Processing
August 21-25, 2020, Lviv, Ukraine

Modeling and Forecasting Financial Heteroscedastic Processes

Vira Huskova
Department of Institute of Applied Systems Analysis,
Igor Sikorsky Kyiv Polytechnic Institute
Kyiv, Ukraine
guskovavera2009@gmail.com

Petro Bidyuk
Department of Institute of Applied Systems Analysis,
Igor Sikorsky Kyiv Polytechnic Institute
Kyiv, Ukraine
pbidyuke@gmail.com

***Abstract* – Context: Nowadays in the field of economics, ecology, demographics, we observed nonlinear non-stationary processes. To evaluate the models and forecasts of the processes under consideration, it is necessary to develop special methods. Modeling methodologies include potential uncertainties that arise during data filtering and processing, as well as model structures and parameter estimates.**

Objective: To develop a modified methodology for constructing mathematical models of nonlinear non-stationary processes that will help to achieve high-quality forecasts. The analyzed heteroskedastic processes create a wide class of nonlinear non-stationary processes and are considered in many areas of research.

Methods: To achieve the main goal of the work, a systematic approach is used to build models and forecasts. Used modified methodologies for modeling non-linear non-stationary processes and methods for identifying and accounting for all uncertainties. To account for uncertainties, used methods to refine the order of the model with an adaptive approach to modeling and automatically search for the most optimal structure using complex statistical criteria. A description of non-linearities is also performed using alternative analytical forms with the ability to perform a subsequent assessment of the predictions received.

Results: The work proposed modified modeling methods, built new models and forecasts. As a result of experiments, it was an achievement that nonlinear models and forecasts with using filtering (exponential, elliptical, Kalman filtering) provide a possibility for computing high-quality results and hight quality indicators for the process under study and their variance.

Conclusions: The application of the proposed modeling methodology allows you to simulate the structure and parameters of the model, which are built on the basis of statistical data. The resulting models show high accuracy and good quality for short-term forecasts.

Keywords – nonlinear non-stationary processes, mathematical model, high-quality forecasting, data filtering, short-term forecasts.

I. INTRODUCTION

This study is focused on development of combined models for forecasting nonlinear non-stationary heteroscedastic processes in finances, economy and other areas. It can be easily shown that most of the modern processes under study are nonlinear and non-stationary. The problem of forecasting dynamics of these processes requires from researchers more and more efforts directed to reaching high quality of respective forecast estimates because most of them belong to the class of nonlinear and non-stationary due to availability of multiple random factors influencing their evolution in time. In financial area we can observe influence of various random shocks happening inside of all countries. Moreover, the influence can be caused by outside shocks in the form of general economic crisis, fast variability of prices on energy resources etc. When analyzing such processes, the main problem is to construct adequate mathematical model capable to generate short-term forecasts of volatility and variance. Here special model structures are required that include dynamics of process variance as the main variable [1 – 4] .

Forecasting of processes dynamics could be reached with market available data processing instruments. However, very often such systems are very costly, inconvenient for use and poorly accessible. The systems usually require long special training for solving specific problems and need highly developed modern computers for implementation. All these specific features of the systems available create unfavorable conditions for their wide usage except for banks and large enterprises. In this work practical possibilities are proposed for constructing data processing procedures based on modern principles of systemic approach [5,6].

II. PROBLEM STATEMENT

For the main process and conditional variance constructed generalized methodology for modeling and forecasting nonlinear non-stationary heteroscedastic processes. Also, an important part is the development of efficient data processing and filtering scheme based on systemic principles and performing a comparative analysis of the results achieved on the basis of various models constructed with the computational experiments.

Modern statistical data characterizing evolution of nonlinear non-stationary processes with probabilistic distribution $\{y(k)\} \sim Dist(\mu(k), \sigma_y^2(k))$ $k=1,\ldots,N$ $\mu(k) \neq const$, where $\mu(k)$ is the varying meantime; $\sigma_y^2(k) \neq const$ is the variance that is changed on the time interval of the process of studying. The parameters of statistical data are subject to the following: $\mu(k) < \infty$ $0 < \sigma_y^2(k) < \infty$; on the interval of studying. For this process, it is necessary to develop mathematical models with the general structure: $y(k) = F[y(k), x(k), \theta, \varepsilon(k)]$, where $x(k)$ is possible independent variable; θ is vector of model parameters; $\varepsilon(k)$ is a stochastic process caused by external disturbances and measurement errors. The received model selection criteria are based upon determination co-

efficient, R^2, Durbin–Watson statistic (DW), Mean absolute percentage error (MAPE), Theil coefficient (U), and alternative combined criterion: $J=\left|1-R^2\right|+\left|2-DW\right|+U\to\min_{\hat{\theta}}$. Received models will be constructed according to the proposed methodology.

III. Methods and Approaches

An approach to modeling heteroscedastic processes consists of the following specific steps: (1) application of exponential smoothing algorithms, application of Kalman data processing or elliptical filtering procedures to available data to perform necessary smoothing so that to achieve preliminary data processing necessary for the model creating procedures; (2) if it is necessary to perform heteroscedastic processes analysis the following actions are required: – creating the regression model that describes correctly the initial data itself; – building adequate model describing conditional variance behavior; – computing short- and middle-term forecasts using the models estimated [8, 9].

The basic steps of the methodology that is proposed for modeling the processes (nonlinear and non-stationary) under consideration are given below.

1. Preliminary data processing before model creating with making use of the following basic operations: missing measurements imputation, normalizing the measurements when necessary, optimal and/or elliptical filtering, appropriate processing of extreme values etc.
2. Detecting and suppressing of influence possible uncertainties (negatively influencing quality of models) that data may contain using appropriate algorithms for data pre-processing.
3. Model structure estimation by making use of the correlation and probabilistic (for example, mutual information) measurement processing techniques that provide a possibility for estimating components of a process model under consideration.
4. Constructing nonlinear components of the model using available possibilities for describing nonlinear process behavior. Linear part of a model is fitted on the basis of characteristics of nonlinear model residuals.
5. Model coefficients (parameters) estimation using the algorithms providing unbiased and consistent estimates of parameters on the basis of specific distributions of variables and their parameters.
6. Determining statistics (quality parameters) that characterize candidate models adequacy and selecting the most suitable (adequate) one.

The model constructing for linear and nonlinear time series data is also based upon proposed representation of model structure [10]:

$$S = \left\{ d,\ mo,\ n,\ nl,\ l,\ sd,\ rs \right\}, \qquad (1)$$

where *d* is dimensionality of a model being built; *mo* is order of the model; *n* is a number of exogenous variables; *nl* represents nonlinearity and its possible type; *l* is input lag (delay time); *sd* is stochastic system disturbance and its probability distribution; *rs* characterizes possible restrictions and limitations imposed on variables and parameters.

The components of a model structure are determined by making use of appropriate statistical tests, correlation estimation methods and expert estimates. The nonlinear models that were used successfully for forecasting usually resulted from the former research of econometric processes. For example, nonlinear regression shown below was used to describe formally Gross National Product and tax capital flows [10].

$$y(l) = a_0 + a_1 y(l-1) + b_{12}\exp(z(l)) + \\ + a_2 y(l) z(l) + \varepsilon_1(l); \qquad (2)$$

$$z(l) = c_0 + c_1 z(l-1) + b_{21}\exp(y(l)) + \\ + c_2 y(l) z(l) + \varepsilon_2(l). \qquad (3)$$

A convenient for practice model structure can also be presented in the form of a known bilinear model:

$$z(k) = a_0 + \sum_{i=1}^{p} a_i z(k-i) + \sum_{j=1}^{q} b_j w(k-j) + \\ + \sum_{i=1}^{m}\sum_{j=1}^{s} c_{i,j} z(k-i) w(k-j) + \varepsilon(k). \qquad (4)$$

Here p, q, m and S are the numbers that characterize the model order (autoregression and moving average) [10, 11].

Modeling nonlinear processes rather often is based upon appropriate combination of linear and nonlinear terms of the kind:

$$y(k) = \beta^T \mathbf{z}(k) + \sum_{i=1}^{p} \alpha_i \varphi_i(\theta_i^T \mathbf{z}(k)) + \varepsilon(k), \qquad (5)$$

Here $\mathbf{z}(k)$ is time delayed vector of dependent variable $y(k)$ values; the same touches upon former and current values of exogenous independent vector variables $\mathbf{x}(k)$ that are used with actual time delays. The functions $\varphi_i(x)$ represent a set of linear (and possibly nonlinear) functions that may include the following function types: power function $\varphi_i(x) \equiv x^i$; trigonometric periodic functions $\varphi_i(x) = \sin x$, or $\varphi_i(x) = \cos x$ etc. Also, to this equation can be added quadratic form of the following type $\mathbf{z}^T(k)\,\mathbf{A}\mathbf{z}(k)$; $\varphi_i(x) = \varphi(x),\ \forall i$, where $\varphi(x)$ is so called link function, for example, adequate probability density or nonlinear logistic function.

Regression models (autoregression or autoregression models with moving average) are used for generating forecast for further use by exponential smoothing, optimal Kalman or elliptical filtering. With this approach we can combine successfully widely known classic statistical approach with the modern intellectual data processing techniques. Usually the best result of process development forecasting is achieved in the cases when the variances of actual forecasting errors for the selected separate forecasting techniques do not exhibit high differences [12, 13].

IV. EXPERIMENTS

The data analysis included a few computational experiments with positive results. The first experiment was devoted to the time series analysis using data on prices for precious metals in the period 2014-2015 years. In table 1 given the statistical characteristics that show constructed models and forecasts quality without filtering. Table 2, 3 and 4 indicate results with the application of filtering: exponential smoothing, optimal Kalman filter and elliptical filtering.

In this work considered the case for preliminary data processing with filtering and without it. The main purpose of using a filtering approach - performing data smoothing (suppressing noisy high-frequency components) and their preparation for modeling.

TABLE I. TYPES OF MODELS AND QUALITY RESULTS OF FORECASTS WITHOUT FILTER

Model type	Results of model quality			Forecast quality			
	R^2	$\sum e^2(k)$	DW	MSE	MAE	MAPE	Theil
AR(1)	0,99	25644.6	2.15	49.8	41.3	8.37	0.046
AR(1, 4)	0,99	25588.1	2.18	49.1	40.3	8.12	0.046
AR(1) + 1st order	0,99	25391.3	2.13	34.3	25.1	4.55	0.032
AR(1) + 4th order	0,99	25173.7	2.12	25.9	17.6	3.19	0.024

The Theil coefficient and MAPE for AR(1) + 4th order trend shows the suitability of using this model for short-term forecasting.

TABLE II. TYPES OF MODELS AND QUALITY RESULTS OF FORECASTS WITH APPLICATION OF EXPONENTIAL SMOOTHING

Model type	Results of model quality			Forecast quality			
	R^2	$\sum e^2(k)$	DW	MSE	MAE	MAPE	Theil
AR(1)	0,99	23355,5	2,10	44,8	39,6	7,44	0,036
AR(1, 4)	0,99	24132,1	2,08	46,4	38,3	6,78	0,034
AR(1) + 1st order	0,99	23861,6	2,07	31,0	22,0	3,12	0,027
AR(1) + 4th order	0,99	21887,5	2,05	21,2	13,1	3,15	0,017

The MAPE for AR(1) + 1st order trend and the Theil coefficient for AR(1) + 4th order trend shows the suitability of using this model for short-term forecasting. In this case, exponential smoothing shows a positive effect. It can be noted on statistical indicators of quality.

TABLE III. MODELS AND FORECASTS QUALITY WITH APPLICATION OF KALMAN FILTER

Model type	Results of model quality			Forecast quality			
	R^2	$\sum e^2(k)$	DW	MSE	MAE	MAPE	Theil
AR(1)	0,99	24335,1	2,13	50,0	40,1	7,98	0,040
AR(1, 4)	0,99	24453,1	2,12	47,1	39,5	7,68	0,042
AR(1) + 1st order	0,99	25061,1	2,10	34,0	23,2	3,55	0,033
AR(1) + 4th order	0,99	23881,2	2,07	24,1	16,58	3,05	0,022

The MAPE and the Theil coefficient for AR(1) + 4th order trend display the suitability of using this model for short-term forecasting. It supported by statistical indicators of quality. But in this case, Kalman filter shows similar results with exponential smoothing.

TABLE IV. MODELS AND FORECASTS QUALITY WITH APPLICATION OF ELLIPTICAL FILTERING

Model type	Results of model quality			Forecast quality			
	R^2	$\sum e^2(k)$	DW	MSE	MAE	MAPE	Theil
AR(1)	0,99	24235,0	2,14	51,0	42,5	7,67	0,038
AR(1, 4)	0,99	24467,1	2,15	48,2	40,1	7,56	0,041
AR(1) + 1st order	0,99	25045,5	2,07	32,1	25,6	3,67	0,032
AR(1) + 4th order	0,99	23869,9	2,06	23,1	18,7	3,25	0,025

In the first experiment revealed that the best model turned out to be AR(1) + trend of 4th order with the application of exponential smoothing and Kalman filter. Due to this, it is possible to forecast for one step ahead with mean absolute percentage error of about 3.05% (for Kalman filter) and 3,15% (for exponential smoothing), and Theil coefficient is 0.022 (for Kalman filter) and 0,017 (for exponential smoothing).

In the second experiment, a statistical analysis of the selected time series was performed. It proved that the heteroscedastic processes with time-varying conditional variance generated by different metal prices data. Table 5 shows appropriate forecasting models for performing trading operations as the variance is one of the key parameters. This problem was solved using GARCH models with a description of a high order process (processes trend which is rather sophisticated).

TABLE V. RESULTS OF MODELING AND FORECASTING CONDITIONAL VARIANCE

Model type	Results of model quality			Forecast quality			
	R^2	$\sum e^2(k)$	DW	MSE	MAE	MAPE	Theil
GARCH (1,7)	0,99	153646	0.11	973.5	–	510.6	0.113
GARCH (1,10)	0,99	102122	0.16	461.7	–	207.3	0.081
GARCH (1,15)	0,99	80378	0.32	409.3	–	117.6	0.058
EGARCH (1, 7)	0,99	45023	0.43	64.8	–	7.15	0.023

Based on the results the best-received model was exponential EGARCH (1, 7). The MAPE result was 7.15%,

which is a good result regarding forecasting conditional variance.

In the third experiment presented a combination (linear model and nonlinear model) for forecasting the financial process (table 6). This approach includes results with Kalman filtering, nonlinear logit models and linear regression model.

TABLE VI. FORECASTING RESULTS DIRECTION FOR MINIMUM PRICE OF EVOLUTION

Model type	Back testing Results					
	95%		97%		99%	
	number of excess	correct forecasts	number of excess	correct forecasts	number of excess	correct forecasts
GARCH (1,7)	35	91,86 %	36	90,65 %	29	92,08%
GARCH (1,10)	48	85,13 %	45	87,07 %	40	88,96%
GARCH (1,15)	56	77,25 %	51	78,98 %	47	82,45%
EGARCH (1, 7)	62	79,01 %	64	78,76 %	58	78,92%

The best results were achieved using the additional forecast with GARCH(1,7) model and indicators were equal to 92,089%. Also, an acceptable result was achieved with GARCH(1,10) model, where the percentage of correct forecasts = 88,96%. The proposed approach can be used on the subsequent steps of decision making since the statistical quality characteristics show the high quality of the forecasts.

V. CONCLUSIONS

The systemic approach was proposed for constructing decision support system for mathematical modeling and forecasting modern financial processes. For obtaining high-quality final results with adequate model the methodology was proposed to describe formally linear and nonlinear parts of a process under study.

The scientific novelty of the study is in a new methodology approach of constructing mathematical models for nonlinear non-stationary processes. The methodology based and implemented in the form of appropriately constructed decision support systems. It is was provided the experiments with data filtering, which provided and confirmed the adequacy of the constructed model and high quality of the forecast.

This methodology is applicable not only for mathematical modeling and forecasting nonlinear non-stationary processes in economy and finances using statistical data but also for constructing models in demography, healthy, ecology, and in other human activity spheres where statistical/experimental data is available in the time series form.

The practical significance of the results obtained in the possibility of achieving a high quality of intermediate and final results through tracking procedures at all stages of data processing and model development with appropriate sets of quality statistics.

Further research prospects will be directed on the improvement and evolution of current methodology. The refinement will concern the model's structure, parameter estimation, the creation of new model structures, and active usage of different approaches to modeling and forecasting in the frames of a similar system.

REFERENCES

[1] O. Trofymchuk et al., "Probabilistic and statistical uncertainty Decision Support Systems", Visnyk Lviv Polytech. Natl Univ, 826, 2015, pp. 237–248.

[2] Y. Affet-Sahalia, "Transition Densities for Interest Rate and Other Nonlinear Diffusions", The Journal of Finance, vol. 54, No. 4, 1999, pp. 1361–1395.

[3] S. Allen, Financial risk management: A practitioner's guide to managing market and credit risk, Hoboken, N.J.: John Wiley & Sons, Inc., 2003.

[4] P.I. Bidyuk and E.O Matros, Models of risk assessment of crediting of physical, Cybernetics and computer engineering, №153, 2007 pp. 87–95.

[5] E. Xekalaki and S. Degiannakis, ARCH models for financial applications, Chichester: John Wiley & Sons Ltd., 2010. DOI: 10.1002/9780470688014

[6] R.F. Engle, "Autoregressive conditional heteroskedasticity with estimates of the variance of U.K. inflation", Econometrica, 50, 1982, pp. 987-1008.

[7] F.X. Diebold and F.X. Mariano "Comparing predicting accuracy", Journal of Business and Economic Statistics, vol. 13, 1995, pp. 253 – 263. DOI: 10.1080/07350015.1995.10524599

[8] T.S. Breusch and A.R. Pagan, "A simple test for heteroskedastic city and random coefficient variation", Econometrica, 46, 1978, pp. 1287-1294.

[9] P.I. Bidyuk, O.M. Trofymchuk and O.A. Kozhukhivska, "Probabilistic and statistical uncertainty processing using decision support systems", Visnyk of Lviv Polytechnic National University, No. 826, 2015, pp. 237-248.

[10] J.H. Herbert and P.S. Kott, "Robust variance estimation in linear regression", Journal of Applied Statistics 15:3, 1988, pp. 341-345.

[11] P.I. Bidyuk, S.O. Dovgij and O.M. Trofymchuk, DSS based on statistical and probabilistic procedures, Kyiv: Logos, 2014.

[12] S.E. Said and D.A. Dickey "Testing for unit roots in autoregressive-moving average models of unknown order," Biometrika, 71, 1984, pp. 599-608

[13] E.A. Bashkow, O.A. Dmitrieva, N.H. Huskova, M. Michalska and Y. Amirgaliyev "Parallel multiple blocked methods of Bickart type", Proc. SPIE 11176, Photonics Applications in Astronomy, Communications, Industry, and High-Energy Physics Experiments, 111761X (6 November 2019). https://doi.org/10.1117/12.2536798

IEEE Third International Conference on Data Stream Mining & Processing
August 21-25, 2020, Lviv, Ukraine

Determination of Heart Disease Based on Analysis of Patient Statistics using the Fuzzy C-means Clustering Algorithm

Ievgen Meniailov
Mathematical Modeling and Artificial Intelligence department
National Aerospace University "Kharkiv Aviation Institute"
Kharkiv, Ukraine
evgenii.menyailov@gmail.com

Dmytro Chumachenko
Mathematical Modeling and Artificial Intelligence department
National Aerospace University "Kharkiv Aviation Institute"
Kharkiv, Ukraine
dichumachenko@gmail.com

Ksenia Bazilevych
Mathematical Modeling and Artificial Intelligence department
National Aerospace University "Kharkiv Aviation Institute"
Kharkiv, Ukraine
ksenia.bazilevich@gmail.com

***Abstract*— The statistical characteristics of patients have a great influence on determining the likelihood of heart disease. To determine the disease in medical diagnostics, statistical methods are most often used Data Mining, which with large amounts of information and complex relationships can give more accurate estimates, especially with a large number of similar characteristics. In this paper, we consider the task of clustering data to determine the likelihood of heart disease for patients with similar characteristics. To solve the problem, the Fuzzy C-means method with normalization was used, and Python and the Pandas library were chosen to implement the algorithm. The developed software module allows visualization of the algorithm. This system also supports downloading and uploading data with the service according to safety requirements of data.**

Keywords*— *Fuzzy C-means; data mining; cluster analysis; differential diagnosis

I. INTRODUCTION

Cardiovascular disease (CVD) is the leading cause of death worldwide: for no other reason, as many people die every year as from CVD.

According to the World Health Organization [1], an average of 17.9 million people die because of CVD each year, accounting for 31% of all deaths in the world. 85% of these deaths occurred as a result of a heart attack and stroke.

Over 75% of CVD deaths occur in low- and middle-income countries [2].

From the 17 million deaths from noncommunicable diseases under the age of 70, 82% are in low- and middle-income countries, and 37% are caused by CVD.

Most cardiovascular diseases can be prevented by addressing risk factors such as tobacco use, unhealthy diets and obesity, physical inactivity, and harmful use of alcohol through strategies that cover the entire population.

People with CVD or at high risk for these diseases (due to one or more risk factors, such as high blood pressure, diabetes, hyperlipidemia, or an already developed disease) need early detection and assistance through counseling and, if necessary, taking medication [3].

The processing of medical data is always devoted to specific goals, one of which, perhaps the most important, is the classification of the object or diagnosis [4]. All further actions and the quality of life of the patient depend on the results of the study. The diagnosis has long been considered in some sense an art based on the experience and intuition of a doctor.

With the mathematization of medicine, the diagnosis can be formulated as a mathematical problem, therefore, automated [5]. Since making a diagnosis means classifying an object (recognizing it as belonging to a class), the medical problem of diagnosis (classification) becomes the mathematical task of recognizing samples [6]. In the general case, the task of classifying (recognizing) an object reduces to the following: if we introduce an n-dimensional feature space $\{X_i\}$, where i = 1,2, ... n, then every j^{th} (j = 1,2 ... m) object in this space is represented by a point with coordinates $x_1, j, x_2, j, ..., x_n, j$, and each class of objects is represented by a set of such points. To classify an unknown object, that is, to recognize an image, means to determine which class the object belongs to based on an analysis of the values of its features.

A comparative classification of cluster analysis methods is presented in table 1.

TABLE I. ADVANTAGES AND DISADVANTAGES OF CLUSTERING METHODS.

Method	Advantages	Disadvantages
Clustering Using REpresentatives [7]	It performs clustering at a high level even in the presence of outliers, identifies clusters of complex shape and various sizes, has linearly dependent requirements for the data storage location and temporary complexity for high dimensional data.	There is a need to set threshold values and the number of clusters.
Balanced Iterative	Two-stage clustering, clustering of large	Working with only numerical

Method	Advantages	Disadvantages
Reducing and Clustering using Hierarchies [8]	amounts of data, runs on a limited amount of memory, is a local algorithm, can work with one scan of the input data set, uses the fact that the data is not equally distributed in space, and processes areas with high density as a single cluster.	data, it distinguishes only clusters of convex or spherical shape, there is a need to set threshold values.
Algorithm based on Minimum Spanning Trees [9]	Allocates clusters of arbitrary shape, including clusters convex and concave shapes, selects from several optimal solutions the most optimal.	Emission sensitive.
k- means [10]	Ease of use; speed of use; understandability and transparency of the algorithm.	The algorithm is too sensitive to outliers that can distort the average; slow work on large databases; it is necessary to set the number of clusters; the impossibility of applying the algorithm to data where there are intersecting clusters.
Partitioning around medoids [11]	Ease of use; speed of use; understandability and transparency of the algorithm, the algorithm is less sensitive to outliers in comparison with k-means.	It is necessary to set the number of clusters; slow work on large databases
CLOPE (Clustering with sLOPE) [12]	High scalability and speed of work, as well as the quality of clustering, which is achieved using the global optimization criterion based on maximizing the gradient of the height of the cluster histogram. It is easily calculated and interpreted. During operation, the algorithm stores in RAM a small amount of information for each cluster and requires a minimum number of scans of the data set. CLOPE automatically selects the number of clusters, and this is regulated by one single parameter - the repulsion coefficient.	
Self-organizing maps of Kohonen [13]	A universal approximator is used - a neural network, training a network without a teacher, self-organization of a network, ease of implementation, guaranteed receipt of an answer after passing data through layers.	Working only with numerical data, minimizing the size of the network, you must specify the number of clusters.
HCM Algorithm (Hard C - Means) [14]	Ease of implementation, computational simplicity.	Setting the number of clusters, lack of guarantee in finding the optimal solution.
Fuzzy C-means [15]	Blurring when defining an object in a cluster allows you to define objects that are on the border into clusters.	Computational complexity, setting the number of clusters, there is uncertainty with objects that are remote from the centers of all clusters.

II. RATIONALE AND PURPOSE OF THE RESEARCH

The C-Means clustering algorithm was proposed by J. Dunn in 1973 and finalized by J. Bezdek in 1981. Unlike most existing clustering algorithms, this algorithm is fuzzy – each of the objects does not belong uniquely to any cluster, but belongs to all clusters with different degrees of membership [16]. This gives advantages as a partition in cases where clusters are close to each other and a large number of points are located at their boundaries. However, the cost of such fuzziness is greater computational cost than in clear algorithms such as K-Means [17]. While maintaining such a drawback as a priori determination of the number of clusters, there is a possibility of the absence of a guarantee of a global optimum of the result.

Output data: array of objects $X_k \in \backslash R^n, k = \overline{1, M}$, number of clusters c, exponential weight $m \in [1, \infty)$. Stop parameter $\varepsilon > 0$.

Measured data: partition matrix F with size M x c (elements $\mu_{ki} \in [0,1], \sum_{i=1}^{c} \mu_{ki} = 1$). Cluster centers V_i, distances D_{ki} between objects and cluster centers.

The algorithm of the method (steps are calculated sequentially at each iteration):

1 step. Clarification of the centers of clusters by degree of affiliation:

$$V_i = \frac{\sum_{k=1}^{M} \mu_{ki}^m * X_k}{\sum_{k=1}^{M} \mu_{ki}^m}, i = \overline{1, c}, \quad (1)$$

2 step. Calculation of distances between new cluster centers and data points:

$$D_{ki} = \sqrt{\left\| X_k - V_i \right\|^2}, k = \overline{1, M}, i = \overline{1, c}, \quad (2)$$

At each iteration of the algorithm, matrix elements are refined. The output of the algorithm is the matrix into which the algorithm converges.

The fact that the algorithm converges is established by checking the form $\max_{k=\overline{1,M},i=\overline{1,c}}(\left|\mu_{ki} - \mu_{ki}^*\right|) < \varepsilon$ or $\max_{i=\overline{1,c}}(\left|V_i - V_i^*\right|) < \varepsilon$, where $\mu_{ki}^*(V_i^*)$ is the value $\mu_{ki}(V_i)$ calculated at the previous iteration.

The computational core of the C-Means algorithm consists of steps for calculating the centers of clusters, the distances between them and data points, and especially the recalculation of the matrix of degrees of belonging of data points.

When implementing the algorithm, some repetitive calculations can be eliminated. So, for the step of calculating the centers of clusters, the values μ_{ki}^m can be calculated once and multiplied by X_k when included in the amount recorded in the numerator. For the step of calculating the distances between the centers of the clusters and points, the operation of taking the square root is not necessary, since later on at the stage of calculating the degrees of belonging, you can directly use the squares of these distances, and the sum in the denominator will take the form:

$$\sum_{j=1}^{c} (\frac{D_{ki}^2}{D_{kj}^2})^{1/m-1}, \quad (3)$$

The algorithm includes three main stages - calculation of cluster centers, calculation of distances between cluster centers and data points (which includes macro operations, subtraction of vectors and calculation of their norms) and recalculation of the membership matrix.

The sequential complexity of the algorithm - when clustering M data objects represented by points in $\backslash R^n$ clusters, the C-Means algorithm in the sequential version has computational complexity:

$$O(c^2MI + cMnI), \quad (4)$$

where I is quantity of iterations.

If we assume that the data dimension is small, then this complexity reduces to $O(c^2MI)$. The main part of the algorithm in this case is the recalculation of the membership matrix, which requires the calculation of cM sums of the terms at each iteration.

III. Experiments and results of the modeling

As input data is a dataset of information about patients (CardiologyCategorical.xls) who have heart disease [8], a complete list of parameters can be seen in table 2.

TABLE II. Statistical data for analysis.

Age	Metrical scale	29..77
Sex	Nominal scale	Male/Female
Chest pain type	Nominal scale	Asymptomatic, Abnormal Angina, Angina, NoTang
Blood Pressure	Metrical scale	94..200
Cholesterol	Metrical scale	126..564
Fasting blood sugar < 120	Nominal scale	True/false
Resting ecg	Nominal scale	Normal/Hyp
Maximum heart rate	Metrical scale	71..202
Angina	Nominal scale	True/false
Peak	Metrical scale	0..6,2
Slope	Nominal scale	Flat, Down, Up
Colored vessels	Metrical scale	0, 1, 2, 3
Thal	Nominal scale	Normal, Rev, Fix

Figure 2 shows a portion of the imported data.

```
Initial data:
   age   sex  chest pain type  blood pressure
0   60  Male      Asymptomatic             130
1   49  Male  Abnormal Angina             130
2   64  Male           Angina             110
3   63  Male      Asymptomatic             130
4   53  Male      Asymptomatic             140

     slope  #colored vessels    thal    class
      Flat                 2     Rev     Sick
        Up                 0  Normal  Healthy
      Flat                 0  Normal  Healthy
      Flat                 1     Rev     Sick
      Down                 0     Rev     Sick
```

Fig. 1. Imported data

Figure 2 shows results of normalization of imported data.

```
Normalized data:
        age   sex  chest pain type  blood pressure
0  0.645833  Male      Asymptomatic        0.339623
1  0.416667  Male  Abnormal Angina        0.339623
2  0.729167  Male           Angina        0.150943
3  0.708333  Male      Asymptomatic        0.339623
4  0.500000  Male      Asymptomatic        0.433962

         slope  #colored vessels    thal    class
          Flat          0.666667     Rev     Sick
            Up          0.000000  Normal  Healthy
          Flat          0.000000  Normal  Healthy
          Flat          0.333333     Rev     Sick
          Down          0.000000     Rev     Sick
```

Fig. 2. Normalized data

As can be seen after normalization, the data (in the metrical scale) are in the interval [-3; 3]. For data in the nominal scale, normalization was not performed.

In fig. 3 you can see a scatter plot of the data.

The first iteration of the C-means method randomly selects points from the set date. In Figure 4 you can see the result of the choice of centers.

As you can see, the number of points is 2, that there are 2 clusters of data - the patient either has heart disease or not. In Figure 5, the data of the centers are presented in the form of green stars (*).

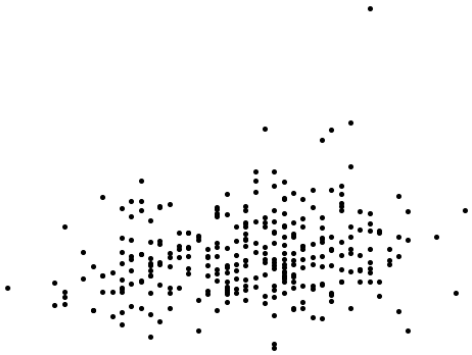


Fig. 3. Plot of normalized data

```
Initial centroids:
[[0.3125     0.25799087]
 [0.7083333  0.3127854 ]
```

Fig. 4. Centers of mass at the first iteration

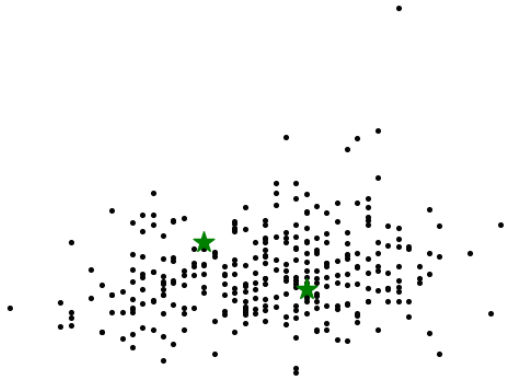


Fig. 5. Normalized data graph with centers

Figure 6 shows data separated by clusters.

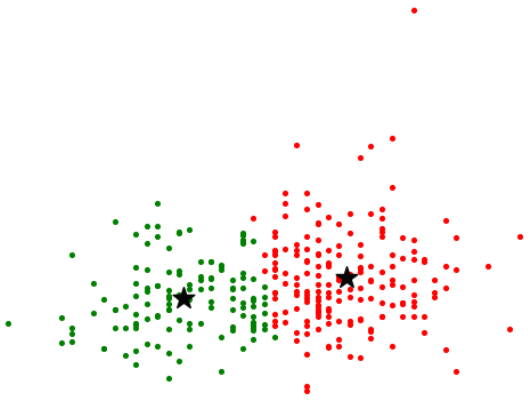


Fig. 6. Result of data clusterization

IV. Conclusions

The study analyzed the statistical relationship between variables that determine the condition of patients. It is shown how it is possible to determine whether a patient belongs to a certain class on the basis of statistical data - registered state variables.

The developed software solution, based on the Pandas library of the Python language, allows classification based on statistical data on patients, which provides a high recognition rate, as well as revealing the precedent belongs to one of several clusters.

It should be noted that the described approach is universal and can be used not only for biomedical systems, but also for technical, economic, etc.

References

[1] S. Kaptoge et. al. “World Health Organization cardiovascular disease risk charts: revised models to estimate risk in 21 global regions” The Lancet Global Health, Volume 7, Issue 10, 2019, e1332 - e1345.

[2] Yu. Polyvianna, D. Chumachenko, T. Chumachenko “Computer Aided System of Time Series Analysis Methods for Forecasting the Epidemics Outbreaks”, 2019 15th International Conference on the Experience of Designing and Application of CAD Systems (CADSM), 2019, pp. 7.1-7.4.

[3] H. D. Park, Y. Han, J. H. Choi, "Frequency-Aware Attention based LSTM Networks for Cardiovascular Disease," 2018 International Conference on Information and Communication Technology Convergence (ICTC), Jeju, 2018, pp. 1503-1505.

[4] Y. Yang, T. Chen, "Analysis and Visualization Implementation of Medical Big Data Resource Sharing Mechanism Based on Deep Learning," in IEEE Access, vol. 7, 2019, pp. 156077-156088.

[5] D. Chumachenko, et. al., "On Intelligent Decision Making in Multiagent Systems in Conditions of Uncertainty," 2019 XIth International Scientific and Practical Conference on Electronics and Information Technologies (ELIT), Lviv, Ukraine, 2019, pp. 150-153.

[6] D. Chumachenko, et. al., “Intelligent expert system of knowledge examination of medical staff regarding infections associated with the provision of medical care,” CEUR Workshop Proceedings, vol. 2386, 2019, pp. 321-330.

[7] M. Alfian, A. R. Barakbah and M. Febrian Ardiansyah, "Representative News Generation using Automatic Clustering in Big Data Environment," 2019 International Electronics Symposium (IES), Surabaya, Indonesia, 2019, pp. 655-659.

[8] K. Bazilevych, et. al., “Determining the probability of heart disease using data mining methods,” CEUR Workshop Proceedings, vol. 2488, 2019, pp. 383-394.

[9] A. Khan, A. Aesha, J. Sarker, “A New Algorithmic Approach to Finding Minimum Spanning Tree,” 2018 4th International Conference on Electrical Engineering and Information & Communication Technology (iCEEiCT), Dhaka, Bangladesh, 2018, pp. 590-594.

[10] I. Meniailov. et. al., “Using the K-means method for diagnosing cancer stage using the Pandas library,” CEUR Workshop Proceedings, vol. 2386, 2019, pp. 107-116.

[11] W. Zhou, H. Tang, Z. Ji, "A Task Partition Algorithm Based on Grid and Graph Partition for Distributed Crowd Simulation," 2014 Fourth International Conference on Instrumentation and Measurement, Computer, Communication and Control, Harbin, 2014, pp. 522-526.

[12] Z. Wang, "Determining the Clustering Centers by Slope Difference Distribution," in IEEE Access, vol. 5, 2017, pp. 10995-11002.

[13] I. Hammami, J. Dezert, G. Mercier, "Kohonen-Based Credal Fusion of Optical and Radar Images for Land Cover Classification," 2018 21st International Conference on Information Fusion (FUSION), Cambridge, 2018, pp. 1623-1630.

[14] R. C. de Araújo, F. d. A. T. de Carvalho, Y. Lechevallier, "Multi-view hard c-means with automated weighting of views and variables," 2017 International Joint Conference on Neural Networks (IJCNN), Anchorage, AK, 2017, pp. 2792-2799.

[15] R. Shang, et. al, “Sar Image Change Detection Based on Mean Shift Pre-Classification and Fuzzy C-Means,” IGARSS 2019 - 2019 IEEE International Geoscience and Remote Sensing Symposium, Yokohama, Japan, 2019, pp. 2358-2361.

[16] D. Chumachenko, T. Chumachenko, “Intelligent Agent-Based Simulation of HIV Epidemic Process,” Advances in Intelligent Systems and Computing, vol. 1020, 2019, pp. 175-188. (2019).

[17] K. Bazilevych et al. “Stochastic modelling of cash flow for personal insurance fund using the cloud data storage”, International Journal of Computing, vol. 17, iss. 3, 2018, pp. 153-162.

IEEE Third International Conference on Data Stream Mining & Processing
August 21-25, 2020, Lviv, Ukraine

Application of Deep Learning in Neuromarketing Studies of the Effects of Unconscious Reactions on Consumer Behavior

Bogdan Glova
Department of Software Engineering
Ternopil Ivan Puluj National Technical University
Ternopil, Ukraine
bogdanglova105@gmail.com

Ivan Mudryk
Department of Software Engineering
Ternopil Ivan Puluj National Technical University
Ternopil, Ukraine
i1mudryk@ukr.net

***Abstract* — The prospects of development of neuromarketing are investigated and the essence of neuro-technologies is considered. A number of benefits of neuromarketing research have been highlighted. In particular, the possibility of obtaining specific information about peculiarities of consumer buying behavior and scientific substantiation of consumer reactions to various incentives are included in the key ones. Prospects for the development of neuromarketing methods and tools are considered. By means of neuromarketing, we will have the chance to understand the basics of the decision-making mechanism.**

Keywords — neuromarketing, neuromarketing methods, neuromarketing research, neuromarketing tools, emotions, neuroimaging

I. INTRODUCTION

Neuromarketing trend has a short history, but in recent years it is spreading all over the world. Over the last 100 years, the situation in the consumer market has changed dramatically. Young companies took the lead position and moved the undefeated giants to the lower sabers. They have obliged the current built-up marketing strategy by success. Marketing plays a key role in the life of any company. In recent years, a promising area of research called neuromarketing has emerged. Neuromarketing researching is gaining popularity due to speed and productivity that require fewer resources for the organization.

The purpose of neuromarketing research is to obtain objective information about the personal preferences of consumers without resorting to subjective data obtained by traditional marketing means The essence of the study is to measure the respondent's attention, emotions, and memory. During the study, the respondent's physiological responses to various stimuli (advertising samples, product packaging, logos) are recorded, such as changes in the activity of different brain areas, pulse and respiratory rate, skin moisture, pupil movement, etc. For their registration special equipment is used - electroencephalographs (EEG), magnetic resonance imaging, pupil tracking systems (the so-called eye-tracker), etc. This study used EEG data.[1]

II. ADVANTAGES AND DISADVANTAGES OF NEUROMARKETING

Analyzing the notion of neuromarketing, as well as ethical concepts of its application, clearly distinguish advantages and disadvantages. Unlike traditional marketing, the advantage of neuromarketing is innovation and the ability to obtain new information. Since the brain is not visible to the naked eye, neuromarketing allows him to find out more about it. Using his techniques, he can explore subconscious responses and watch brain activity in humans and their reaction to marketing tools that consumers are not aware of. In addition, neuromarketing can add value to consumers.

Neuro-marketing information makes it easier to connect consumers with products or services. Given that consumers often give subjective responses to preferences and tastes, the advantage of neuromarketing is to give more objective measurements and results. It is known that emotions play an important role in making decisions, but it is often difficult to measure them. With neuromarketing techniques such as facial encoding, emotional reactions and emotions related to packaging, web site or color can be measured. On the other hand, the primary drawback is the fact that it is difficult to generalize, or to make conclusions pertaining to every consumer.[2]

The reason is that relatively small samples (about 15 people) are often used for analysis with fMRI. The great lack of fMRI technique is that it does not give "live" images. Researchers can not see how the brain responds to advertising or real-time photography. For researchers, the brain is still a big secret and they know nothing 100% safe, because different functions will always show overlapping with different brain areas. With the development of neuromarketing, some of its disadvantages may disappear, and techniques will become more advanced but more accessible. It's already possible to buy glasses for eye tracking! Therefore, traders must keep in mind the needs of their consumers. Both traditional marketing and neuromarketing have their advantages and disadvantages (Odekerken, 2018). They tell us something about how consumers (or people in general) make decisions, were they aware or subconscious.The neural network models the work of the human nervous system, a feature of which is the ability to self-study taking into account previous experience. Thus, every time the system makes fewer mistakes. Like our nervous system, the neural network consists of separate computational elements - neurons located on several layers. The data that entering into the neural network input is sequentially processed at each layer of the network.

In the course of the study data that were taken from the respondent, namely, his brain activity in the emergence of

978-1-7281-3214-3/20/$31.00 ©2020 IEEE

extraneous stimuli of taste, smell, hearing, emotional background when viewing different video sequences, as well as brain activity during the movement of the thumb or index finger. When transferring them to different neural network models (such as: MLPClassifier, RandomForestClassifier, ExtraTreesClassifier, DecisionTreeClassifier, AdaBoostClassifier, GradientBoostingClassifier, BaggingClassifier (KNeighborsClassifier)), the calculations were calculated, and time of experience.

TABLE I. NUMBERS FOR MLPCLASSIFIER

	MLPClassifier			
Classifier metric	precision	recall	f1-score	accuracy
A sad and funny video	0.90	0.91	0.90	0.90
Chocolate and onion	0.97	0.88	0.93	0.93
Heavy and light music	0.96	0.98	0.97	0.97
Forefinger and thumb	0.79	0.77	0.78	0.78
Nashir and perfume	0.88	0.94	0.91	0.91

After analyzing the numerical values of the f1-score column and other metrics, the most accurate and effective classifier that is suitable for this task was selected, the MLPClassifier, its numerical values are listed in Table 1. The F1 score can be interpreted as the weighted average of the accuracy and response of the irritation, where the F1 score reaches its best value at 1 and a worse value at 0. The impact on the F1 accuracy and response score are equal. F1:

$$F1 = 2 * \frac{(precision * recal)}{(precision + recal)} \quad (1)$$

Based on the f1-score interpretation, as this metric is fundamental to the algorithm's evaluation, the above classifier was selected. [3]

Was done selecting attributes to simplify model learning, increase the ability to interpret and simplify the model, and increase accuracy, the data obtained were analyzed and selected among the best, indicating a specific brain region when irritating the respondent with various factors. The data obtained during the work of olfactory receptors revealed the most relevant attributes of the entire set of objects to predict the best model.

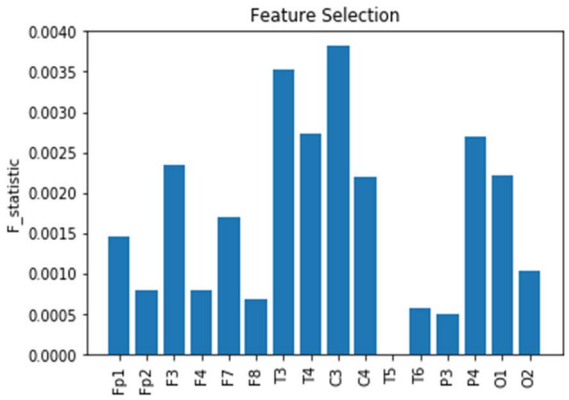
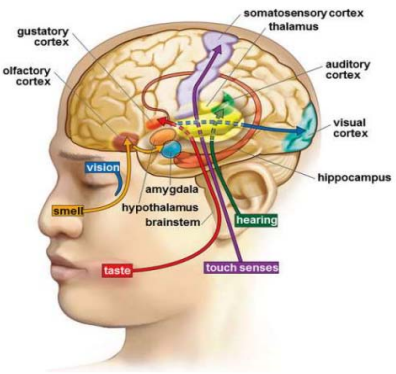


Fig.1 Statistical Data Sampling

On Figure 1 clearly shows the different electromagnetic potentials at different points of contact of the electrolytes to the main ones. The electrode that has the highest potential is key in training the model. Therefore, you should see the data in the highest statistics view.

a)



a)

b)

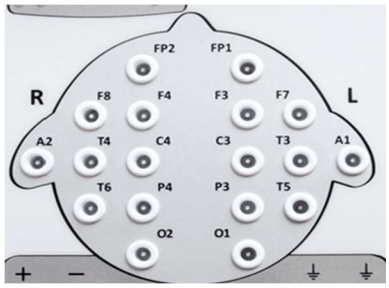


Fig.2 (a) Brain map with parts of sensitive zone
(b) Encephalography map

Figure 2 notes the great potential on the T3, T4, C3, C4 electrodes. The main part of this encephalography in Picture 2 for the use of the activated brain region in the sense of smell. They were also addressed in the same fields of the brain.

III. NEUROANATOMY / STRUCTURE

Just as neural activity can be indicative of disease, so can neuroanatomy. As such, it is often possible to take anatomical measurements and use machine learning to diagnose disease. For example, researchers can distinguish between Alzheimer's disease and healthy brains of older adults using MRI scans (Sarraf, Tofighi, and Others 201). More generally, neuroanatomical measurements such as structural MRI and diffusion tensor imaging (DTI) can distinguish healthy from unhealthy patients across many conditions including schizophrenia, depression, autism and ADHD (Arbabshirani et al. 2017; Rathore et al. 2017; Vieira, Pinaya, and Mechelli 2017).

Sometimes, ML enables surprising opportunities. For example, using deep convolutional neural networks, we can surprisingly predict cardiovascular risk factors from retinal fundus photographs (Poplin et al. 2018). The future will undoubtedly see ongoing efforts to automatically detect disease from biological data. Since the majority of research in neuroanatomy is based on imaging techniques, recent advances in computer vision using ML are becoming important tools for neuroanatomy. Thus, segmenting and labeling parts of an image, which usually requires manual annotation, is an especially important task. However, as imaging techniques improve and the volume of data increases, it will become infeasible to rely on manual annotation. [4]

TABLE II. INSTRUMENTAL NEUROMARKETING RESEARCH METHODS

Eye tracking system	
The essence of the method	An apparatus that tracks the direction of eye movement and changes in pupil size
Opportunities for marketing	- determining the time of attention; - determination of speed of perception; - construction of the route of study of the object, object; - degree of usability or usability;
Advantages	- low price; - easy to interpret results; - non-invasiveness of the method; - there is a great deal of research using this method; - possible use of the apparatus in experiments with others research methods;
Disadvantages	- psychological discomfort; - Qualified staff required and dependency of results from the method of interpretation;
Electroneuromyography (EMG)	
The essence of the method	Method for the study of the nervous system through the study of electrical
Opportunities for marketing	Investigation of the emotional response of the subject
Advantages	- relatively cheap and sensitive method; - registration of specific reactions; - possible use of the apparatus in experiments together with other methods of research; - there is a great deal of research using this method; - non-invasiveness of the method;
Disadvantages	- psychological discomfort in the subject; - the need for qualified personnel;
Electroencephalography (EEG)	
The essence of the method	The essence of the method is to record the total bioelectric activity of the brain.
Opportunities for marketing	Studies of attention, emotional reactions, memory and other studies of functional states of the brain.
Advantages	- relatively cheap and sensitive method; - possible use of the apparatus in experiments together with other methods of research; - there is a great deal of research using this method; - the possibility of obtaining a three-dimensional localization of the detected activity; - non-invasiveness of the method;
Disadvantages	- psychological discomfort in the subject; - the need for qualified personnel; - high price for expert class apparatus; - non-specificity of the received data;
Caused potentials	
The essence of the method	Electrical activity of brain neurons in response to stimuli.
Opportunities for marketing	Studies of attention, emotional reactions, memory and other studies of functional states of the brain.
Advantages	- low price for devices not belonging to the expert class; - non-invasiveness of the method; - there is a great deal of research using this method; - possible use of the apparatus in experiments together with other methods of research; - high availability and sensitivity of the device;
Disadvantages	- high price for expert class apparatus; - psychological discomfort in the subject; - the need for qualified personnel; - the data are obtained mainly by the activity of the cortical area of the brain;
Magnetoencephalography (MEG)	
The essence of the method	This technology is capable of measuring the visualization of magnetic fields that result from brain activity.
Opportunities for marketing	Studies of attention, emotional reactions, memory and other studies of functional states of the brain.
Advantages	- possible use of the apparatus in experiments together with other methods of research; - high sensitivity of the device; - the possibility of obtaining a highly detailed three-dimensional model of the active site;
	- non-invasiveness of the method;
Disadvantages	- the high price; - psychological discomfort in the subject; - the need for qualified personnel; - the data are obtained mainly by the activity of the cortical area of the brain
Functional Magnetic Resonance Imaging (fMRI)	
The essence of the method	Allows you to map the brain to determine the features of the brain areas responsible for language, movement, vision, memory, and more.
Opportunities for marketing	Studies of attention, emotional reactions, memory and other studies of functional states of the brain.
Advantages	- the ability to explore almost all structures of the central nervous system; - high sensitivity of the device; - the possibility of obtaining a highly detailed three-dimensional model of the active site; - non-invasiveness of the method;
Disadvantages	- the high price; - psychological discomfort in the subject; - the need for qualified personnel; - complicated use of the apparatus in experiments together with other research methods; - the need for the absence of various metal prostheses, implants, neurostimulants;
Magnetic resonance spectroscopy (MRS)	
The essence of the method	Allows non-invasive information on metabolism in the brain
Opportunities for marketing	Studies of attention, emotional reactions, memory and other studies of functional states of the brain.
Advantages	- high sensitivity of the device; - the possibility of obtaining a highly detailed three-dimensional model of the active plots; - non-invasiveness of the method; - the ability to explore almost all structures of the central nervous system;
Disadvantages	- the high price; - psychological discomfort in the subject; - the need for qualified personnel; - complicated use of the apparatus in experiments together with other research methods; - the need for the absence of various metal prostheses, implants, neurostimulants;
Positron emission tomography (PET)	
The essence of the method	With this method it is possible to trace the distribution of biologically active compounds in the human body.
Opportunities for marketing	Studies of attention, emotional reactions, memory and other studies of functional states of the brain.
Advantages	- high sensitivity of the device; - the possibility of obtaining a highly detailed three-dimensional model of the active site; -the ability to explore virtually all structures of the central nervous system
Disadvantages	- Due to the high degree of invasiveness and, as a consequence, complicated nervous system performance, this method is extremely rarely used in neuroeconomic studies.
Transcranial magnetic stimulation (TMS)	
The essence of the method	Allows non-invasive stimulation of the cerebral cortex with short magnetic pulses.
Opportunities for marketing	Studies of attention, emotional reactions, memory and other studies of the functional states of the brain, as well as simulation of changes in the consumer's mind.
Advantages	- high sensitivity of the device; - selective action of the device on specific areas; - non-invasiveness of the method;
Disadvantages	- the high price; - psychological discomfort in the subject; - the need for qualified personnel; - complicated use of the apparatus in experiments together with other research methods; - it is possible to study only brain structures at depths up to 2 cm

For the solving of this problem, many ML techniques have been developed to automatically segment or label new images based on a dataset of previously labeled images. The vast majority of techniques in this developing literature are based, at least in part, on convolutional neural networks (Litjens et al. 2017; Ronneberger, Fischer, and Brox 2015). This approach has been employed to label medical images, such as in identifying white matter tracts from MRI scans (Ghafoorian et al. 2017). They have also been used to understand the connections and morphologies of neurons from electron microscopy (Helmstaedter et al. 2013; Funke et al. 2017; Turaga et al. 2010). As imaging data improve in resolution and volume, ML is becoming a crucial and even necessary tool for reconstructing and mapping neuroanatomy.

IV. Applying Artificial Intelligence to Marketing

In complex systems that operate in an environment of uncertainty, they try to use Data Mining for optimal processing of input information.

Equally important for such decision-making systems. These tasks are in many ways close to the problems that are constantly being solved by living organisms.[5]

Therefore, given the relative novelty of the considered problems for complex systems and the evolution of living organisms to perfection the relevant mechanisms of memory and decision making, there is an increased interest in the analysis of the work of these mechanisms in living organisms and their emulation to create complex technical systems. In the process of information perception, awareness and decision-making (PR), the memory plays a decisive role and, therefore, in his actions and intentions. In addition to neurophysiologists and behavioral psychologists, the organization of human memory is actively interested in specialists in the field of cognitive psychology and decision theory. As well as the creators of artificial neural networks, neurocomputers, and artificial intelligence (USI) systems. When constructing a computer-brain model, we will take into account the peculiarity of living organisms, due to the fact that they, as predicted by the US, are constantly operating in conditions of considerable uncertainty of the environment.

Talking about applying artificial intelligence to marketing, it is important to point out that the contemporary business of every company in greater or lesser intensity implies some form of artificial intelligence. At the growing level of integration of artificial intelligence into marketing, the following is achieved:

A. Mass Data Analysis

In the past private labels and agencies have employed data team analysts. However, while the team of analysts prepared the database, the data in them was already outdated. With the appearance of artificial intelligence analysis and data processing becomes less and more demanding and increasingly efficient. This means that analysts need less time to analyze data and have more time to find answers to questions about them. Today, science becomes more and more accessible because of modern computer infrastructure that supports artificial intelligence and machine learning such as Amazon AWS plus Apache Spark, Google Cloud Machine Learning Engine, and Microsoft Azure Machine Learning Studio. [6]

B. Understanding Buyer Behavior

In marketing, machine learning is increasingly helping to predict human behavior, which allows it to create additional value for the consumer. Advertising campaigns that apply artificial intelligence can analyze whether a consumer responds well to a particular ad and so adjusts it for the next release.

C. Improving user experience

Artificial intelligence already has a major impact on user experience with tools such as Google Assistant, Amazon Alexa, and Apple Siri. These tools have become a major part of our daily life and their role will grow more and more. Artificial Intelligence Marketers can use to improve data processing, mapping consumer times, optimizing bids, and improving overall user experience. [4-5]

V. Conclusion

Over the last several years, the use of machine learning (ML) in neuroscience has been rapidly increasing. Here, we review ML's contributions, both realized and potential, across several areas of systems neuroscience

By means of neuromarketing, we have the chance to understand the basics of the decision-making mechanism. Thus, we will be able to access a superior level of knowledge on the consumer's behavior. They will be the basis of understanding the way in which consumers respond to various marketing stimuli. This approach is going to be a difficult one, given the context, as per the way it is presented in the specialized literature, according to which the decision-making process can activate various cortical areas.

The study used a range of neuromarketing technologies and neural algorithms to simplify the classification of the effects of unconscious reactions on consumer behavior. The use of technology records in real-time the response of the respondents' subconscious to the investigated materials and allows to consider practically any third-party factors that may affect the course of the research. This allows us to measure the perception and, in the aggregate of the results obtained, improve the accuracy and give an objective assessment.

References

[1] T.Z. Ramsøy, "A foundation for consumer neuroscience and neuromarketing", 2019. 10.13140/RG.2.2.12244.45446.

[2] A. Emić and S. Čabro, Artificial Intelligence and Neuromarketing Vještačka Inteligencija i Neuromarketing, 2019.

[3] Library documentation «scikit-learn». Access mode: [https://scikit-learn.org/stable/].

[4] E. Choy, "Neuromarketing: 3 ways neuromarketing help you captivate audiences", 2018. Available at: [https://www.america-retail.com/neuromarketing/neuromarketing-3-ways-neuromarketing-help-you-captivate-audiences].

[5] D. Mykhalyk, I. Mudryk, A. Hoi and M. Petryk, "Modern hardware and software solution for identification of abnormal neurological movements of patients with essential tremor", proceeding of 2019 9th International Conference on Advanced Computer Information Technologies (ACIT), Budejovice, Czech Republic, pp. 183-186, 2019.

[6] V. Sebastian, "Neuromarketing and evaluation of cognitive and emotional responses of consumers to marketing stimuli", Procedia - Social and Behavioral Sciences, 127, 2014, pp. 753 – 757.

IEEE Third International Conference on Data Stream Mining & Processing
August 21-25, 2020, Lviv, Ukraine

Evaluating Power Consumption Model and Load Deficit at Different Temperatures Using Clustering Techniques and Presenting a Strategy for Changing Production Management

Mesbaholdin Salami
Department of Industrial Engineering
Central Tehran Branch
Islamic Azad University
Tehran, Iran
mesbah.salami@yahoo.com

Farzad Movahedi Sobhani
Department of Industrial Engineering
Science and Research Branch
Islamic Azad University
Tehran, Iran
farzad.movahedi@gmail.com

Mohammad Sadegh Ghazizadeh
Department of Electrical Engineering
Abbaspour School of Engineering
Shahid Beheshti University
Tehran, Iran
ghazizadeh.ms@gmail.com

***Abstract*—The present study evaluated the consumption model among the subscribers of the power industry at a certain temperature range by focusing on the amount of load deficit and using clustering techniques in data mining. This study an initial dataset including the load demand, power plant reserve, power import, and power export at the range of -12 to -50 degrees centigrade. Then, the optimal algorithm was selected by evaluating three clustering algorithms of Single Linkage, Complete Linkage, and Average Linkage and comparing to K-means algorithm using the key indices of clustering performance. Finally, a power export model was proposed to reduce load deficits and an appropriate solution to compensate for this deficit. The results indicated that the supply and distribution network faced load deficit and caused electric breakdowns due to the increased demand for power consumption. This study identified the desired temperature range and proposed some suggestions for reducing the volume of load such as the reduced amount of export, increase of import, and the increase in power plant production in accordance with capacity.**

Keywords— load deficit, clustering algorithms, single linkage, complete linkage, average linkage, K-means

I. INTRODUCTION

In the studies on power systems, electrical loads are typically determined by definitive or probabilistic models. In definitive models, the system loads are modeled by fixed power values. These fixed values include: (1) the average load demand curve in daily, monthly, and seasonal periods obtained from historical data (Omran et al., 2010); (2) the fixed values being related to different loading conditions such as peak load, average or minimum load. These definitive models are not suitable for evaluating long-term load behaviors. In fact, using the probabilistic models can solve this problem. For this reason, the probabilistic clustering techniques were used in this study to analyze load deficit. The main purpose of presenting the dataset to the above-mentioned algorithms was to classify the current load status into two categories of normal status and load deficit status and determine the extent to which the temperature ranges of this load deficit are more obvious. After calculating each algorithm, the outputs of the algorithm were indicated in the relevant sections. Here is the literature review in this area.

A study using modified-follow-the-leader, K-means, fuzzy K-means, and hierarchical clustering algorithms to classify the subscribers of power in non-residential sectors and compared the above-mentioned methods using a set of clustering validation parameters. The results indicated that modified follow-the-leader from clustering algorithms and average linkage from hierarchical clustering group have a very high accuracy for clustering the customers. In addition, some techniques were used to reduce the size of data (Chicco et al., 2006). In another study, the hourly power consumption data were analyzed in the household sector using k-means, k-medoid and SOM algorithms. Then, the best technique was used to partition the families into different clusters (McLoughlin et al., 2015). Another study discovered power consumption models in the household sector by providing a data mining framework. First, the data were classified into three classes using the k-mean clustering algorithm and then the changes in the power consumption model of each class were examined. The results of this study indicated that the volatility of power consumption in winter and summer is higher than in spring and fall while instability was observed in household power consumption models. Three types of load profile were observed in spring festivals while two types of load profile were observed on the workday and national day. High temperature in summer and low temperature in winter indicated a significant effect on power consumption (Guo et al., 2018). Panapakidis et al (2014) used the clustering techniques of SOM, FCM and K-means to examine the load curve in nine faculties of AUTH university. Other researchers examined the features of MV power consumers using the data mining technique. In fact, they analyzed different types of consumer load profiles and created a set of rules for the automatic classification of new consumers (Sérgio et al., 2015). In another study, the probabilistic load was modeled using the clustering algorithms. The presented approach began using 24 data points indicating the hourly loading per day. In this regard, 365 data were collected within a year to conduct the analyses. Then, the data were clustered using K-means ،fuzzy K-means ،Ward's linkage ،Single linkage و Complete linkage, and Average linkage. Then, the most optimal methods were selected using the performance parameters. The results indicated that the proposed model can represent the data for or loading and cover the results for the whole year (ElNozahy et al., 2013).

The research method is explained in detail in the next sections of the study and then the results of applying the above-mentioned algorithms are analyzed. Finally, the best clustering technique is selected using the various parameters and a power export model is proposed to reduce load deficits

978-1-7281-3214-3/20/$31.00 ©2020 IEEE

and present a suggestion to take the solution to compensate for such a deficit.

II. Research Method

In this study, Single Linkaget, Complete Linkage and Average Linkage which are among the most common clustering algorithms and K-mean algorithm were used to investigate the power consumption model and load deficit at different temperatures and present a solution for changing the production management, load import and export to reduce load deficits.

IIA. Data Collection

The desired data in this study were collected from statistical databases on the coldest and hottest days of the year. Such data included load demand, power plant reserve, power export and import at the temperature range of -12 to 50 ° C, being collected in an Excel file for 1600 samples.

IIB. Data Preparation

Since there are several fields from the above-mentioned parameters at a given temperature, it was necessary to average the records for the specified temperatures in the first step. Thus, one record including the average parameters of load demand, power plant reserve, power export, and power import was created for each temperature at the temperature range -12 to 50 ° C. In this situation, clustering algorithms are applied on the data set more effectively. In order to achieve this issue, first the input data should be modified. Then, the average data of each parameter were averaged in an Excel file. Then, the data were presented to the desired algorithms for secondary calculations. Here are the results of this averaging in Figures 1, 2, and 3.

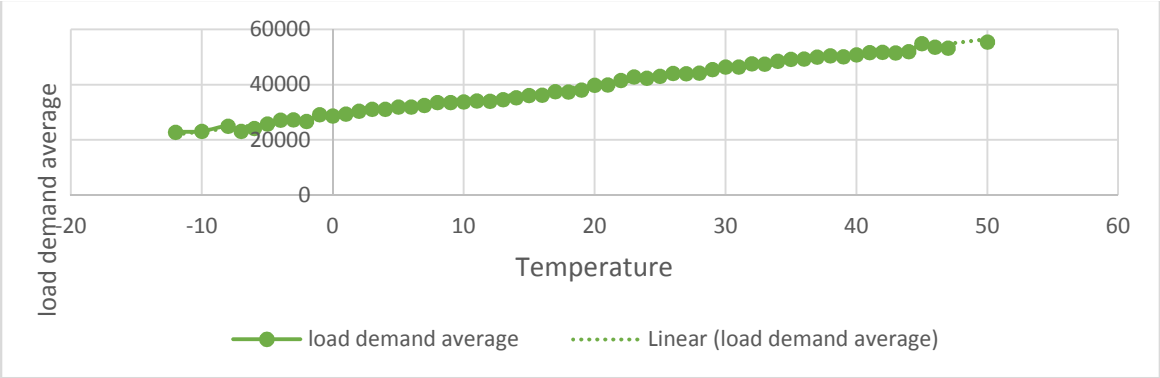


Fig. 1. Average load demand at different temperature

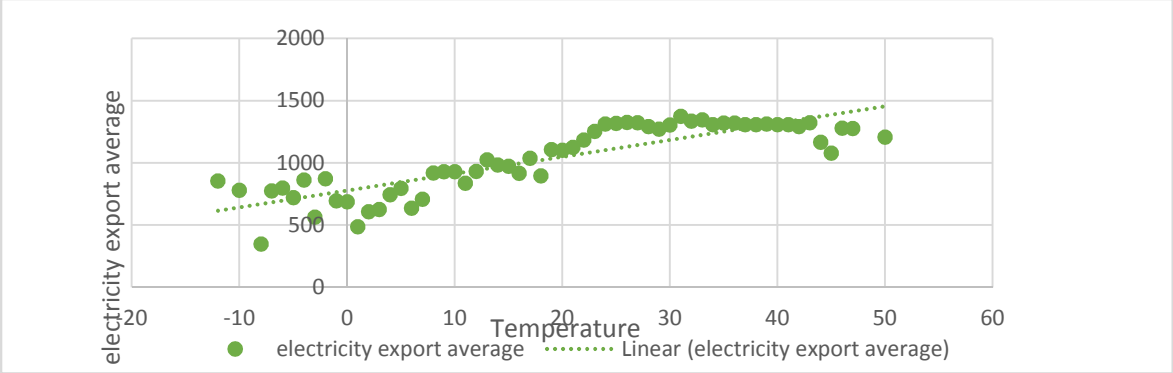


Fig. 2. Average power exports at various temperatures

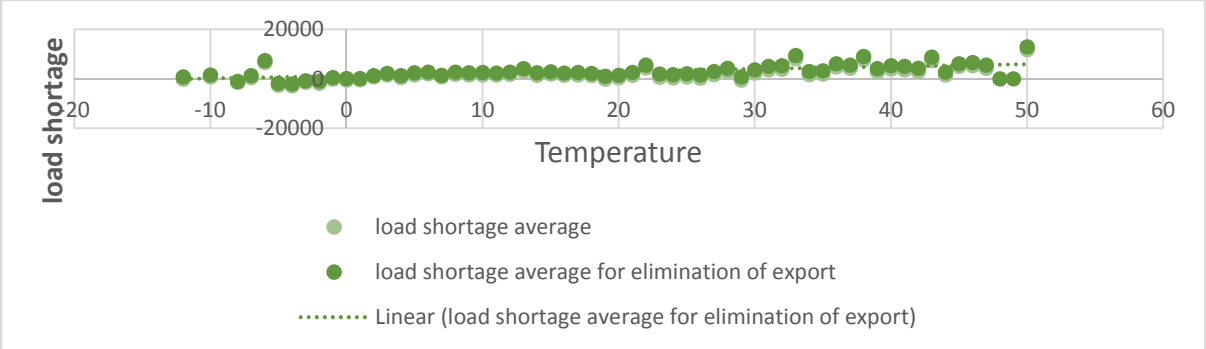


Fig. 3. Average load deficit in case of removing export at different temperatures

IIC. CLUSTERING ALGORITHMS

Clustering algorithms are divided into two groups of exclusive and overlapping. In exclusive clustering algorithms, each data segment belongs to only one cluster while in overlapping algorithms, each data segment may belong to more than one cluster with different member degrees. Exclusive clustering algorithms can be divided into two groups of hierarchical and partitional. Data segments in partitional clustering algorithms are divided into a predetermined number of segments without creating a hierarchical structure. However, hierarchical clustering aims to create a hierarchy of clusters with a sequence of distributed partitions. In fact, it moves from each cluster into a cluster containing all data segments or vice versa. The first mode of clustering is called agglomerative hierarchical in which each datum is considered as a separate cluster. Then, the clusters which are more similar to each other are combined together during a repetitive process at each stage to result in cluster or a certain number of clusters. The second mode of clustering is called divisive hierarchical in which all data are first considered as a cluster. Then, the data which are less similar to each other are broken into distinct clusters during a repetitive process at each stage and this process s continues until the clusters with one member are achieved. (Xu, R., & Wunsch, 2009).

Single linkage clustering

Single linkage clustering is a hierarchical clustering approach in which each point is compared to other points using a bottom-up strategy. In this method, each component is placed in a separate cluster and the closest pair of clusters is merged together in each stage and this process continues until the conditions are satisfied (patra et al., 2011). This method is considered as the nearest neighbor and the distance between a pair of clusters is determined by two of the closest objects (Panapakidis et al., 2012). In fact, each datum is considered as a cluster and finding the nearest cluster is finding the least distance between the data.

$$D\left(C_l,\left(C_i,C_j\right)\right)=\min\left(D\left(C_l,C_i\right),D\left(C_l,C_j\right)\right)$$
$$=\frac{1}{2}D(C_l,C_i)+\frac{1}{2}D\left(C_l,C_j\right)-\frac{1}{2}\left|D(C_l,C_i)-D\left(C_l,C_j\right)\right| \quad (1)$$

Complete linkage clustering

This technique is known as the furthest neighbor technique. In fact, this technique uses the farthest distance between a pair of attributes for determining the distance of the inner cluster (Panapakidis et al., 2012).

$$D\left(C_l,\left(C_i,C_j\right)\right)=\max\left(D\left(C_l,C_i\right),D\left(C_l,C_j\right)\right)$$
$$=\frac{1}{2}D(C_l,C_i)+\frac{1}{2}D\left(C_l,C_j\right)+\frac{1}{2}\left|D(C_l,C_i)-D\left(C_l,C_j\right)\right| \quad (2)$$

K-Means Clustering

This method is considered as one of the easiest learning algorithms solving well-known clustering problems. In addition, this method is an easy and simple method classifying the data by a certain number of clusters (k clusters hypothecially) (MacQueen, 1967). The stages of this algorithm are as follows:

1. *K points are randomly selected as the initial cluster centers*
2. *The number of k clusters is tested by putting each point in the nearest cluster*
3. *Cluster centers are renewed*
4. *Stages 2 and 3 are repeated*
5. *Clustering results are obtained when the convergence criterion is obtained.*
6. Data analysis

Based on the K-Means clustering technique, the best clustering minimizes the total similarity between the cluster centers and all cluster members. In addition, it minimizes the total similarity between the cluster centers. In order to select the best cluster, this range is usually selected. Then, the ρ (k) value was calculated for each k value. A part of k, where ρ (k) is maximized, was selected as the optimal number of clusters. In this regard, a number of clusters was selected for which the distance between the cluster centers and the similarity of cluster centers with the members within each cluster was maximum. The quality of clustering results with K clusters is defined as Equation 3. In this equation, O indicates the cluster centers, Cn indicates the cluster centers, On indicates the set of elements which were not selected as cluster centers.

$$O=\left\{c^n\middle|n=1,\ldots,k\right\}$$
$$O^n=\left\{C_i\middle|i=1,\ldots,\left\|T^c-O\right\|\right\}$$
$$\rho(k)=\frac{1}{k}\sum_{n=1}^{k}\left(\min\left\{\frac{\eta_n+\eta_m}{\delta_{nm}}\right\}\right)$$
$$\eta_n=\frac{1}{\left\|O^n\right\|}\sum_{c_i\in O^n}Sim(c_i,c^n)$$
$$\eta_m=\frac{1}{\left\|O^m\right\|}\sum_{c_j\in O^m}Sim(c_j,c^m)$$
$$\partial_{nm}=Sim(c^n,c^m) \quad (3)$$

The following results are obtained due to the above-mentioned calculations on the data sets.

Since it was necessary to identify the temperatures with load deficit and temperatures of load- deficit y temperatures, the k value was considered as 2 in the current algorithm. Based on Figure 4, the initial data set was divided into two parts. The green part indicates the temperatures in which a lot of load deficits occurred.

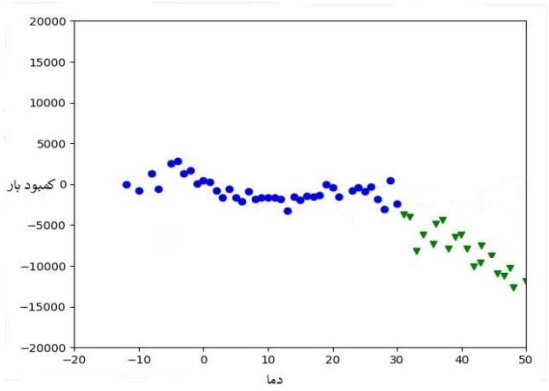


Fig. 4. Applying the K-means algorithm to the dataset with K = 2.

The necessary measures should be adopted at these points to increase the power reserve, reduce export, and increase import in line with the amount of deficit. Here are the results of applying clustering algorithms on the initial data set.

As indicated in the above figures, the hierarchical algorithms applied on the initial data set had almost the same function. Such algorithms considered the load deficit values related to

different temperatures. At the first step, the temperatures at which load deficit values were close to each other were placed in one cluster. After placing the pair of values close to each other related to temperatures in the cluster of level one, the mean load deficit values related to the first-level clusters were considered again and the clustering process was applied to this set of clusters resulting in the creation of the second-level clusters. The cluster integration process continued until all data were placed in form of a cluster. For example, two clusters were created in complete linkage algorithm in which he loads deficit values for the temperatures in each cluster were more similar to each other in comparison to all load deficit values associated with the temperatures in another cluster.

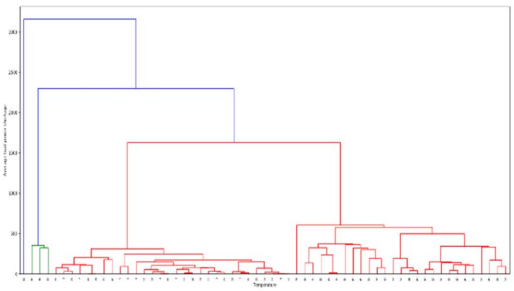


Fig. 5. Applying the single linkage algorithm on the initial data set

Table 1 summarizes the export and import results in each cluster. Based on this table, the seasons of winter, fall, and spring were placed in cluster 1 and the average temperature and average load deficits of each season were shown. Accordingly, the amount of power generation, import and export was determined.

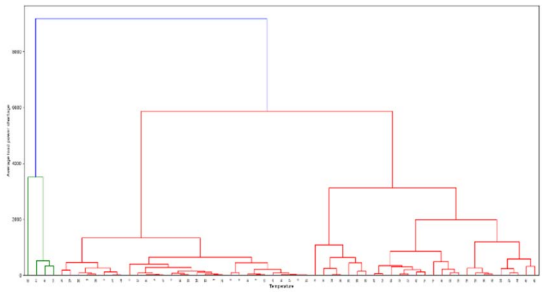


Fig. 6. Applying the average linkage algorithm on the initial data set

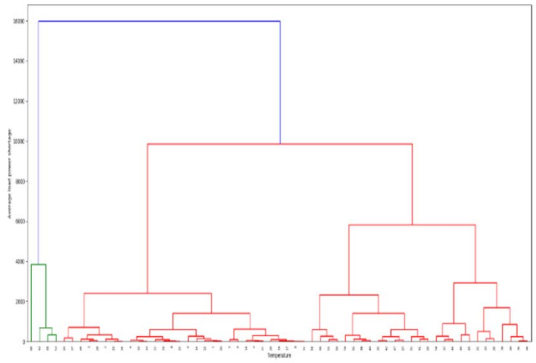


Fig. 7. Applying the complete linkage algorithm on the initial data set

As displayed in the table, no deficit occurred in the first three seasons of the year and power could be exported while there was a lot of load deficit in summer when power export should be prevented and its generation and import should be added.

TABLE I. EXPECTED STATUS OF POWER GENERATION AND DISTRIBUTION NETWORK FOR SOLVING THE PROBLEM OF LOAD DEFICIT

row	Season	Cluster number	Average temperature	Average deficit	Expected increase in generation	Expected increase in import	Expected increase in export
1	winter	1	-5	-971	0	0	300
2	autumn		11	-1115	0	0	450
3	Spring		26	-1311	0	0	280
4	Summer	2	42	5581	4600	400	-580

Based on the analysis on the results obtained from the K-means algorithm, in summer when an increase was observed in temperature, the status of power generation and distribution network should be selected based on the results presented in Table 2. Due to the average temperature of 42 degrees centigrade, the average deficit should be set as 5581. In fact, the expected increase in generation should equal 4600, the expected increase in import should equal 400, and the expected increase in export should equal -580 to solve the problem of load deficit.

III. VALIDATING AND COMPARING THE TECHNIQUES WITH DIFFERENT PARAMETERS

This section examines and compares the desired clustering algorithms being already applied to the initial data set based on 10 separate parameters. All technical aspects of the algorithms were covered in selecting the parameters. These parameters include:

- High intra-cluster similarity
- Low inter-cluster similarity
- Determining the number of clusters
- The independence of clusters from each other
- The reduction of overall distortion
- Local optimization
- Multiple initializing and selecting the best answer
- Appropriate run time
- Real optimization
- Convergence speed

Table 1 shows the numerical value between 0 and 10 for each of the above-mentioned parameters based on the conducted tests. Larger numbers indicate the better performance in this attribute by the desired algorithm.
Based on the results of Table 2, the K-mean algorithm can be a better option for solving the above-mentioned problem.

TABLE II. COMPARING THE ALGORITHMS BASED ON THE PROPOSED PARAMETERS

	High within-cluster similarity	Low inter-cluster similarity	Need the number of clusters	Independent of Partitions	Minimizes the total distortion	Local optimum	Multiple initializations and choose the best result	Efficient run-time wise	Exact optimization	Convergence speed
K-mean	10	9	10	10	10	10	10	10	5	6
Single Linkage	8	7	0	0	5	3	0	4	5	8
Average Linkage	8	7	0	0	5	3	0	4	5	8
Complete Linkage	8	7	0	0	5	3	0	4	5	8

IV. CONCLUSION

The present study aimed to investigate the consumption model of power industry at the specified temperature range by focusing on the amount of load deficit and proposing some suggestions to adopt a solution for compensating this deficit. This study used an initial data set involving the parameters of load demand, power plant reserve, power import, and power export at the temperature range of -12 to 50 ° C. Since there are many fields from the mentioned set of parameters at a given temperature, it was first necessary to average the records for the specified temperatures. Thus, a record including the average of the above-mentioned parameters was created for each temperature. In this situation, the clustering algorithms could be applied on the data set more effectively. In this study, four clustering algorithms of K-means, Single Linkage, Average Linkage, and Complete Linkage were used to apply to the data set. Of these algorithms, Single Linkage, Average Linkage, and Complete Linkage algorithms were considered as hierarchical. Based on the findings of this study, the supply and distribution network faced load deficit and caused electrical breakdowns at high temperatures due to increased power demand. In this study, the final status of the load generation and distribution network was indicated in the form of clusters according to the seasons of the year. Based on the analyses conducted on the results obtained from the K-means algorithm, no load deficit occurred in the first three seasons and power could be exported while there was a lot of load deficit in summer when power export should be prevented and its generation and import should be added. In summer, when the temperature increases significantly, the power generation and distribution network status should be set according to the average temperature of 42 degrees centigrade and average deficit of 5581. In fact, the expected increase in generation should equal 4600, the expected increase in import should equal 400, and the expected increase in export should equal -580 to solve the problem of load deficit.

REFERENCES

[1] G. Chicco, R. Napoli and F. Piglione, "Comparisons among clustering techniques for electricity customer classification", IEEE Transactions on Power Systems, 21(2), 2006, pp. 933-940.

[2] M. S. ElNozahy, M.M.A. Salama and R. Seethapathy, "A probabilistic load modelling approach using clustering algorithms", in Power and Energy Society General Meeting (PES), pp. 1-5, July 2013.

[3] Z. Guo, K. Zhou, X. Zhang, S. Yang and Z. Shao, "Data mining based framework for exploring household electricity consumption patterns: A case study in China context", Journal of Cleaner Production, 2018.

[4] J. MacQueen, "Some methods for classification and analysis of multivariate observations" in Proceedings of the fifth Berkeley symposium on mathematical statistics and probability, vol. 1, no. 14, pp. 281-297, June 1967.

[5] F. McLoughlin, A. Duffy and M. Conlon, "A clustering approach to domestic electricity load profile characterisation using smart metering data", Applied energy, 141, 2015, pp. 190-199.

[6] W. A. Omran, M. Kazerani and M.M.A. Salama, "A Clustering-Based Method for Quantifying the Effects of Large On-Grid PV Systems", Power Delivery, IEEE Transactions on, vol. 25, 2010, pp. 2617-2625.

[7] I. P. Panapakidis, M.C. Alexiadis and G.K. Papagiannis, "Load profiling in the deregulated electricity markets: A review of the applications", In European Energy Market (EEM), 2012 9th International Conference, pp. 1-8, May 2012.

[8] I.P. Panapakidis, T.A. Papadopoulos, G.C. Christoforidis and G.K. Papagiannis, "Patternrecognition algorithms for electricity load curve analysis of buildings", Energy and Buildings, 73, 2014 pp. 137-145.

[9] B. K. Patra, S. Nandi and P. Viswanath, "A distance based clustering method for arbitrary shaped clusters in large datasets" Pattern Recognition, 44(12), 2011, pp. 2862-2870.

[10] S. Ramos, J. Duarte, F. Duarte and Z. Vale, "A Data-mining-based Methodology to support MV Electricity Customers' Characterization", Journal of Energy and Buildings, 91, 2015. 91. 10.1016/j.enbuild.2015.01.035.

[11] R. Xu and D.C. Wunsch, "Clustering", Hoboken. NJ: Wiley, IEEE Press, 6, 2009, pp. 583-617.

Topic #4

Big Data & Data Science Using Intelligent Approaches

IEEE Third International Conference on Data Stream Mining & Processing
August 21-25, 2020, Lviv, Ukraine

The Use of Social Engineering in Developing the Concept of "Smart Packaging"

Iryna Biskub
Lesya Ukrainka Eastern European Nationa University
Lutsk, Ukraine
0000−0001−5844−7524

Lyubov Krestyanpol
Lutsk National Technical University
Lutsk, Ukraine
0000−0003−3617−7900

***Abstract*—The article presents the process of developing the model of the information gathering system. The authors suggest the possibility of designing the product monitoring system working on the wireless data transmission methods. RFID technology is viewed as an instrument for gathering information about the condition of particular products. Three main stages of system functioning have been singled out. At the first stage, the programming and product marking-up is carried on. Corresponding product conditions and the systems for their coding are defined. The performance principles of the product monitoring system are described.The second stage presupposes defining the mechanisms of reading the coding system from RFID tags. At the third stage, the information is gathered from the customers. For this, the authors suggest to integrate into the developed system an additional element – a web framework. For the developed system, the authors have constructed a simulation model of gathering information from the customers, and the adequacy of the model has been checked experimentally. The developed system enables gathering and processing information in the real time mode, and the decision-making process concerning product utilization depends on the input data. Thus, the authors suggest the possibility of applying the developed system for resolving global food waste and food shortage issues.**

Keywords— simulation model, web framework, automated information system, smart packaging.

I. INTRODUCTION

Within the recent twenty years packaging recycling and disposal have one of the most often discussed topics. Therefore, new packaging methods, packaging recycling technologies and ecological ways of packaging emerge [5]. However, the humanity has always been concerned about the problem of the rational use of natural resources, which has become a global issue, and requires a careful scientific discussion [8]. According to the research of the influential international organizations (The United Nations, FAO, and UNEP), food waste has become the world's main source of the environmental pollution. Every year more than four billion tons of food products is produced by the food industries, one third of which turns into food waste (1.3 billion tons) [9]. Trying to resolve the problem at the global level, The UN developed a special "Save Food" program which aims at finding the e□ective ways of decreasing the level of food waste and food loss, the preservation and the rational use of the natural resources owned by the humanity, minimizing man's destructive impact on nature. Global prevention of malnutrition and starvation has long ago become an urgent topic, which is being constantly discussed by many international organizations. These facts predetermined the urgency of the present article, the authors of which have developed a concept of the "smart packaging" providing the relevant conditions for preserving food products and preventing food wastage [6, 7]. One of the basic elements of developing the concept of "smart packaging" is the creation of the specific information gathering system. Below we will consider three main stages of' developing such a system which include new technical solutions, as well as social and communicative aspects of processing information.

II. MATERIALS AND METHODS

At first stage, we will discuss the process of developing automated information system the using RFID technologies. It should be noted, that modern food manufacturing market is characterized by a strong tendency towards increasing food wastage. Food waste emerges even before food products reach the customers. The reasons for food losses and food waste are di□erent in di□erent corners of the globe. Food products manufacturing consists of a definite set of technological operations . This set presupposes some basic stages similar for all type of products. Food loss and wastage occur at all stages of the food supply chain.

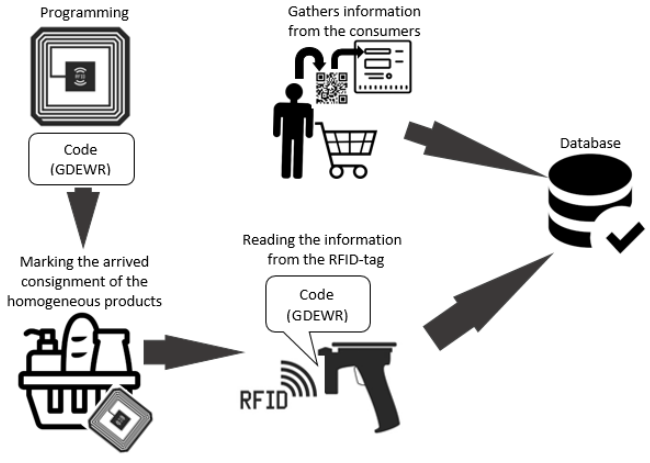


Fig.1.Sheme of "smart packaging" concept.

If we compare the percentage ratio of the reasons for food loss and the emerging of food waste, research has shown that food loss constitutes 20 and 30 per cent at the stages of retailing and consuming correspondingly [7]. One of the possible ways to reduce global food loss and waste is to develop new automated systems for monitoring all the stages of food producing, manufacturing, retailing and consuming.

In the present paper, we suggest developing a new concept of "smart packaging" integrated into a RFID system that will be permanently processing the input data from the consumers Figure 1. RFID (Radio Frequency Identification) is a wireless control system using electromagnetic induction as the main method of communication tranmitting data transmission. It di□ers from other known wireless technologies such as Wi-Fi, Bluetooth, 3G that are based on radio transmitters [1, 4, 11]. Object identification is performed according to the unique digital code recorded on a special electronic tag (RFID-tag) and is later read from it with the help of special RFID-tag readers. Radio frequency tags are attached to the identified object enabling the product quality control and making product movement in space possible [4, 13]. Standard RFID System consists of the RFID-tag reader with an antenna, and RFID-tags with the corresponding information. RFID-tag reader, in its turn, consists of the transceiver and an antenna sending and receiving back a signal from the RFID-tag, a microprocessor for data checking and decoding, and memory for data storage [15, 16].

The work of the Automated information system consists in three main stages where at which particular operations are performed. The first stage of the Automated information system presupposes programming and marking the arrived consignment of the homogeneous products by the specific tags. At this stage each product item is marked by the RFID-tag. The tag, in its turn, has the recordings of the three types of codes.

The second stage of the automated information system consists in reading the information from the RFID-tag on the required period of time.

The suggested identification method minimizes the time spent on checking products' expiration dates. It also simplifies the accounting procedures. The retailing administrators acquire the possibility of examining the product's condition just on time and modifying its further consumption procedure. Regardless of its obvious benefits, this method is only based on one-side control performed by the retailing administrators, whereas the consumer in not involved in the control procedure at all.

The violation of the shelf life neighborhood conditions, ignoring temperature, humidity, and other important criteria influence on the quality and product's expiry date. This will eventually lead to increasing the amount of food waste. Thus, we suggest involving the consumers to the process of the product's quality control.

This can be done using web frameworks for gathering information for the Automated Information Systems.

At the third stage of AIS's functioning, the system gathers information from the consumers. One of the e□ective ways of gathering data from the consumers is using web frameworks. All big retailing networks have their own websites, through which they provide the information for the consumers and do the online trading. Thus, information gathering can be done at the stage of storing the products within the retailing network, as well as at the stage of consumption in case of, in online trading.

In case of defect detection, damaged products, or other product discrepancies, the consumer may fill-in a special web framework to report the product's condition. Having visually examined the product in the trading room, the consumer may send the following types of information to the retailing administrator. Information about the product:

- lack of product;
- mechanically damaged goods;
- contaminated goods;
- the signs of product use;
- the signs the spoiled product.

Information abot product storage conditions:

- violation of the commodity neighborhood rules;
- violation of the storage conditions (temperature, humidity, direct sunlight);
- lack of price tags.

The retailing administrator may gather all the above-mentioned types of information using a web framework. The main purpose of using web frameworks is sending the information from the user to the mainframe computer. As a rule, the result of the information exchange is the new HTML-document generated by the mainframe for the user, based on the previously sent information [10, 17]. Web frameworks can be created using programming codes, Google forms, plugins [14, 2]. The user can fill in a web framework using websites or mobile applications.

III. Experiment

Developing a web framework and integrating it into the simulation model of AIS shown in Figure 2.

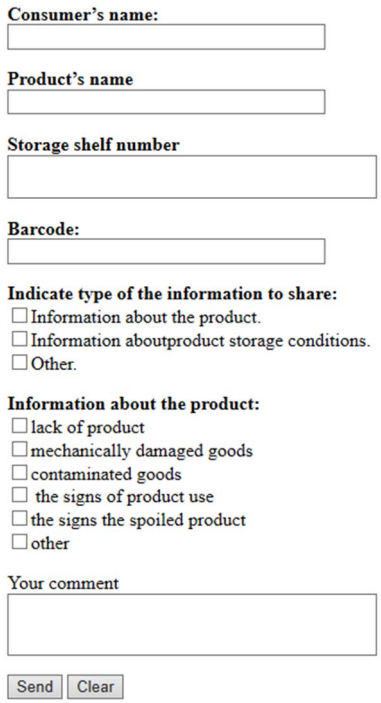


Fig.2. The developed web framework for gathering information in the AIS.

In order to specify the process of gathering information about the products, the authors of this paper have developed a simulation model of the process. Our simulation model defines the process of sending the information from the user (i.e. a consumer) to the retailing administrator. This model

makes it possible to prognose the behavior of the system as a whole. The suggested simulation model aims at simulating the process of visiting a store by the consumer, finding violations of product quality, and informing retailing administrators about them [12]. The construction of the model presupposes the following set of actions. Creating the plan of the trading room and virtual marking of the operation zones. Figure 3 presents a 2-D model of the trading room plan. The dotted line highlights and marks six (1-6) zones where the consumer is supposed to stay and perform some preconditioned algorithm of actions. These zones correspond to the shelvings with goods.

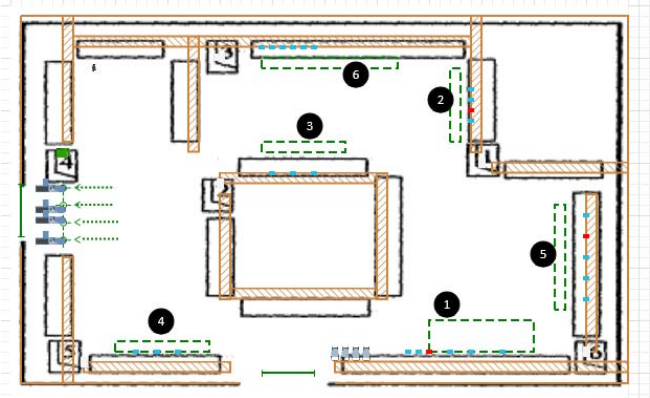


Fig.3. 2-D model of the trading room plan with the virtually marked zones.

The plan also indicates consumer's entrance and exit points. To make the model more realistic, we have added 3-D images of the cashiers at the exit, trading machines, and boxes with goods.

Having monitored customers' route schemes showing their movements between the zones of the suggested simulation model, we have clearly defined five route schemes of the customers and modelled five independent customer's flows. 2-D model of the trading room plan is shown in Figure 3. While analyzing the above presented simulation model, we have also designed customer's behavior in the zones marked by the deviations in the product conditions.

IV. Result and Discussion

The presented research resulted in the development of the concept of an automated information system for gathering information. The developed simulation model is based on modern technical requirements, as well as it takes into consideration the existing patterns of customers' social interactions.

Technical requirements include RFID-tags, RFID-tag readers, computer frameworks for processing and retrieving the input data, and some additional computer equipment for the system functioning.

While analyzing the specifications of the suggested simulation model, we have concluded that customer's behavior may be simulated by two stages depending on the indicated deviations in the condition of the goods. The customer may perform two possible actions regarding the particular specifications within the zone of his location:

1. The customer finds deviations in the product condition (changes in the storage conditions, signs of the spoiled products) and fills in the suggested web framework to inform the retailer about the deviations.
2. The customer ignores the web framework and follows the indicated route scheme. In this case he stays in the particular zone for about 1 minute.

The first option is, obviously, more desirable, since both sides – the customer and the retailer – will benefit from the indicated and reported deviations. Thus, another important aspect in developing our simulation model is finding the mechanisms of encouraging the customers to fill in the web framework.

For this, we suggest to apply some specific communication strategies to facilitate the communication between the customer and the retailer via web framework. In modern sphere of Human-Computer

Interaction, communication strategies predetermine the choice of language means to be represented in the web framework. The appropriate choice of the names of the fields will make web framework's interface user-friendly. A well-designed user-friendly interface will encourage the customer to inform the store administrators about the indicated problems.

Communication strategy used in the web framework's interface can be viewed as:

- global strategy (i.e. being polite, being positive, being honest);
- local strategy (i.e. greetings, request for the personal information, detailed description, appreciation of the customer's help).

All this has an important analytical and optimization potential for the development of "smart packaging" concept in order to essentially reduce food loss and food waste. In addition, the results of the application of the developed simulation model will make an important contribution into the creation of Decision Support Systems.

Conclusions

The article presents the analysis of a new information gathering system aimed at refining the concept of "smart packaging". The authors suggest an innovative approach to understanding the concept of "smart packaging" as an integral unity of modern technical solutions and human social responsibilities.

The paper highlights the possibility of using RFID technology of data transmission to support the performance of the information gathering system. The procedures of coding RFID tags and retrieving the received information are clearly defined and illustrated by the experimental data. The main benefit of the suggested simulation model is the possibility of using time ID-tags in monitoring product expiration date limits in the real-time mode. Just like any other information technology, the developed simulation model strongly depends on an appropriate input data.

In this paper, we have suggested the mechanism of modeling automatic communication between the customer and information gathering system by means of using specific

communication strategies. Therefore, the authors suggest using web frameworks to gather the requested information directly from the customers in the real-time mode. This will help preventing modern retailing networks from unnecessary food loss and food waste, which is a very disturbing issue in modern world. The suggested simulation model for "smart packaging" will enable companies, countries, cities to predict and quantify the unnecessary food loss, as well as to develop relevant strategies to benefit from "smart packaging" technologies.

REFERENCES

[1] A.V. Butrin, "Use of RFID technology in the warehouse of temporary storage", Actual problems of aviation and cosmonautics, vol 2 (11), 2015, pp. 847–849.

[2] I. Biskub, and L. Krestyanpol. Web Forms as a Tool for Automated Information Collection. Lutsk, 2020.

[3] Z. Cujan. Packaging technique and identification. Publish info: Prerov: Vysoka skola logistiky, 2012. ISBN 978-80-87179-18-52

[4] K. Finkenzeller. RFID handbook: radiofrequency identification fundamentals and applications. J. Wiley and Son, Ltd, 1999.

[5] O.M. Gavva, S.V. Tokarchuk, and O.O. Kokhan, "Smart Packing for Food", Packaging, vol. 2, 2013, pp. 36–40.

[6] L. Krestyanpol, "Economic feasibility of smart packaging under the global initiative "SAVE FOOD" ", Technological complexes, vol. 1(14), 2017, pp. 68–74.

[7] L. Krestyanpol, "The developing of "smart packagind". Technological complexes, vol. 1(13), 2016, 70–74

[8] 8. I. Miroshnik. Why SAVE FOOD is important for Ukraine. [Electronic resource]. Access mode: http://savefood-ua.com/blog/92/

[9] O□cial website of the Food and Agriculture Organization of the United Nations. [Electronic resource]. Access mode: http://www.fao.org/savefood/resources/keyfindings/infographics/fruit/en/

[10] A. Pavlenko. "UX review: Developing the right forms of information input". [Electronic resource]. Access mode:: https://medium.com/@pavljenko/ux-viewdevelopment-right-form-introduction-information-3844211d1e17

[11] B. Palchewskyi, O. Krestyanpol, and L. Krestyanpol. Information technologies in the design of the system of protection of packaged products. Lutsk, 2015.

[12] B. Palchewskyi, O. Krestyanpol, and L. Krestyanpol, "The use of Anylogic system for modelling a flexible automated packing system in training engineering students". Information Technologies and Learning Tools, vol. 75(1), 2020, pp. 225–236. https://doi.org/10.33407/itlt.v75i1.2714

[13] T.V. Plugina, and O.L. Reut, "Design models of radio frequency identification systems", Bulletin of Kharkov National Automobile and Highway University, vol. 55, 2011, pp. 23–28.

[14] O. Romaniuk. Web Design and Computer Graphics: A Textbook for Students of Software Engineering in All Specialties. Vinnytsia National Technical University, 2007.

[15] T. Scharfeldt. Low cost RFID systems. Moscow, 2006.

[16] C. Scholz, S. Doerfel, M. Atzmueller, A. Hotho, and G. Stumme, "Resource-aware on-line RFID localization using proximity data", Machine Learning and Knowledge Discovery in Databases, vol. 6913, 2011, pp. 129–144. https://doi.org/10.1007/978-3-64223808-69

[17] Y.A. Zavyalets. Web-technologies and web-design. Chernivtsi, 2014.

IEEE Third International Conference on Data Stream Mining & Processing
August 21-25, 2020, Lviv, Ukraine

Noisy Multiple-Control Fredkin Gate in Nuclear Spin Based Qubits Chain

Vitaly Deibuk
Computer Systems and Networks Dept.
Yuriy Fedkovych Chernivtsi National University
Chernivtsi, Ukraine
v.deibuk@chnu.edu.ua

Ivan Yuriychuk
Physics of Semiconductors and Nanostructures Dept.
Yuriy Fedkovych Chernivtsi National University
Chernivtsi, Ukraine
i.yuriychuk@chnu.edu.ua

***Abstract*— The Fredkin gate is a reversible self-inverse logic gate that maintains parity, so it is widely used in DNA computing, low-power CMOS, quantum computing, and nanotechnology. The paper proposes a design of the multiple-control Fredkin gate based on the model of a chain of atoms with nuclear spins one-half in a spinless semiconductor matrix. The allowed transitions realizing the gate operation for two π-pulses are analyzed. The effect of frequency noise and decoherence on the fidelity of the multiple-control Fredkin gate are discussed.**

Keywords—quantum computing, multiple-control, Fredkin gate, Toffoli gate, fidelity

I. INTRODUCTION

Processing of quantum information due to significant advances in modern nanotechnology has recently been moving from a purely theoretical field to the field of practical use [1]. The IBM Q project [2], which allows a wide set of users to have cloud access to a 50-qubit quantum processor is a confirmation of this. At the same time, quantum computing, low-power CMOS devices, bioinformatics, and optical computing require both efficient algorithms and fault-tolerant basic logic elements. However, system noise significantly limits the range of elements that can be used for reliable computations. As was shown in [3], the use of reversible logic in digital circuits design leads to zero power dissipation. It is the reversible logic that allows us to come closer to solving these problems through the use of the universal reversible gates, in particular, the Toffoli and Fredkin gates [4]. Many different schemes have been proposed for their implementation on the basis of elementary one- and two-qubit controlled gates (NOT, CNOT, CV, CV+ etc.), whose quantum cost is taken to be unity [5-8]. However, the attempts to build reversible logic circuits, which is optimal in the number of primitives (quantum cost) continue. In particular, the quantum cost of the Toffoli and Fredkin gates is known to be 5 [7]. The three-qubit Toffoli gate consists of two control inputs, which are transmitted without changes to the output, and one information input signal, which is added by modulo two to the product of the control signals and appears at the output (Control-Control-NOT). The Fredkin gate also has three input qubits, one of which is control and is transmitted to the output without changing. Two other input lines, depending on the size of the control qubit, can exchange information with each other or are transmitted to the output without changes (Control-SWAP). Along with the functional completeness of the Fredkin gate, unlike the Toffoli gate, it maintains parity, that is, the sum by modulo two of the inputs signals is equal to the sum of the output signals, which is an important advantage.

Despite a large number of methods [9] for the synthesis of reversible circuits, their practical implementation is significantly dependent on the quantum processor manufacturing technologies. Semiconductor quantum processors with a nuclear spin chain in a spinless matrix are a promising version of the NMR quantum computers, due to the large coherence times, relative simplicity in qubit manipulation, and the ability to scale [10,11]. In this case, the allowed atomic transitions are carried out under action of a sequence of the control pulses. Experimental implementation of the simplest three-qubit gates [12] showed that the number of pulses in such a sequence corresponds to the complexity of the corresponding computing device. In this case, the number of control pulses is a characteristic of the quantum cost of the circuit, and these quantities are not always consistent with each other. An important challenge in building a scalable quantum computer on the base of a spin qubits chain is an accuracy of the implementation of the reversible lossless quantum gates. In particular, it is an optimal choice of quantum elements of the reversible logic, both in terms of their physical implementation and functional versatility. In our opinion, such gates may be the multiple-control Toffoli and Fredkin gates, which together with the one-qubit Hadamar gate form a universal basis [10]. In addition, these gates are important components of the quantum error correction schemes [13]. The correct practical implementation of the reversible quantum gates can also be estimated by the fidelity of the operations [14], which is caused by various technological factors (an inaccuracy in tuning of the magnetic fields in amplitude and frequency, the number of control pulses, the nature of the interaction between qubits), and decoherence, which, in particular, can lead to broadening of the system energy levels, etc.

Recent advances in quantum computing show that machine learning can be a very fruitful field of application [18]. Since machine learning is usually reduced to a form of multivariable optimization, it can be effective for quantum computing. Quantum equipment designed for machine learning can become a reality much faster than a general-purpose quantum computer. In this paper, we will focus on the issues of correct operation of quantum devices under noise conditions, associated with the actual operation of control devices. Obviously, for such high-precision systems, correct data processing is impossible without their proper

978-1-7281-3214-3/20/$31.00 ©2020 IEEE

processing at the physical level. The main purpose of the paper is to study the effect of noise on the correct operation of the multiple-control Fredkin gate. We discuss the implementation of the multiple-control Fredkin gate in the model of a nuclear spin chain in a spin-less semiconductor matrix. Comparison of the number of the required transitions, which realizes correct operation of the gate, on the number of the control qubits is performed.

II. System Spectrum and Dynamics

First, let's define the reversible multiple-control Fredkin gate at the logical level. The sets of the multiple input signals are specified as $C = \{x_1, x_2, \ldots, x_{N-2}\}$ and $T = \{x_{N-1}, x_N\}$, moreover $C \cap T = \varnothing$. The Fredkin gate is mapped them onto their own, namely, the corresponding tuples of the input signals $(x_1, x_2, \ldots, x_{N-2})$ at the output remain unchanged, while the tuples (x_{N-1}, x_N) of the set T are mapped on the output in (x_{N-1}, x_N) if and only if (iff) the product of the input variables in C is unity, otherwise they also remain unchanged. The set C is called the set of control signals and the set T is the set of the target signals. Thus, the multiple-control Fredkin gate implements the $N \times N$ bijective logic function. Such gates are called reversible because the bijectivity does not lead to the loss of the input information and therefore to the power dissipation [3].

As the physical model, we use the quantum-mechanical Ising model for a system of the N interacting nuclear spins one-half, which are linearly spaced along the x-axis and are under the action of a strong magnetic field directed along the z-axis and a control transverse radio-frequency (RF) field with a circular polarization in the (xOy) plane. The Hamiltonian of such system corresponding to the full energy operator can be written in the form:

$$\hat{H} = \hat{H}_0 + \hat{W}\,, \tag{1}$$

where

$$\hat{H}_0 = -\hbar\left(\sum_{k=0}^{N-1}\omega_k I_k^z + 2J\sum_{k=0}^{N-2} I_k^z I_{k+1}^z + 2J'\sum_{k=0}^{N-3} I_k^z I_{k+2}^z\right), \tag{2}$$

$$\hat{W} = -\frac{\hbar\Omega}{2}\sum_{k=0}^{N-1}\left[I_k^+ \exp(i\omega t) + I_k^- \exp(-i\omega t)\right]. \tag{3}$$

We use the following notations in (1) – (3): I_k^z is the spin projection operator of the k-th nucleus on the z-axis; $\omega_k = \gamma B(x_k)$ is the Larmore frequency; γ is the proton gyromagnetic ratio; J and J' are the coupling constants between the nearest and second-nearest neighbors exchange interactions. The magnetic induction vector can be written as $\boldsymbol{B} = (b\cos\omega t, -b\sin\omega t, B(x_k))$; $\Omega = \gamma b$ is the Rabi frequency corresponding to a circularly polarized transverse field with the amplitude b. The latter can be considered as a perturbation since it is much smaller than the static magnetic field $B(x_k)$ and can be used to control the direction of the spin. This process can be described by the ascent and descent operators $I_k^+ = |0_k\rangle\langle 1_k|$ and $I_k^- = |1_k\rangle\langle 0_k|$ for the k-th spin [15]. The energy states of the (2^N) spin system were found from the solution of the stationary Schrödinger equation on the basis of the nuclear spins eigenstates $|i_{N-1}\ldots i_2 i_1 i_0>$, which equal to 0 or 1 and represent the content of each digit, depending on the spin orientation of each qubit. The transitions between the obtained energy levels under the action of the external control field with the frequency ω, correspond to the operation algorithm of the logical element and can be described by the non-stationary, i.e. time-dependent, Schrödinger equation:

$$i\hbar\frac{\partial\Psi(t)}{\partial t} = \hat{H}\Psi(t) \tag{4}$$

The complete wave function $\Psi(t)$ is presented in the full basis $|i_{N-1}\ldots i_2 i_1 i_0>$. The technique of finding the numerical solution of (4) is described in detail in [15].

III. Results and Discussion

Implantation of the ^{31}P ions with $|\uparrow\rangle$ and $|\downarrow\rangle$ nuclear spins states into a spinless isotope-enriched ^{28}Si semiconductor matrix revealed that such spin qubits have a record-breaking decoherence time exceeding 30 seconds, and fidelities of the single-qubit gates reach 99% [16]. This intriguing fact stipulates a modeling of more complex reversible fault-tolerant multiple-control Fredkin gate on the base of the above mentioned simple model. In particular, the following parameters were selected: the Larmor frequencies of the qubits varied as $\omega_k = 100\cdot 2^k$, (k = 0, 1, …, N-1). This is achieved by changing the magnitude of the magnetic field gradient $B(x_k)$ along the x-axis. The parameters of the exchange interaction J and J' between the first and second neighbors were chosen to satisfy the condition of the selective excitation ($\Omega << J << \omega$). The energy diagram of the simulated system together with the allowed transitions realizing the four-qubit Fredkin gate is presented in Fig. 1. Here the first two qubits are control and the last two qubits are target.

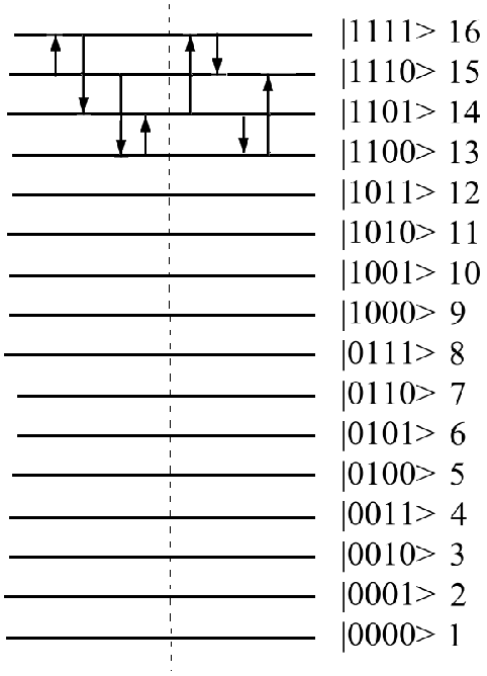


Fig. 1. Energy levels and allowed transitions for the four-qubit Fredkin gate

There are only two transitions that implement the four-qubit Fredkin gate, namely:

$$15 \rightarrow 14 \quad \text{and} \quad 14 \rightarrow 15.$$

However, according to the Pauli exclusion principle, such transitions are forbidden, so their implementation is possible through some intermediate levels. As can be seen from Fig. 1, the following two-step transitions implemented under the action of two π-pulses of duration π/Ω are permitted when tuning the RF control magnetic field into the resonance:

$$15|1110\rangle \rightarrow 16|1111\rangle \rightarrow 14\ |1101\rangle,$$
$$15|1110\rangle \rightarrow 13|1100\rangle \rightarrow 14|1101\rangle,$$

and

$$14|1101\rangle \rightarrow 16|1111\rangle \rightarrow 15|1110\rangle,$$
$$14|1101\rangle \rightarrow 13|1100\rangle \rightarrow 15|1110\rangle.$$

Increasing the number of control qubits does not change the number of the π-pulses which provide algorithmic gate transitions. This effect is also valid for the multiple-control Toffoli gate [17]. Along with pure digital states, the quantum Fredkin gate also implements transitions between superposition states. However, both the former and the latter are negatively affected by the frequency noise, due to an inaccurate tuning of the radio frequency pulse generator both in the frequency and in the phase. These factors will inevitably affect the gate fidelity. In particular, the fidelity value in optical systems of about 68% for the Fredkin gate has recently been experimentally achieved [10].

A. Effect of frequency noise

Firstly, consider the negative effect of frequency noise on the correct operation of the multiple-control Fredkin gate, depending on the number of control signals. Such noise may be caused by an inaccuracy in tuning the frequency ω of the RF generator, which is different from the resonant frequency ω_{lm} of the transition:

$$\omega = \omega_{lm}(1+\eta), \qquad (5)$$

where η is the frequency imbalance parameter. We can quantitatively describe the correctness of the Fredkin gate quantum algorithm using the fidelity, which we define as

$$F = \left\langle \Psi(t) \middle| \Psi_0(t) \right\rangle. \qquad (6)$$

Here, $\Psi(t)$ and $\Psi_0(t)$ are the proper wave functions obtained from the solution of the equation (4) at frequency ω and ω_{lm}, respectively. In our interpretation, the value of $|F|^2$ is the probability of correct operation of the gate. In Fig. 2 we present the fidelity as a function of the parameter η for the 15→14 transition (15→16→14 and 15→13→14 two-step transitions) in the four-qubit Fredkin gate. For the digital states, the fidelity decreases sharply with increasing η (Fig. 2), with the maximum imbalance corresponding to $F = 0.9$, which is independent of both the number of the control qubits of the Fredkin gate and the type of the transition (see Table 1). At the same time, the calculation for the uniformly filled superposition states with the probability 1/16 gives a much more complicated dependence $F(\eta)$. In particular, the maximum value of the parameter η, which corresponds to the correct operation of the gate, significantly increases and depends on the type of the algorithmic transition (Fig. 2).

N	IMBALANCE OF THE RESONANCE FREQUENCY η, RELATIVE UNITS		MODULATION OF RESONANCE FREQUENCY δ, □П MHZ	
	DIGITAL STATES	*SUPERPOSITION STATES*	*DIGITAL STATES*	*SUPERPOSITION STATES*
3	$\|\eta\| \le 2.63 \cdot 10^{-4}$	$\|\eta\| \le 3.52 \cdot 10^{-4}$	$\delta \le 5.68 \cdot 10^{-4}$	$\delta \le 6.87 \cdot 10^{-4}$
4	$\|\eta\| \le 2.65 \cdot 10^{-4}$	$\|\eta\| \le 3.54 \cdot 10^{-4}$	$\delta \le 1.33 \cdot 10^{-4}$	$\delta \le 1.79 \cdot 10^{-4}$
5	$\|\eta\| \le 2.57 \cdot 10^{-4}$	$\|\eta\| \le 3.48 \cdot 10^{-4}$	$\delta \le 6.78 \cdot 10^{-5}$	$\delta \le 7.56 \cdot 10^{-5}$

TABLE I. CORRECTNESS OF THE OPERATION OF THE GENERALIZED FREDKIN GATE WITH THE NUMBER OF QUBITS *N*

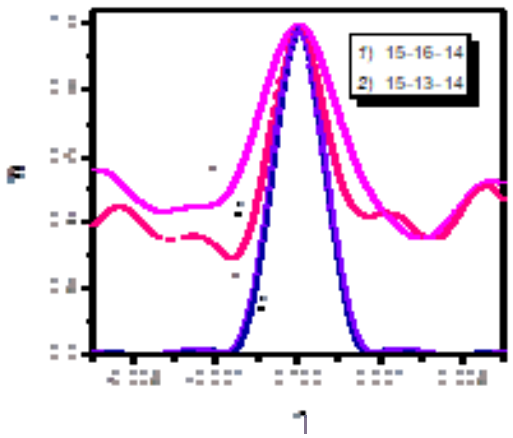


Fig 2. Fidelity as a function of the relative error η for the 15→14 transition in the four-qubit Fredkin gate. Superposition states: magenta and red color curves; digital states: violet and blue color curves

Another type of frequency noise is small periodic frequency oscillations over time that can be described as

$$\omega = \omega_{lm}\cos(\delta t), \qquad (7)$$

where the parameter δ characterizes the time deviation of the frequency of the RF generator relative to the resonant frequency of the transition which implements the Fredkin gate. Analysis of the fidelity dependence for this type of noise on the correct operation of the gate makes it possible to state that, as in the previous case for the digital states, $F(\delta)$ decreases rather sharply with increasing of the parameter δ (Fig. 3) and does not depend on the type of the transition. For the superposition states, an increase in the possible values of the parameter δ for correct operation ($F \ge 0.9$) was found (Table 1). As follows from Table 1, an increase in the number of control qubits for the generalized

Fredkin gate leads to a decrease in the frequency imbalance parameter δ of the RF control field. From the obtained dependencies $F(\delta)$ and $F(\eta)$ it is found, that the two-step transition 15→16→14 is preferable compared to the 15→13→14 transition because it has higher values of the fidelity (see Fig 2, 3).

In order to take into account an inaccuracy in tuning the frequency of the RF generator it is necessary to choose the optimal values of the imbalance parameters. In Fig. 4 we present the dependence of the imbalance parameter η on the dimensionless parameter ω_0/J, which allows us to choose the best values of both the magnitude of the magnetic field and the parameter of the exchange interaction. The choice of the latter is determined by the physical properties of the impurity atom (qubit), the distance between qubits, etc. [17].

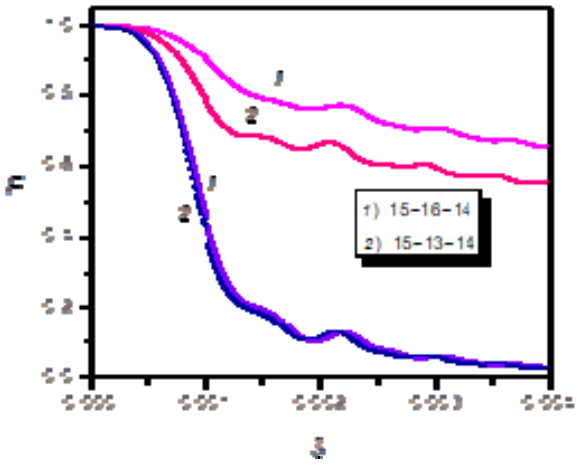


Fig 3. Fidelity as a function of the parameter δ for the 15→14 transition in the four-qubit Fredkin gate. Superposition states: magenta and red color curves; digital states: violet and blue color curves

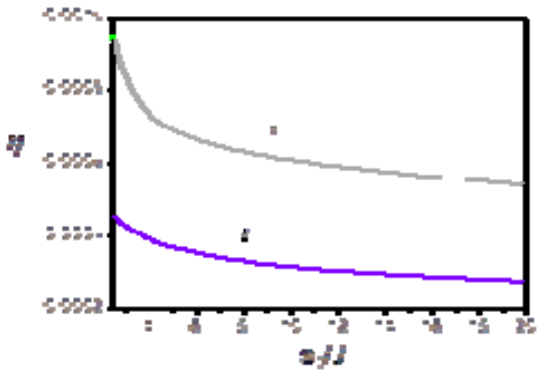


Fig. 4. Dependence of the imbalance parameter η on the ratio ω_0/J for the 15→16→14 two-step transition in the four-qubit Fredkin gate. 1 – superposition states; 2 – digital states

In particular, for the correct operation of the four-qubit Fredkin gate on pure digital states, the imbalance parameter $\eta = 2.65\cdot10^{-4}$ corresponds to $\omega_0/J = 20$, while for the superposition states $\eta = 3.54\cdot10^{-4}$ corresponds to $\omega_0/J = 8$. The value of ω_0/J obtained for the four-qubit gate agrees with the estimates [14], where it is ~ 23.

B. Effect of decoherence

Consider the effect of the decoherence phenomenon on the Fredkin gate operation. This phenomenon means that dissipative interaction of the qubits with an external environment leads to energy losses. As a result, being in any of the excited states, the system attempts to return to the state with the lowest energy, which corresponds to the logical zeros in all digits of the quantum register. Thus, information losses can occur already at the stage of recording the input data in the quantum register. We will describe the decoherence by broadening of the energy levels using the ratios:

$$\Gamma_0 = 0, \quad \Gamma_k = -i\beta\left(E_k - E_{k-1}\right), \quad k = 1...16. \qquad (8)$$

An imaginary unit with a negative sign characterizes the fact that, in the presence of a dissipative interaction with the environment the considered quantum bits system may be in any of the excited states over a finite period of time, and not for an indefinitely long time, as in the previous cases. A typical dependence of the probability of the correct operation of the Fredkin gate on the broadening parameter β is shown in Fig. 5. The probability of a correct answer is reduced to 0.5 at $\beta \approx (5 \div 6)\cdot10^{-5}$. In this case, the minimum distance between the levels of the system (Fig. 1) is Δν = 94.492 MHz. If the broadening is thermal in nature then its critical values correspond to the temperature $T = 5\cdot10^{-5} h\Delta\nu/k = 2.27\cdot10^{-7}$ K.

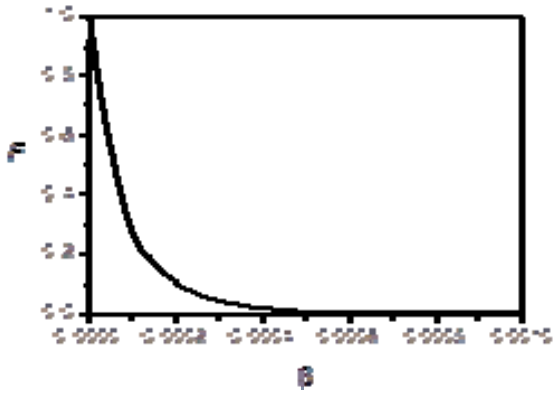


Fig 5. Fidelity as the function of the relative broadening parameter β

In the quantum mechanical approximation at low temperatures, for example on the base of the Debye formula, we obtain that $T = \pi^{-1}\sqrt[4]{25\cdot10^{-5}\, h\Delta\nu T_D^3/3k} = 1.01$ K, where the Debye temperature is $T_D = 645$ K (for the silicon matrix according to [17]). As follows, the Fredkin gate can work correctly only at sufficiently low temperatures.

IV. Conclusion

In the present paper, we present the design of the multiple-control Fredkin gate in the model of a nuclear spin chain in a spin-less semiconductor matrix. The comparison of the number of the required transitions, which realizes the correct operation of the gate, on the number of the control qubits, is performed. The allowed transitions, which realize the gate operation for two π-pulses are determined. The analysis of the frequency noise associated with an inaccuracy

in tuning the RF field generator suggests that the two-step 15→16→14 transition is preferable.

REFERENCES

[1] J. Preskill, "Quantum Computing in the NISQ era and beyond," Quantum, vol. 2, pp. 79, 2018.

[2] "IBM QX backend information." https://github.com/QISKit/ibmqx-backend-information.

[3] C. H. Bennett, "Logical reversibility of computation," IBM J. Res. Develop., vol. 17, pp. 525–532, 1973.

[4] P.D.Picton, "Modified Fredkin gates in logic design," Microelectron. J., vol. 25, pp.437-441, 1994.

[5] D.Maslov, G.V.Dueck, D.M.Miller, "Synthesis of Fredkin-Toffoli reversible networks," IEEE Transactions on VLSI Systems, vol. 13, pp.765-769, 2005.

[6] O. Golubitsky and D. Maslov, "A study of optimal 4-bit reversible Toffoli circuits and their synthesis," IEEE Transactions on Computers, vol. 61, pp. 1341-1353, 2012.

[7] M. Saeedi and I.L. Markov, "Synthesis and optimization of reversible circuits—a survey," ACM Comput. Surv., vol. 45, article 21(34 pages), 2013.

[8] Md B. Ali, T. Hirayama, K. Yamanaka, Y. Nishitani, "Function design for minimum multiple-control Toffoli circuits of reversible adder/subtractor blocks and arithmetic logic units," IEICE Trans. Fundamentals, vol. E101–A, pp. 2231-2243, 2018.

[9] E. Testa, M. Soeken, L. G. Amar and G. De Micheli, "Logic synthesis for established and emerging computing," in Proceedings of the IEEE, vol. 107, no. 1, 2019, pp. 165-184.

[10] R. B. Patel, J. Ho, F. Ferreyrol, T.C. Ralph, G.J. Pryde, "A quantum Fredkin gate," Science Advances, vol. 2, e1501531, 2016.

[11] QiuBo Fan, "Implementation of the Fredkin gate with a three-qubit mixed-spin Heisenberg model," Sci. China Phys. Mec. Astron., vol. 53, pp.1276-1280, 2010.

[12] F. Xue, J.-F. Du, M.-J. Shi et al., "Realization of the Fredkin gate by three transition pulses in a nuclear magnetic resonance quantum information processor," Chin. Phys. Lett., vol.19, pp. 1048-1050, 2002.

[13] A. K. Singh, H. M. Gaur and U. Ghanekar, "Fault detection in multiple controlled Fredkin circuits," IET Circuits, Devices & Systems, vol. 13, pp. 723-729, 2019.

[14] V. M. Stojanović, "Feasibility of single-shot realizations of conditional three-qubit gates in exchange-coupled qubit arrays with local control," Phys. Rev. A, vol. 99, 012345, 2019.

[15] O. I. Rozhdov, I. M. Yuriychuk, and V. G. Deibuk, "Building a generalized Peres gate with multiple control signals," Advances in Intelligent Systems and Computing, vol. 754, pp. 155-164, 2019.

[16] Morello A, Tosi G, Mohiyaddin FA, Schmitt V, Mourik V, Botzem T et al., "Scalable quantum computing with ion-implanted dopant atoms in Silicon," in IEEE International Electron Devices Meeting, IEDM 2018, pp. 6.2.1-6.2.4. 8614498, 2019.

[17] I. M. Yuriychuk, Z. Hu, and V. G. Deibuk, "Effect of the Noise on Generalized Peres Gate Operation," Advances in Intelligent Systems and Computing, vol. 938, pp. 428-437, 2020.

[18] J Biamonte, P Wittek, N Pancotti, P Rebentrost, "Quantum machine learning," Nature, vol.549, pp. 195-202, 2017.

IEEE Third International Conference on Data Stream Mining & Processing
August 21-25, 2020, Lviv, Ukraine

Mobile Decision Support System To Take Into Account Qualitative Estimation By The Criteria

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

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

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

Abstract— **The paper is devoted to the improvement of the mobile-based decision support system (DSS) based on the modified heuristic Smart method for Android platform for solving the multiple criteria decision-making (MCDM) problems in which the criteria's estimates can be either quantitative or qualitative in nature and display the presence or absence of relevant characteristics. The three- layer architecture was used when developing this mobile-based DSS for reuse of code: presentation level, application level and data level. Changes were made in the decision-making subsystem at the application level. It was proposed the modified heuristic Smart method improvement to take into account the criteria whose estimates can be either quantitative or qualitative character and display the presence and absence of relevant characteristics. The implementation of the improved modified heuristic Smart method is described. The developed improved mobile decision support system was applied to the task of choice by the students the courses for their inclusion in the educational process. A criterion as the relevance of skills and knowledge that students will receive as a result of studying course was taken into account. It allowed students to receive support in the choice of courses offline, using mobile devices, at any convenient time, to include courses in the individual training plan.**

Keywords—decision support system, Smart, Android, the alternative, the criteria

I. INTRODUCTION

The task of choice arises in any human activity every day. During decision A person identifies and takes into account the various characteristics of the choices (alternatives). These characteristics of alternatives (criteria) can be quantitative and / or qualitative in nature (good, bad, satisfactory). Also, the presence or absence of relevant characteristics can be taken into account, when solving the task of choice [1]. The models that describe the preferences of decision-makers are constructed when, the number of criteria by which alternatives are evaluated, are increased and when the estimates of the alternatives by these criteria are contradictory. The best choice is made, based on these models. The methods MAUT (Multi-Attribute Utility Theory), AHP (Analytic Hierarchy Process), Electre (ELimination Et Choix Traduisant la REalité), Topsis, PROMETHEE, VIKOR, ANP [2] are the most famous methods for constructing such models. These methods are used to solve the multiple criteria decision-making problems in various application areas. So the problem of choosing mobile phones using the AHP and Topsis methods is solved in [3], the AHP, ELECTRE, and TOPSIS methods are used for evaluation of competitive ability of lighting equipment manufacturers in [4]. The method AHP is used for evaluation of competitive ability of lighting equipment manufacturers in [5]. In addition, these methods are used in many DSS, the main task of which is to support the decision- makers in solving the multiple criteria problems. The problem creation of DSS, which based on the modified AHP method, is considered in [6]. The authors considered the task of creating a user interface. This interface can be implemented in different types of DSS, such as desktop DSS, Web-DSS, and mobile-based DSS later. The DSS of a certain type are used, depending of the tasks and the conditions in which the decision- makers solve them. The compactness of the mobile devices, the ability to support in decision-making at any time and offline were highlighted in [7], as the advantages of the mobile-based DSS. The mathematical methods on which the mobile-based DSS based should be non-resource-intensive. Thus, the development of mobile-based DSS based on a non-resource-intensive mathematical method to take into account criteria which have quantitative and / or qualitative estimates and / or display the presence or absence of relevant characteristics is relevant.

II. PROBLEM DESCRIPTION

Set of m alternatives $A = \{A_i\}$, $i = \overline{1,m}$ and set of n criteria $C = \{C_j\}$ $j = \overline{1,n}$, which used to estimate alternatives are known when solving multiple criteria decision-making problems. Need to rank alternatives A_i, $i = \overline{1,m}$ by use generalized weights of alternatives calculated on the set of criteria C [8]. The construct the decision matrix, which contains estimates of alternatives by the criteria [2], is one of the steps in many mathematical methods of MCDM, which are used in DSS for solving the multiple criteria problems. Elements of the decision matrix are used to count the general estimation for each alternative. However, the use of these methods in DSS is limited if a qualitative estimates alternatives by criteria are used. There are two important problems during the comparison of alternatives: the qualitative nature of the criteria estimates; the criterion may reflect the presence or absence of relevant characteristics [1].

978-1-7281-3214-3/20/$31.00 ©2020 IEEE

One of the most important tasks of choice is the tasks related with obtaining higher education. As part of student-centered learning [9], it is proposed to provide freedom to students to choose courses for increase the effectiveness of students' learning. In addition, it is important the choice of courses, which give to students knowledge and skills, that meeting employer's needs [10]. The problems of building mobile-based DSS are considered for choosing higher education studies [11].

The problems of building mobile-based DSS on the Android platform based on the modified heuristic Smart method [12] was considered for solving multiple criteria problems in [7]. The calculations in this method are non-resource-intensive. This mobile-based DSS was development using three levels architecture: presentation level, application level and data level. A decision-making subsystem based on the modified heuristic Smart method was development at the application level. In addition, this mobile-based DSS is invariant to the subject area.

Thus, the aim of the work is to improve mobile-based DSS on the Android platform [7], namely, the decision-making subsystem by improving the modified heuristic Smart method to take into account criteria estimates that are qualitative in nature and / or display the presence or absence of relevant characteristics. This will reduce the complexity of evaluating and comparisons so data type for the decision-makers when solving multi-criteria problems which contain different data types.

To accomplish the aim, the following tasks have been set:

– to improve the modified heuristic Smart method to take into account criteria estimates that are qualitative in nature and / or display the presence or absence of relevant characteristics;

– to perform the user interface change in mobile-based DSS on the Android platform;

– to perform the decision-making subsystem change in mobile-based DSS on the Android platform;

– to experimentally research of the improved mobile-based DSS on the example of the task of choice by the students the courses for their inclusion in the educational process.

III. SOLVING THE PROBLEM

The algorithm of the modified heuristic Smart method [12], based of which the decision-making subsystem of mobile-based DSS was developed, included the following steps:

Step 1: Construct the decision matrix X, the elements of which are numerical values, based on the introduced alternatives, criteria, and estimates of alternatives by the criteria.

Step 2: Each the criterion is appointed weight, so that the most important is estimated at 100 points.

Step 3: The weights of the criteria are normalized.

Step 4: Select the action over the criteria - maximization or minimization. Depending on the selection, the elements of the decision matrix are normalized:

– for criteria that are maximized (Equation 1)

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

– where r_{ij} – normalized value of the decision matrix X element x_{ij},

– for criteria that are minimized (Equation 2)

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

Step 5: Determine the overall rating of each alternative.

Step 6: Rank the preference order.

The following improvement of the modified Smart method is proposed to take into account criteria estimates that are qualitative in nature and / or display the presence or absence of relevant characteristics.

Step 1: Construct the decision matrix X, evaluating alternatives by the criteria. The criteria's estimates can be either quantitative or qualitative in nature and display the presence or absence of relevant characteristics.

Step 2 and Step 3 do not change.

Step 4 does not change for criteria's estimates that are qualitative.

Step 5: The use of the E. Harrington's scale for evaluating criteria that are of a qualitative nature is proposed (Table I) [13]:

TABLE I. SCALE FOR MEASURING THE QUALITATIVE ESTIMATES

The qualitative estimate	**E. Harrington's scale**
very bad	0-0,19
bad	0,2 - 0,36
satisfactory	0,37 - 0,62
good	0,63 - 0,79
excellent	0,8 - 1

The Harrington's scale is a multi-interval discrete verbal-numerical scale that transforms linguistic estimates into quantitative ones in the range from 0 to 1. The Harrington's scale is universal and can be used to evaluate different quality indicators.

The expert evaluation method's estimates are proposed to use for evaluating the criteria that reflect the presence or absence of relevant characteristics (Table II) [13]:

TABLE II. SCALE FOR MEASURING THE ESTIMATE «YES-NO»

The estimate «yes-no»	**Numerical estimate**
yes (desirable)	0,67
no (not desirable)	0,33
yes (not desirable)	0,33
no (desirable)	0,67

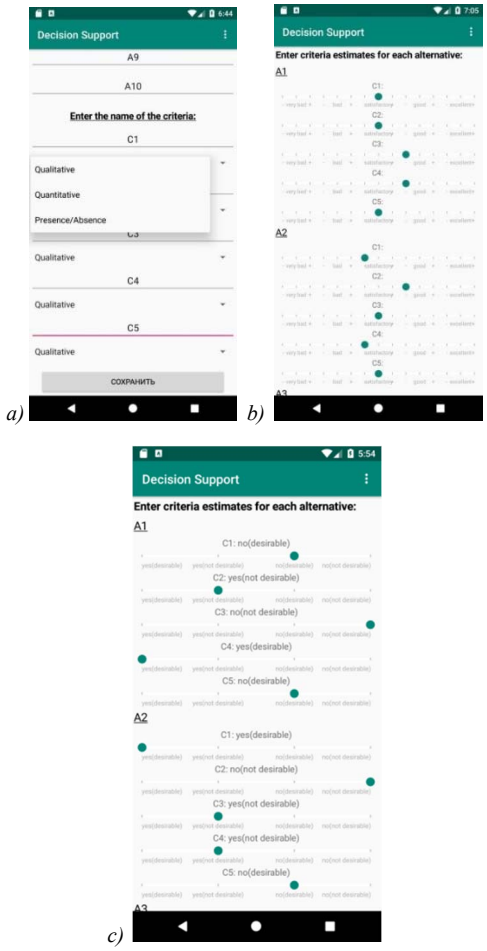
Step 6 and Step 7 do not change.

Step 8: Rank the preference order.

Thus, the improved modified Smart method allows you to process data that are quantitative or qualitative in nature, and also display the presence or absence of relevant characteristics.

IV. Description of Information System

Changes in the modified Smart heuristic method led to a change in the user interface of the mobile-based DSS, which consists of four windows that are displayed sequentially as a result of the user actions during the information processing (input, editing, deleting, and displaying the result). In the second window, the ability to select the nature of the input data (quantitative, qualitative, reflecting the presence or absence of relevant characteristics) is added (Fig. 1, a). In the third window, data depending on their type is entered. Qualitative data is entered using Seek Bar, which allows you to select ratings in the range from "very bad" to "excellent" (Fig. 1, b). Data, which display the presence or absence of relevant characteristics, is selected using Seek Bar (Fig. 1, c). The improved Smart method is implemented in the decision-making subsystem; the entered data is received in it after clicking the "Get a solution" button in the third window. The calculation results are saved automatically in the database and displayed in the fourth window, from which it is possible to go to the first window to solve a new task or load previously solved tasks from the database for viewing or editing data.



a) b) c)

Fig. 1. The user interface windows: *a* – the second window; *b,c* – the third window

The following changes and additions have been made to the presentation level program code where the user interface is implemented.

In the class Grid.java, which implements the second window interface, and in the class SecondActivityForDb Item.java, which implements a user interface that allows editing data displayed from the database in the second window an implementation of the graphic elements for choosing the data type has been added: quantitative, qualitative and data, showing the presence or absence of relevant characteristics.

In the class Thirdactivity.java, which implements the third window interface, and in the class ThirdactivityForDbItem.java, which implements a user interface that allows editing data displayed from the database in the third window an implementation of graphic elements has been added to enter qualitative data and data, showing the presence or absence of relevant characteristics.

The class info_from_list.java, which implements the ability to make changes in evaluating alternatives by the criteria, if necessary, has been changed.

Program code class Solve.java, which is implemented in the decision subsystem in mobile-based DSS on the Android platform, has been supplemented by the methods that allow to present the data entered in the third window using Seek Bar and Check Boxes as numerical data, and also take this data into account when obtaining a solution.

V. Experimental Research of the Application of Developed System

The task of choice by the students the courses for their inclusion in the educational process was considered. To solve this problem, it was necessary to determine the set of alternatives and criteria by which these alternatives were evaluated. Elective courses from the syllabus were seen as alternatives. Ten disciplines were considered, which we will designate as A_i, $i=\overline{1,10}$. Criteria were identified as a result of a survey of graduate students: Demand for knowledge. (criterion C_1), Pithiness of course materials (criterion C_2), Methodical level of material presentation (criterion C_3), The quality of methodical support course (criterion C_4), Relevance (criterion C_5). The input data are presented in Table 3.

TABLE III. The Qualitative Estimates of the Input Data

Name of course	C_1	C_2	C_3	C_4	C_5
A_1	Satisfactory	Satisfactory	Good -	Good -	Satisfactory
A_2	Satisfactory -	Good -	Satisfactory	Satisfactory –	Satisfactory
A_3	Satisfactory -	Good	Good	Good -	Satisfactory -
A_4	Good	Good +	Satisfactory	Excellent –	Satisfactory -
A_5	Satisfactory -	Satisfactory +	Good	Satisfactory -	Satisfactory
A_6	Satisfactory	Good	Good -	Satisfactory –	Satisfactory
A_7	Satisfactory	Satisfactory	Satisfactory	Satisfactory	Satisfactory
A_8	Satisfactory -	Good	Satisfactory +	Satisfactory –	Satisfactory

Name of course	C_1	C_2	C_3	C_4	C_5
A_9	Bad	Satisfactory	Satisfactory	Excellent -	Satisfactory -
A_{10}	Good	Good -	Satisfactory +	Satisfactory –	Satisfactory

The fourth window of the mobile-based DSS user interface, in which the results of the calculations are displayed, is shown in Figure 2.

As a result of applying the proposed improved method Smart, we obtained the following ranking of courses:

$$A_4 \succ A_{10} \succ A_1 \succ A_6 \succ A_3 \succ A_7 \succ A_2 \succ A_5 \succ A_8 \succ A_9 .$$

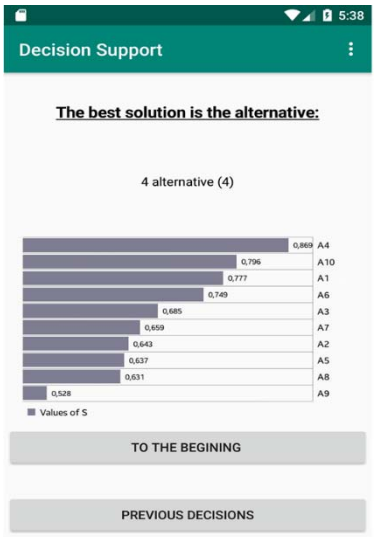


Fig. 2. The fourth window of the user interface

Also, the integrated assessment method [14] was applied to solve this task of choice by the students the courses for their inclusion in the educational process. The input data were presented as quantitative estimates according to table 1. We designate these estimates by P_{ij}, where i – the number of the criterion, j – the number of the alternative, and we write it in table 4.

At the first stage, the experts determined the importance of the criteria by direct assessment on a 10-point scale. 20 graduate students were as experts. Then, weights of the criteria were determined, for which the estimates mean values were calculated and normalized they. As a result, the following criteria weights were obtained: $V_1 = 0{,}213$; $V_2 = 0{,}187$; $V_3 = 0{,}209$; $V_4 = 0{,}184$; $V_5 = 0{,}207$.

TABLE IV. THE QUANTITATIVE ESTIMATES OF THE INPUT DATA

Name of course	C_1	C_2	C_3	C_4	C_5
A_1	0,56	0,5	0,644	0,63	0,56
A_2	0,44	0,63	0,475	0,37	0,455
A_3	0,4	0,66	0,63	0,63	0,395
A_4	0,68	0,75	0,495	0,7	0,42
A_5	0,43	0,63	0,5	0,36	0,428
A_6	0,52	0,73	0,63	0,37	0,45
A_7	0,42	0,65	0,5	0,5	0,459
A_8	0,41	0,63	0,5	0,37	0,468
A_9	0,26	0,515	0,343	0,8	0,41
A_{10}	0,68	0,645	0,5	0,3	0,459

At the second stage, weights of the criteria, which reflect scatter, were determined. For this were calculated:

– Average ratings for each criterion:

$$P_i = \frac{1}{N}\sum_{j=1}^{N} P_{ij}, \quad i = \overline{1,M} \quad , \tag{3}$$

where M – the number of the criterion, N – the number of the alternative.

– Scatter values on each criterion:

$$R_i = \frac{1}{N \cdot P_i}\sum_{j=1}^{N} \left| P_{ij} - P_i \right|, \quad i = \overline{1,M} . \tag{4}$$

– Sum of scatter values:

$$R = \sum_{i=1}^{M} R_i \, , i = \overline{1,M} . \tag{5}$$

- Weights of the criteria, which reflect scatter:

$$Z \; = \frac{R_i}{R} , i = \overline{1,M} \tag{6}$$

As a result, we received following values: $Z_1 = 0{,}37$; $Z_2 = 0{,}11$; $Z_3 = 0{,}16$; $Z_4 = 0{,}1$, $Z_5 = 0{,}27$.

At the third stage, generalized weights of the criteria were determined.

$$W_i = \frac{V_i + W_i}{2} , i = \overline{1,M} . \tag{7}$$

As a result, we received following values: $W_1 = 0{,}29$; $W_2 = 0{,}146$; $W_3 = 0{,}186$; $W_4 = 0{,}13$, $W_5 = 0{,}24$.

In the fourth stage, weighted scoring of alternatives were found (table 5):

$$E_{ij} = P_{ij} \cdot W_i , \quad i = \overline{1,M} , \; j = \overline{1,N} . \tag{8}$$

TABLE V. WEIGHTED SCORING OF ALTERNATIVES

Name of course	A_1	A_2	A_3	A_4	A_5	A_6	A_7	A_8	A_9	A_{10}
C_1	0,16	0,13	0,12	0,20	0,13	0,15	0,12	0,12	0,08	0,20
C_2	0,07	0,09	0,10	0,11	0,09	0,11	0,10	0,09	0,08	0,09
C_3	0,12	0,09	0,12	0,09	0,09	0,12	0,09	0,09	0,06	0,09
C_4	0,08	0,05	0,08	0,09	0,05	0,05	0,07	0,05	0,11	0,04
C_5	0,13	0,11	0,09	0,10	0,10	0,11	0,11	0,11	0,10	0,11

At the fifth stage, complex object estimates (sums of weighted scoring) were found:

$$E_j = \sum_{i=1}^{M} E_{ij} \, , \; j = \overline{1,N} . \tag{9}$$

As a result, we received following values: $E_1 = 0{,}576$; $E_2 = 0{,}469$; $E_3 = 0{,}511$; $E_4 = 0{,}596$; $E_5 = 0{,}463$, $E_6 = 0{,}534$; $E_7 = 0{,}489$; $E_8 = 0{,}468$; $E_9 = 0{,}421$; $E_{10} = 0{,}538$.

We obtained the following ranking of courses by the integrated assessment method: $A_4 \succ A_1 \succ A_{10} \succ A_6 \succ A_3 \succ A_7 \succ A_2 \succ A_5 \succ A_8 \succ A_9$.

Thus, as a result of the experiment, it was received that the course A_4 is the most recommended for study both by the improved modified Smart method, and by the integrated assessment method. The difference in ranking is only for alternatives A_1 and A_{10}. However, these alternatives are among the three best alternatives. Thus, mobile-based DSS on the Android platform which was improved by the improving the Smart method can be recommended for solving multi-criteria problems, in which criteria's estimates can be either quantitative or qualitative in nature and display the presence or absence of relevant characteristics. And, decision-makers can use this mobile-based DSS offline.

VI. CONCLUSION

In the work mobile-based DSS on the Android platform [7] was improved, namely the user interface and the decision-making subsystem. The decision-making subsystem based on the improved modified Smart method which was implemented and made possible to take into account evaluations of criteria that are qualitative in nature and / or reflect the presence or absence of relevant characteristics. This mobile-based DSS was applied to solve the multi-criteria problem of choice by the students the courses for their inclusion in the educational process.

Thus, the decision-maker received a tool that reduces the complexity of evaluating and comparisons data that are qualitative in nature and / or reflect the presence or absence of relevant characteristics. This allows decision-makers to solve multi-criteria problems that contain different data types. Thus, the area of tasks, in which the decision-maker can use this mobile-based DSS, has been expanded.

REFERENCES

[1] P.A. Gudkov “Methods for comparative analysis”, Penza: Izd-vo Penz. gos. un-ta, pp. 81, 2008.

[2] Udaykumar S. Kashid, Archana U. Kashid, Saurabh N. Mehta. A Review of Mathematical Multi-Criteria Decision Models with A case study// Proceeding of the ICESTMM 2019 (International Conference on Efficacy of Software Tools for Mathematical Modeling). Mumbai, India, April 19-20, pp. 1-15.

[3] Işiklar G., Büyüközkan G. Using a multi criteria decision making approach to evaluate mobile phone alternatives/ Gülfem Işıklar, Gülçin Büyüközkan // Computer Standards & Interfaces 29 (2007) pp. 265–274. doi:10.1016/j.csi.2006.05.002.

[4] Schinas, Orestis. "Examining the use and application of multi-criteria decision making techniques in safety assessment." International symposium on maritime safety, security and environmental protection, pp. 1-9, 2007. doi: 10.13140/2.1.4533.9841.

[5] I. A. Ivanova, E. A. Syisoeva, “Estimation of competitiveness of Russian enterprises of lighting based on the analytic hierarchy process”, Ekonomicheskiy analiz: teoriya i praktika, no. 26, pp. 47–53,2014.

[6] D.G. Furtsev, “About creation of decision support system which based on the modified analytic hierarchy process”, Nauchnye vedomosti BelGU. Serija: Jekonomika. Informatika, no. 7, pp. 125–131, 2015.

[7] Development of a mobile decision support system based on the Smart method for ANDROID platform/ D. Horpenko, N. Volkova, M. Polyakova, V. Krylov// Eastern-European Journal of Enterprise Technologies. – 2019. – Vol.3. – № 2(99), May. – p. 6–14. doi:http://doi.org/10.15587/1729-4061.2019.168163.

[8] N. D. Pankratova, N. I. Nedashkovskaya, “The hybrid method of multi-criteria evaluation of decision-making alternatives”, Kibernetika i sistemnyiy analiz, no. 5, pp. 58 – 70, 2014.

[9] A. Attard, E. Di Iorio, K. Geven, R. Santa.: “Student-centered learning. Toolkit for students, faculty and universities”. Instrumentariy dlya studentov, professorsko - prepodavatelskogo sostava i vuzov, Astana: NKAOKO-IQAA, pp. 64, 2017.

[10] Y.Y. Kozina, E.V. Verbitskaya, “Analysis of decision making methods to include a set of disciplines in the educational process”, Proceedings of the International Scientific Conference Intellectual Systems of Decision Making and Problems of Computational Intelligence ISDMCI'2019 HNTU, Ukraine, May 21-25, pp. 81–82, 2019.

[11] Kostoglou, V., Kafkas, K. Design and development of an interactive mobile-based decision support system for selecting higher education studies// Proceeding of the 8th Balkan Region Conference on Engineering and Business Education. Sibiu, Romania, 240–248. doi:https://doi.org/10.1515/cplbu-2017-0032

[12] Kozina, Y., Volkova, N., Horpenko, D. Mobile application for decision support In multi-criteria problems// Proceeding of the 2nd DSMP 2018 (Data Stream Mining&Processing). Lviv, Ukraine, August 21–25, pp. 56 – 59, 2018. doi:https://doi.org/ 10.1109/ DSMP.2018.8478499

[13] D. V. Petrov, E. V. Panina, “Development of alternative utility estimation method using linear functions”, Mezhdunarodnyiy nauchno-issledovatelskiy zhurnal, no. 5, pp. 136–142, 2019. doi:https://doi.org/10.23670/IRJ.2018.71.021.

[14] D. V. Petrov, E. V. Panina, “Comparison of Self-propelled Feed Mixer-distributors using Integrated Assessment Method”, Sustainable development of science and education, no. 12, pp. 197–203, 2017.

IEEE Third International Conference on Data Stream Mining & Processing
August 21-25, 2020, Lviv, Ukraine

Strategy of Construction of Intellectual Production Systems

Bogdan Palchevskyi
Eastern European Scientific Society
Lutsk, Ukraine
0000−0002−4000−4992

Lyubov Krestyanpol
Lutsk National Technical University
Lutsk, Ukraine
0000−0003−3617−7900

***Abstract*—An important issue to produce the technological equipment, including packaging, is the improvement of production automation systems, which involves improving the control and guidance of industrial equipment, providing its self-diagnosis. Technological equipment is used in many spheres of industry. Its malfunction leads to a significant decrease in the economic e□ciency of production. To eliminate material losses directly during the operation of the equipment, its proactive maintenance and repair should be carried out. To solve this problem, the methods of Big data analysis, artificial intelligence principles, digital technologies are applied more widely in the industry, that make it possible to create Digital Twin (Digital Twin) technological equipment and to use them for self-diagnose manufacturing systems. These Digital Twin model the internal processes, technical specifications and behavior of technological equipment under the conditions of influence of interference and the environment, that allows to perform data analysis to identify the sources of e□ciency loss. At the stage of operation, the Digital Twin model of the technological equipment can be used to provide feedback in order to adjust the diagnostics and forecasting of malfunctions and to increase the operating e□ciency of technological equipment. It is proposed the method of using digital models of the state of technological equipment, based on the measurement and comparison of the parameters that characterize the operability of the equipment, which contributes to the early detection of malfunctions and their elimination.**

Keywords — Digital twin, diagnosis, malfunction matrix, mathematical expectation, system analysis, additive evaluation indicators, parameter.

I. INTRODUCTION

The use of advanced information technologies at all levels of control allows us to proceed to the intellectualization of production systems. Any task, which has the unknown algorithm of a solution, can be attributed to the intellectual, which should be solved by means of applying of artificial intelligence. When solving intelligent tasks of the production, system operates without an exact algorithm for solving the problem.

Adaptation algorithms are being created for work in external conditions that change over time. The first intelligent Systems of Automatic Control (SAC), combining the methods of traditional systems of automatic control and engineering knowledge, became Expert Systems (ES). The simplest intelligent SAC can, for example, consist of a simple SAC and a base of productive rules. They allow you to change the parameters of the functioning of the production system or its structure.

Since during the formation of a control program, it is necessary to take into account possible production situations, then the intelligent subsystems of the SAC should compensate changing of the external conditions by making certain changes in the control algorithm to achieve the optimal characteristics of the technological complex functioning. Obviously, such a SAC should evaluate, above all, the external conditions in order to make the necessary changes in the algorithm of functioning. Therefore, intelligent SAC implements three control functions.

- Identification of the system consists in getting of an instant assessment of quality of the process or functioning of the machine by identifying a certain quality index that can be compared with its specified value.
- The decision is to find the direction of changing the control program in order to improve the quality of the process.
- Adjustment involves a physical or mechanical change in the control of the machine.

Functional scheme of the intelligent SAC is presented below.

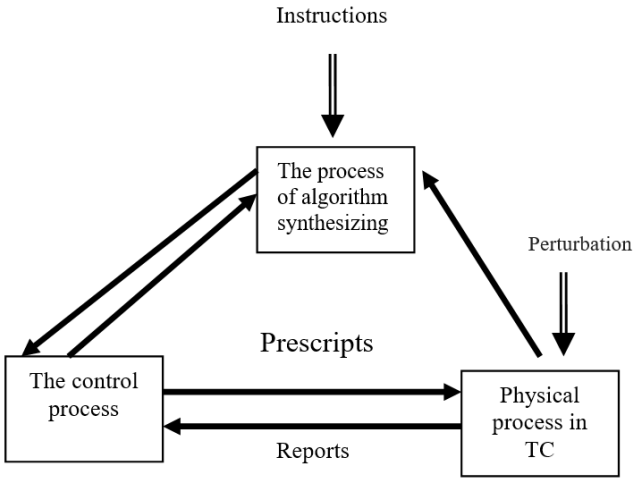


Fig. 1. Processes in Intellectual Technological Complex

The allocation of the technological complex to the executive system, the control system and the intelligent system will enable to tie the tasks, which should be performed by the TC, with the tasks of the control system and the intellectual system.

Recently, with the increasing use of artificial intelligence in the manufacturing systems, it has become possible to take production automation processes to significantly new level. It is the consolidation of integrated manufacturing systems with intelligent control systems has become the main feature of the

creation of modern intellectual manufacture, which has begun to develop in all spheres of industrial production. This development has been called by experts the fourth industrial revolution.

Typical changes in modern packaging manufacture include, first of all, the intellectualization of control systems for technological equipment. The consequence of this change is also the appearance of Intelligent maintenance system. One of the advanced types of manufacturing equipment maintenance is proactive diagnosis model. It allows repairing not according to a prearranged plan, but when there is a need for it. It means not the elimination, but the prevention of the equipment failures by means of interactive evaluation of its technical state based on totality of data from the sensors and determining the optimal timing of repairing work.

Proactive (predictive)service is applied in cases where the degree of mechanism use in the workflow is estimated as high, and its malfunction leads to the long idle time and significant loss of e□ciency. The e□ectiveness of proactive technical equipment service maintenance is achieved by means of:

- collecting data on the technical state of the equipment and their pre-processing,
- early detection of malfunctions,
- forecasting the time of failure.

For the practical implementation of intelligent maintenance, we use the Digital Twin concept of technological equipment, Big Data technology, which allow to predict the time of failure with high accuracy. The combination of these technologies in intelligent technological equipment makes it possible to realize modern advanced methods of technological service.

Digital Twin is used at all stages of the product life cycle, including design, production, operation and utilization. Depending on the scale of operational modeling, the following levels of Digital Twin of manufacturing system are distinguished:

- Digital twin of manufacturing system. – Digital twin of manufacturing line.
- Digital twin of separate technological equipment of the manufacturing line.

An important task of the operating model of the digital twin manufacturing system is to minimize the possible malfunctions of the technological equipment by the timely carrying out of its technical support and repair. To solve this problem, the information from sensors of a real operating device is used as input for the Digital twin.

It allows to compare the information of the virtual sensors of the digital twin with the sensors of the real device, to detect anomalies and reasons of their occurrence.

Due to the equipment of the various sensors, the data on its technical state could be collected continuously, during its operation. The timely detection of even small deviations of the operating parameters allows to take prompt measures to ensure the normal operation of the equipment by its technical support.

Big Data technology (Big Data) allows to identify hidden patterns, that elude from limited human perception by the detection from data of previously unknown, non-trivial, practically useful experience, which are necessary for decision making. For this purpose, we use the special data processing methods, for example, the application of associative rules, classifications for categories, cluster analysis, regression analysis, detection and analysis of deviations, etc.

The aim of the research is the development of the digital twin principles of technological systems, methods of big data for creation of conditions for their predictive (proactive) technical support.

II. Materials and methods

As the experience of operating technological systems of di□erent manufactures shows, one of the most significant disadvantages is the low e□ciency of the main equipment[4, 5]. The key influence on the indication of this equipment is carried out either by idle time because of organizational and technical reasons, or by idle time due to the adjustment caused by the seasonality of the technological lines use in processing and packaging production[9, 12].

Almost always the operation of the manufacturing system is accompanied by the constant need to reduce its operating costs. The modern way to solve this problem is to create intelligent manufacturing system based on the applying of advanced information technology to optimize operating costs [2, 6, 16].

Summarizing the arguments of many researchers [3, 7, 10], we can confirm that the intelligent manufacturing system adapts to work in external conditions that change over time, basing on the appropriate adaptation algorithm. The basis for making management decisions to optimize operating costs for the technical support of intelligent technological systems is to create and apply digital models that describe the state of the manufacturing system [1, 18].

So, recently, there are two key sources for improving the e□ciency of integrated intelligent manufacturing systems:

- High flexibility of production that allows e□cient use of the equipment all year round, and is based on a su□ciently developed concept of flexible manufacturing systems (FMS), which allow within the technological capabilities of the integrated machines to produce a wide range of products at a cost close to cost of mass production [5, 9, 14].
- Complex automation of technological systems, when the transition from e□ciency improvement at each stage of production separately to optimization of the manufacturing process as a whole is made. Here with it is monitored the use of planning and scheduling functions of technological equipment loading, its control and diagnosis of its state to provide the proactive technical support. [13].

The applying of the intelligent systems of control, the increasing quantity of feedbacks in the technological equipment, help to increase the machine processes control and to give the independence in their functioning. [9, 18].

To form a technical support program for a technological system, it is necessary to consider possible manufacturing situations by making certain changes to the control algorithm [2, 18]. Obviously, first of all, we need to evaluate the external conditions to make necessary changes in the algorithm of technical support and to be provided with a digital model that

describes the technological system state in the given time intervals. [1].

The e□ciency of such use of digital models for forming of the technical support program for a technological system, despite the possibility of manufacture analyzing and identifying the necessary reasons for correction, has the disadvantage. Since the equipment is being operated day and night at many manufacturing facilities, its unexpected stoppage can can cause the considerable losses. [1, 8]. To prevent this problem, we have developed a mathematical model to determine the controlled parameters and features that describe the functioning of the technological automated machine, which allows to determine the current state of the machine and to predict the necessary preventive maintenance actions.

The manufacturing system digital model is based on collecting data on the time duration the equipment is in di□erent states. Information about the technological equipment stay in the state of idle time is obtained based on the analysis of the functioning and idle time of the technological equipment, which is placed in the standard equipment malfunctions and idle time log, carried out at each modern enterprise.

Ranking of the malfunction and idle time reasons in units of time, for example, in minutes, for the collectivity aggregate of the manufacturing system over a long observation time is performed by digital data processing. To solve the problem (to improve the manufacturing system) - means to eliminate the reasons that lead to a decrease of its e□ciency. Obviously, there are many such reasons, and eliminating each of them is a very di□cult and expensive.

Therefore, to identify the parameters that most a□ect the unsatisfactory index of the target property of the system, we use a method of data aggregation, known from system analysis. While creating collective objects of analysis, you can determine the e□ciency of the equipment as a fraction of the useful time of its operation. Namely, the improvement and stabilization of these generic technical equipment parameters should be focused on during the following technical support planning steps. Such aggregate parameters (collective) are often found among the target properties of the system.

The logic of system analysis involves identifying the most significant reasons that lead to the problem and concentrating e□ort on them.

To determine the probability of diagnoses by Bayes method it is necessary to make a diagnostic matrix, which is formed on the basis of previous statistical material. This table shows the probabilities of the categories of characteristics and their corresponding probabilities of predicted diagnoses for di□erent combinations of characteristics. The size of the studied values is determined by the number of possible (potential) characteristic failures and malfunctions.

Ideally, to fill this matrix, it is necessary to keep a continuous record of the technical state of the equipment during the passing of each regular technical support with the fixing of the state of its systems and mechanisms.

The term ”technical diagnosis” means recognition of the control object state in the conditions of incomplete information [5, 15]. In this case, it is assumed that the most known are those states that are often met. From the definition it follows that the primary choice is a set of states, which is very large for a complex object. Diagnosis, as a rule, should answer the question:

- What is the malfunction?
- What item of equipment has a malfunction?
- What is the prognosis for further course of the process?
- What is the e□ect of the malfunction on the control object characteristics and the quality of the technological process?

We select five main machine states and three main diagnostic characteristics (parameters) of various malfunctions:

- malfunctions in the system of laminate supply – state D_1;
- malfunctions in the system of cork supply – state D_2;
- malfunctions in the system of supply and dosage of paste – a product – state D_3;
- malfunctions in the system of welding mechanisms – state D_4;
- machine state in working order – D_0.

We accept that the basis for the proactive maintenance of the machine is the extreme(critical) state of the machine due to the productivity, product quality, material consumption. As a diagnostic characteristic we choose the intensity of change of these indicators.

As it is known, they reflect the dynamic characteristics of the machine and depend on its technical state. Each characteristic has three levels of state: good, satisfactory, and unsatisfactory, corresponding to a system malfunction:

- the intensity of time and e□ciency changes – k_1;
- the intensity of quality product change – k_2;
- the intensity of material consumption change – k_3.

The range of index changes from good to unsatisfactory is determined by the corresponding values of these indexes for prolonged operation of the machine. These data must be obtained in the manufacturing environment and they serve to determine the frame conditions in diagnosis of the state of the machine and to determine the moments of predictive technical support and repair.

III. Experiment

The development of a digital model of the technological line was done in AnyLogic.

AnyLogic simulation environment is based on an object-oriented concept and has a number of benefits. Some of them are the existence of all modeling paradigms (high choice flexibility of approach); the possibility of choosing between paradigms or applying an integrated approach; it has all the properties necessary for the development of simulation models [11].

AnyLogic environment is developed in a universal Java programming language, which allows it not to depend on the type of operating system. Simulation modeling consists of two

major stages: the creation of a model and the analysis of the data obtained with the help of a decision model.

The elements of the system, their relationships, parameters and variables, as well as their relationships and the laws of their changes, must be expressed by means of the simulation environment, that is, in this environment, variables and parameters of the model must be defined, procedures for calculating the variables' changes and model characteristics in time are constructed .

First of all, you need to define the input parameters for the simulation model. The input parameters of the model are:

- type of product;
- speed of the conveyor;
- parameters of packing machines (weight of packed products);
- breakage of conveyors.

This model will be done in a discrete-event abstraction.

IV. Result and discussion

The created digital model reflects the operation of the manufacturing line in perfect conditions with 100 % e□ciency. The model, created in a simulation environment.

The authors created the action part of operation of the multi-position packing machine, which includes the following modules:

- Laminate supply System (FFS).
- Welding system (WS).
- System of supply and dosage of product (DS).
- Cork supply system (CS).
- Welding system (WS1).
- Conveyor System (conveyorn).

To reflect the adequate functioning of the technological system under the conditions of malfunctions of its constituent elements, a block for its technical support in case of malfunctions was introduced.

The created model in the further researches gives an opportunity to predict the characteristics of malfunctions, and to optimize the process of under other operating conditions, the normal state of the machine (state D_0) is characterized by other values of the diagnostic characteristics $k_1, k_2, k_3,$ which can be determined statistically. A significant reduction in experimental studies is achieved with a help of a simulation model.

CONCLUSIONS

For manufacturers of modern technological equipment, it becomes clear that to increase the production e□ciency requires more functional, flexible and productive machines that are connected to the information network of the enterprise.

To achieve this, in recent years, more and more enterprises in the area of manufacturers of packing equipment are expanding production automation systems, which already provide for improving not only the control and operation of industrial equipment, but also the introduction of methods of its self-diagnosis and predictive (proactive) technical support.

For this purpose, a number of solutions have already been developed, based on the applying of digital duplicates of technological systems of machine-building, packing and other manufactures based on the use of methods of big data analysis.

Considering that digital twin technologies enable for product manufacturers to make proactive decisions based on big data, e□ciency in the applying of industrial equipment is achieved. The areas of the highest growth in the digital twin apply are high-productive manufactures, which include the machine-building, packaging, pharmacological and food industries.

Obviously, to solve the problems of industrial equipment e□ciency increasing by using the digital twin principles of technological systems, methods of big data for creation of conditions for their predictive (proactive) technical support is an actual scientific and practical task.

References

[1] S. Boschert, C. Heinrich, and R. Rosen, "Next generation digital twin", Proceedings of TMCE, vol. 1, 2018, pp. 209–2017.

[2] B. Gaines, and D. Norrie, "Knowledge systematization in the international ims research program". Proc. of IEEE Conference on Systems, Man and Cybernetics Intelligent Systems for 21st Century, 1995, pp. 958 – 963.

[3] A. Gola, A. Swic, and V. Kramar, "A multiple-criteria approach to machine-tool selection for focused flexible manufacturing systems", Management and Production Engineering Review vol. 2(4), 2011, pp. 21–32.

[4] S.Goldman, R. Nagel, and K. Preiss, Agile competitors and virtual organizations: strategies for enriching the customer, Van Nostrand Reinhold, New York, 1995.

[5] M. P. Groover, and E.W. Zimmers, CAD/CAM: computer-aided design and manufacturing. Prentice-Hall, New Jersey, 1984.

[6] L. M. Camarinha-Matos, H. Afsarmanesh, and V. Marik, "Intelligent Systems for Manufacturing Multi-agent systems and virtual organizations", Kluwver Academic Publishers, IFIP vol 130, 1998, pp. 137–140.

[7] N. J. Nilsson, Principles of Artificial Intelligence. Palo Alto, CA: Tioga, 1980.

[8] B. Palchevskyi, "Application of digital models in intelligent packaging systems", TC2018 International Conference Progressive directions of technological complexes development, vol. 1, pp. 3–5, 2018.

[9] B. Palchevskyi, A. Swic, and H. Krestyanpol, "Increasing e□ciency of flexible manufacturing systems based on computer product grouping", Advances in Science and Technology Research Journal, vol 12(2), 2018, pp. 6–10.

[10] B. Palchevskyi, O. Krestyanpol, and L. Krestyanpol, "Principles of designing and developing intelligent manufacturing systems of packaging", International journal for science, technics and innovations for the industry, vol 11, 2017, pp. 515–519.

[11] B. Palchevskyi, O. Krestyanpol, and L. Krestyanpol, "The use of Anylogic system for modelling a flexible automated packing system in training engineering students", Information Technologies and Learning Tools, vol. 75(1), 2020, 225–236.

[12] B. Palchevskyi, A. Swic, and H. Krestyanpol, Computer integrated designing of flexible manufacturing systems. Lublin University of Technology, Lublin, 2015.

[13] S.J. Russell, and P. Norvig, Artificial Intelligence. A Modern approach. NJ:PrenticeHall, Englewood Cli□s, 1995.

[14] M. Schwaninger, Intelligent Organizations. Powerful Models for Systemic Management. Springer-Verlag Berlin Heidelberg, 2009.

[15] F.P. Tarasenkov, Applied system analysis.Tomsk, 2004.

[16] T. Tolio, Design of Flexible Production Systems. Methodologies and Tools. Springer-Verlag Berlin Heidelberg, 2009.

[17] D. A. Waterman, Design of Flexible Production Systems. Methodologies and Tools. Mir, 1989.

[18] L.S. Yampolsky, P.P. Melnychuk, K.B. Ostapchenko, and O.I. Lisovichenko, Flexible computer-integrated systems: planning, modeling, verification, control. Zhytomyr, 2010.

IEEE Third International Conference on Data Stream Mining & Processing
August 21-25, 2020, Lviv, Ukraine

Embedding technique and network analysis of scientific innovations emergence in an arXiv-based concept network

Serhii Brodiuk
Faculty of Applied Sciences, Ukrainian Catholic University
UA–79011 Lviv, Ukraine
brodiuk@ucu.edu.ua

Vasyl Palchykov
$\mathbb{L}^4$ *Collaboration & Doctoral College for the Statistical Physics of Complex Systems*
Leipzig-Lorraine-Lviv-Coventry, Europe
palchykov@gmail.com

Yurij Holovatch
Institute for Condensed Matter Physics, National Acad. Sci.of Ukraine
UA–79011 Lviv, Ukraine
$\mathbb{L}^4$ *Collaboration & Doctoral College for the Statistical Physics of Complex Systems*
Leipzig-Lorraine-Lviv-Coventry, Europe
Centre for Fluid and Complex Systems, Coventry University
Coventry, CV1 5FB, United Kingdom
hol@icmp.lviv.ua

***Abstract*—Novelty is an inherent part of innovations and discoveries. Such processes may be considered as an appearance of new ideas or as an emergence of atypical connections between the existing ones. The importance of such connections hints for investigation of innovations through network or graph representation in the space of ideas. In such representation, a graph node corresponds to the relevant concept (idea), whereas an edge between two nodes means that the corresponding concepts have been used in a common context. In this study we address the question about a possibility to identify the edges between existing concepts where the innovations may emerge. To this end, we use a well-documented scientific knowledge landscape of 1.2M `arXiv`.org manuscripts dated starting from April 2007 and until September 2019. We extract relevant concepts for them using the ScienceWISE.info platform. Combining approaches developed in complex networks science and graph embedding, we discuss the predictability of edges (links) on the scientific knowledge landscape where the innovations may appear.**

***Index Terms*—complex networks, embedding, concept network, arXiv**

I. Introduction

An idea of scientific analysis of science is not new. It is at least as old as the science itself, see, e. g. [1] and references therein. Contemporary studies in this domain share a common specific feature: besides traditional philosophical and culturological context, such analysis attains quantitative character. The questions of interest cover a wide spectrum, ranging from fundamental, such as: what is the structure of science? How do its constituents interact? How does knowledge propagate? [2]–[4] to entirely practical ones: which fields of science deserve financial investments or how to rate scientists in a particular domain? [5]–[7]. All these and many more questions constitute a subject of a science of science or logology [8].

The problem we consider in this paper concerns an emergence of new scientific knowledge or the so-called scientific innovation. Quantitative investigation and modeling of innovations are not straightforward. On the one hand, one may think of innovation as an emergence of a new idea, see, e. g. [9]. Another approach considers innovation as an atypical combination of existing ideas, see, e. g. [10]. The goal of our work is to suggest a way to quantify analysis of scientific innovations emergence and to propose an approach to identify edges on the graph of knowledge where innovations may emerge. We believe that such analysis, if successful, is useful both from the fundamental point of view, explaining properties of knowledge formation, as well as is of practical relevance, helping to detect innovation-rich fields.

To reach this goal, we will analyze a body of scientific publications taking an `arXiv` repository of research papers [11] and studying its dynamics with a span of time. We will use a specially tailored software, ScienceWISE.info platform [12], to extract a set of concepts from all publications on an annual basis. These are the properties of this set of concepts that will serve us as a proxy of structural features and dynamics of human knowledge. In particular, we will use complex network theory [13]–[16] to track intrinsic connections between concepts that are contained in different papers. Using several completing each other approaches we will construct a complex network of concepts (as a proxy of a complex network of knowledge) and we will calculate its main topological characteristics, paying particular attention to the emergence of new links between existing concepts. These last may serve as a signal about an emergence of atypical combinations between existing ideas, i. e. about scientific innovations. We will refine

978-1-7281-3214-3/20/$31.00 ©2020 IEEE

our analysis by exploiting embedding technique [17], [18] to quantify a proximity measure between different concepts and in this way, we will establish a solid and falsifiable procedure to quantify an emergence of possible scientific innovations in certain fields of science.

The rest of the paper is organized as follows. In the next Section II we describe the dataset used in the analysis, the data is presented in complex network form and analysed in Section III. In Section IV we introduce the concept embedding technique and study dynamics of link appearance. The results are summarized in the last Section V.

II. Dataset

We use the E-repository of preprints arXiv.org [11] as a source of data: at the moment of writing this paper, there are about 1.6M full-text accesible manuscripts uploaded to the arXiv. It makes them an optimal source to extract scientific ideas/concepts. The arXiv covers a variety of scientific fields such as physics, mathematics, computer science, quantitative biology, quantitative finance, statistics, electrical engineering and systems science, and economics. The average daily upload rate is 400 $\approx$ 12.5K new manuscripts per month (every next year there are $\approx$ 5000 more articles than the previous one, starting from 1991). Each paper submitted to the arXiv contains, besides the full-text, different metadata such as authors, subject category (categories), journal reference, DOI if any, submissions history with dates, etc. For the purpose of our study, we need to extract specific words or combination of words that carry a specific scientific meaning (concepts) from each manuscript. The set of concepts to some extent represent the content of the paper, both with respect to the subject of research and methods applied. To this end, we will use a ScienceWISE.info platform, specially tailored for such tasks.

The ScienceWISE.info platform [12], [19] has been built to support the daily activities of research scientists. The goal of the platform is to "understand" the interests of its users and to recommend them relevant newly submitted manuscripts. For this purpose, arXiv serves as one of the data source of new submissions. In order to understand the research interests of the users, the platform extracts scientific concepts from the texts of the manuscripts and compares the concept vector of the manuscript and the corresponding concept vector of the user's research interest. Concept extraction approach implemented into this platform has two phases: i) automatic key phrase (concept candidate) extraction and ii) [optional] crowd-sourced validations of scientific concepts. During the first phase, each manuscript is scanned by the KPEX algorithm [20]. The algorithm extracts key phrases from the text of the manuscript, and these key phrases serve as concept candidates. Then, during the second step, the concept candidates are reviewed by the registered users of the platform who are permitted to validate the concepts. The described procedure arrived at approximately 20,000 concepts as of the date when this paper was written. About 500 of them have been marked as generic concepts assuming their generic meaning (the ones like Energy, Mass or Temperature).

Navigating over ScienceWISE.info platform at the end of September of 2019, we accessed a collection with near 1.2M arXiv manuscripts with metadata and concepts list for each one (from April of 2007 till September of 2019). As data is publicly available (anyone with access to the internet could get it), we scraped it to storage on our side with a convenient structure for further manipulations. A detailed data parsing approach could be found at a GitHub repository [21]. For each manuscript, a set of concepts found within its text has been recorded. The total number of unique extracted concepts is 19,446 and the number of concepts per manuscript varies in range 0 – 1164. A similar dataset (it can be considered as a small subset of the described above) of 36386 articles in Physics domain have been previously investigated in [22]–[24]. Once the dataset is downloaded and prepared for the analysis, the first step of our investigation is to analyze the topological properties of the resulting concept network using the tools of Complex network theory as descibed in the next Section III.

III. Concept networks and their topological features

The above described dataset may be naturally represented as a bipartite network, details of network construction are shown in Fig. 1. Our further analysis is based on a single-mode projection of this network to the concept space, Fig. 1 **d**. We will call the resulting network a concept network. There, a link between two nodes means that the corresponding concepts have appeared together in the lists of concepts for at least one manuscript.

To proceed with the analysis, we will consider two slices of data: the manuscripts submitted during the years 2013 and 2015. This will allow us, in particular, to compare some properties of a concept network as they evolve in time. The first subset (the year 2013) network consists of 16,229 nodes, whereas for the year 2015 we arrived at 16,660 nodes. We will refer to these networks as g-2013 and g-2015, correspondingly. These networks share 15,431 concept-nodes in common.

To take into account link strength, we will implement two filtering procedures that keep only significant links in a network. Within the first procedure, we assign a weight w_{ij} to the link between sites i and j such that it equals the number of manuscripts in the dataset that contains concepts i and j simultaneously. Then the simplest way to filter out insignificant links is to consider the hard threshold on the link weight. Below we will consider threshold value $\omega = 10$ and keep the links for which $w_{ij} > \omega$. In our case, approximately 16.5% of total links remain after such filtering. If such procedure removes all links from a node, the node is removed too, so there are no isolated nodes in the network. We will refer to the resulting networks as w-2013 and w-2015, for the corresponding years.

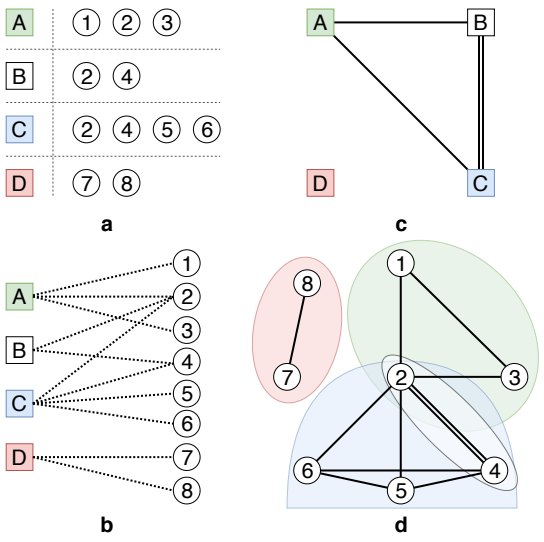


Fig. 1. Illustration of the dataset and three network representations constructed from it. Squared nodes represent manuscripts and circles represent concepts. Panel **a** illustrates a dataset of four manuscripts and the concepts identified within each of them. Panel **b** is a bipartite network representation of the dataset. Panels **c** and **d** illustrate single-mode projections of the bipartite network to the manuscript and concept spaces, correspondingly. Connections between nodes represent how many concepts has a given pair of manuscripts in common (panel **c**) or how many manuscripts shared a given pair of concepts (panel **d**).

A more sophisticated approach for filtering insignificant links is to consider the disparity filter proposed in Ref. [25]. The key idea of the method is to calculate the probability α_{ij} that a given link is as strong or even stronger as observed in a random setting. This probability, known as p-value reads

$$\alpha_{ij} = 1 - (k_i - 1) \int_0^{p_{ij}} (1-x)^{k_i - 2} dx \qquad (1)$$

where k_i is degree (i.e. the number of links) of the i-th node, and $p_{ij} = w_{ij}/(\sum_j w_{ij})$ is the normalized link weight. Then one may set a threshold for p-value, and only the links with small enough p-value are kept, meaning that a random process can not arrive at a link with such weight. In our analysis we set a threshold for p-value $\rho = 0.1$, and keep a link between i and j if $\alpha_{ij} < \rho$ or $\alpha_{ji} < \rho$, i. e. if it is significant from at least one node standpoint. As a result of such procedure, approximately 85% of links are removed as insignificant. The resulting concept networks for the corresponding years will be referred to as `d-2013` and `d-2015`.

With the networks at hand, it is straightforward to compare them measuring standard indices that quantify their different features. In Table I we report some values obtained by us, a more comprehensive comparison can be found in Ref. [26]. There, besides the number of network nodes N and links L and mean and maximal node degrees $\langle k \rangle$, $k_{\max}$, we provide the values, that characterize network size (mean and maximal shortest path lengths l, $l_{\max}$ measured as a shortest number of steps between two different nodes) and correlations in network structure. To quantify correlations, we measured mean clustering coefficient $\langle c \rangle$, global transitivity C and assortativity r. The clustering coefficient c_i of node i describes the level of connectivity among its neighbours: $c_i = \frac{2m_i}{k_i(k_i-1)}$ where m_i is the number of existing connections among k_i neighbouring nodes. Therefore, the mean value $\langle c \rangle$ of c_i, averaged over all nodes in the network, characterizes the local density of neighborhood links in the entire network. Instead of calculating the average value of local measurements, global transitivity C is defined as a ratio between the total number of connected triplets in the network and the number of all possible triangles. In turn, assortativity r is defined as Pearson correlation coefficient between node degrees on both ends over existing link [13], [14], [27].

Our analysis of the topology of the concept network indicates that observed concept networks are heterogeneous graphs that obey internal clustering (community structure) and hierarchical organization. These properties of a concept network are independent of the subset of data used (constructed from 2013 and 2015 year data). These features, however, are more pronounced once weak links have been removed. As it follows from the comparison of data obtained for different years and via different procedures of relevant link determination, cf. Table I, complex networks under consideration attain a range of universal features that do not change with time and characterize the system of concepts as a whole. In particular, they are the small world networks [13], [14], [27]–[30] characterized by a small size (mean shortest path and maximal shortest path values) and large value of clustering coefficient. The last also brings about the presence of strong correlations. Moreover, an essential difference between the clustering coefficient and global transitivity serves as evidence of possible community structure. In turn, the negative value of assortativity suggests that they are disassortative networks where a group of central nodes (hubs) serves as common attraction points for nodes with lower degree values.

IV. Scientific innovations and concept embedding

In this section, we investigate the possibility to detect in advance fields where scientific innovations may emerge. In particular, we are interested in the questions of the prediction power of concept embedding.

Investigation of scientific innovation emergence is not straightforward. The simplification adopted in frames of this paper considers innovations as the appearance of a new statistically significant link between nodes that previously were not linked to each other. In this way, the emergence of such a link is treated as a novelty introduced into the graph of scientific concepts.

We proceed by considering a network of scientific concepts built upon manuscripts submitted to `arXiv` during the year 2013. Let us consider a pair of concepts i and j. In terms of link existence, these concepts may be either connected by a link or disconnected, meaning no link between i and j. The fraction of pairs connected by links equals to the density of links, $\rho = 2L/N(N-1)$, in the corresponding concept network. For a `g-2013` network it has a value $\rho = 8.46\%$, see Table I. Some of the links that carry low weight may be

TABLE I

AGGREGATED CHARACTERISTICS OF CONCEPTS NETWORKS. N, L: NUMBER OF NODES AND LINKS; $\rho = 2L/N(N-1)$: DENSITY OF LINKS; $\langle k \rangle$, $k_{\max}$: MEAN AND MAXIMAL NODE DEGREE; l, $l_{\max}$: MEAN AND MAXIMAL SHORTEST PATH LENGTHS; $\langle c \rangle$: MEAN CLUSTERING COEFFICIENT; C: GLOBAL TRANSITIVITY; r: ASSORTATIVITY. SEE THE TEXT FOR MORE DESCRIPTION.

network	N	$L, \times 10^6$	$\rho, \%$	$\langle k \rangle$	$k_{\max}$	l	$l_{\max}$	$\langle c \rangle$	C	r
g-2013	16,229	11.1	8.46	1,373	15,345	1.92	3	0.77	0.37	-0.324
g-2015	16,660	12.7	9.12	1,520	15,935	1.91	4	0.77	0.38	-0.325
w-2013	9,999	1.8	3.69	369	8,856	2.00	4	0.89	0.28	-0.390
w-2015	10,770	2.2	3.84	414	9,661	2.00	4	0.89	0.28	-0.382
d-2013	13,358	1.6	1.84	246	11,665	2.01	4	0.90	0.14	-0.375
d-2015	13,969	1.9	1.92	268	12,367	2.00	5	0.89	0.14	-0.368

considered as spurious links rather than statistically significant, meaning that they could arise as a result of noise rather than a real coupling between the corresponding concepts. In this paper, we consider two alternative ways to filter out such spurious links: i) naive filtering by setting up a link weight threshold and ii) disparity filtering that employs statistical significance testing, as explained in Section III.

With the thresholds set above ($\omega = 10$ and $\rho = 0.1$), majority of the concept pairs (out of about 130M potential connections) are either disconnected or are connected by spurious links. Namely, 98.6% of concept pairs out of all possible $N(N-1)/2$ pairs are disconnected or connected by weak links ($\omega_{ij} \leq \omega$, referred below as `weak/missing links`) and 98.8% of pairs are either disconnected or connected by a statistically insignificant links ($\alpha_{ij} \geq \rho$, referred below as statistically `insignificant links`).

Some of these pairs may become connected in the future by strong or statistically significant connections. The emergence of such connections is referred in this paper as scientific innovations. Our analysis indicates that only 564,330 pairs (0.43%) became strongly connected ($w_{ij} > 10$) in 2015 out of 129,837,653 weakly connected/disconnected pairs in 2013. Disparity filter arrives at a similar picture. Only 475,788 pairs (0.37%) became statistically significant in 2015 out of 130,039,148 insignificant/disconnected pairs in 2013. To conclude, less than 0.5% of `weak/missing links` or statistically `insignificant links` between concepts in 2013 became strong/significant in the year 2015.

Thus the questions of our interest are related to forecasting the pairs where such innovations may emerge given the number (fraction) of such connections is known. In particular, we are interested in the power of concept embedding technique [17], [18] to distinguish between the pairs of concepts that will become connected vs the pairs that will stay disconnected in the future. The key assumptions are that i) concepts that appear in a similar context will have close enough vectors in embedded space and ii) that the concepts that carry similar content are more likely to become connected in the future.

For this reason we use concept co-occurrence matrix for year 2013 and embedded each concept vector in 100 dimensional space using PyTorch-BigGraph [18]. The whole detailed pipeline we used for described graphs and embeddings formulation can be found at the GitHub repository [21]. As a result, each concept i becomes associated with a vector $\vec{v_i}$ in the embedded space. The similarity s_{ij} between a pair of concepts i and j is then calculated as a cosine similarity between the corresponding vectors $\vec{v_i}$ and $\vec{v_j}$.

Once similarities s_{ij} between concept vectors in embedded space have been calculated for each pair of concepts i and j, we divide all pairs of concepts into two groups: i) `Strong embedding similarity group` and ii) `Weak embedding similarity group`. To distribute pairs of nodes/concepts among the groups, we put an arbitrarily selected threshold of $\zeta = 0.6$. The pairs of concepts for which embedding similarity $s_{ij} \leq \zeta$ are assigned to `Weak embedding similarity group`, for convenience, we will refer to the corresponding pairs as `dissimilar concepts`. Instead, if the embedding similarity between concepts i and j, $s_{ij} > \zeta$, the corresponding pair is assigned to a `Strong embedding similarity group` and will be referred below as `similar concepts`. We expect that the selection of the other value of ζ threshold will not change the qualitative results of our analysis. Especially, because pairs on both extremes of embedding similarity will eventually be assigned to different groups.

The results of our analysis indicate significant differences in the allocation of pairs of concepts among embedding similarity groups for weakly and strongly connected pairs of nodes in the network. While only 1.2% of `weak/missing links` in `g-2013` falls into `similar concepts` group, this fraction is much higher for `strong links`, reaching 22.8%. Similar results have been observed if one uses a disparity filter instead of link weight threshold filter. Thus, we expect that the grouping of pairs of nodes using embedding similarity improves predictions of the pairs of concepts where statistically significant links will be established in the future.

With the data about the concept network for the year 2013 at hand, let us now consider the network of scientific concepts constructed from manuscripts submitted to `arXiv` during the year 2015. Below we perform preliminary analysis rather than propose a predictive model.

Comparing the networks constructed from data of years 2013 and 2015, we see that the majority of strongly connected concept pairs in 2015 were connected by strong links in 2013 too. Table II shows that about 90% of strong links in 2013 remained strong in the year 2015. If we take into account grouping by concept embedding similarity, we observe additional segregation: strong links with low embedding

TABLE II
PERCENTAGE OF CONCEPT PAIRS THAT BELONG TO A SPECIFIC COMBINATION OF LINK WEIGHT GROUP AND EMBEDDING SIMILARITY GROUP IN 2013 THAT EITHER REMAINED OR BECAME STRONG IN 2015.

	dissimilar concepts during 2013	similar concepts during 2013
weak links during 2013	0.4%	3.19%
strong links during 2013	88.89%	94.34%

similarity in the year 2013 remained strong in the year 2015 in almost 89% of cases, while strong links with strong embedding similarity in the year 2013 remained strong in the year 2015 for more than 94% of cases. These results lead us to the following conclusions. First, if a link between two concept-nodes exists and this is a strong link, then it is likely that the link will exist in the future, and it will remain the strong one. In other words, the strength of a link is a good predictor for a link to belong to the same category in the future. Second, strong links with high concept embedding similarity have higher chances to remain strong in the future than strong links that are characterized by low embedding similarity.

On the other side, weak links evolve to strong links quite rarely. Only 0.4% of weak links in 2013 evolved to strong links in year 2015. However, classification of concepts pairs by their embedding similarity allowed us to identify a subgroup of these pairs for which the probability of becoming strong connections raises to 3.19%, i. e. in about 8 times. Even though the concept embedding similarity does not point the "future" emergence of a new strong link in the network exactly, the results of our analysis indicate its power as one of the features to be used in such predictions.

Similar results have been obtained if we use classification of links between pairs of concepts using statistical significance testing instead of link weight threshold, see Table III.

TABLE III
PERCENTAGE OF CONCEPT PAIRS THAT BELONG TO A SPECIFIC COMBINATION OF LINK SIGNIFICANCE GROUP AND EMBEDDING SIMILARITY GROUP IN 2013 THAT EITHER REMAINED OR BECAME STRONG IN 2015.

	dissimilar concepts during 2013	similar concepts during 2013
insignificant links during 2013	0.3%	3.01%
significant links during 2013	82.71%	91.06%

Thus, independent of the method used to classify pairs of concepts, either using link weight threshold or statistical significance testing, the results of our analysis indicate the ability of concept embedding similarity in predicting scientific innovations, i. e. the emergence of strong or statistically significant links in a concept network.

V. CONCLUSIONS AND OUTLOOK

The goal of our work was to analyze the possibilities of innovation emergence in the course of knowledge generation. To this end, we have investigated the structure and dynamics of connections between scientific concepts that constitute a body of research papers, as recorded in the arXiv repository [11]. We have applied two methods, concept embedding and network analysis, to quantify properties of sets of concepts and to predict the emergence of new links (innovations) between different concepts. We have shown that whereas each of the above methods is a powerful tool to define certain features of a system of concepts, it is the combination of these two methods that leads to a synergetic effect and allows to forecast dynamics of new links creation and evolution of a system as a whole. The main results obtained in the course of our analysis include the following:

- We have represented a system of concepts of scientific papers in the form of a complex network. Different nodes in this network correspond to different concepts, and a link between two nodes-concepts means that they were exploited in the same paper. We have determined the quantitative characteristics of a complex network of concepts and their evolution with time, and the data is given in Table I.
- We have used two complementary approaches to define the presence of a strong link between two nodes, i. e. of a link that serves as evidence of a relevant connection. In one approach, the criterion is given by a link weight. The second method takes into account subtle information about network intrinsic structure [25]. Corresponding data is shown in Table I.
- As is follows from the comparison of data obtained for different years and via different procedures of relevant link determination, see Table I, complex networks under consideration attain a range of universal features that do not change with time and characterize the system of concepts as a whole. In particular, they are the small world networks characterized by small size (mean the shortest path and maximal shortest path values) and large value of the clustering coefficient. The last also brings about the presence of strong correlations. Moreover, an essential difference between the clustering coefficient and global transitivity serves as evidence of possible community structure. In turn, the negative value of assortativity suggests that they are disassortative networks where a group of central nodes (hubs) serves as common attraction points for nodes with lower degree value.
- Concept embedding technique enabled us to find out proximity (by context, by subject, or related in any other way) between different concepts. With a measure of proximity at hand, we were in a position to compare it with the dynamics of new links emergence between different concepts. In turn, this enables one to reveal groups of concepts (subsequently – fields of knowledge) where innovations are probable to emerge. Corresponding

statistical analysis is summarized in Tables II and III.

The results obtained in this study may be useful both from the fundamental point of view, contributing to our understanding of how the knowledge is formed, as well as they may have the practical implementation. In particular, the methodology elaborated in the course of our analysis can be used to detect fields where innovations have a higher probability of appearing. A natural way to continue the analysis presented here is to evaluate practical outcomes (i. e. impact) of papers, where the higher probability of innovation is predicted. With the scientometric data at hand, such a task is not much time consuming and will be a subject of future work. Another work in progress is to suggest a model predicting the emergence of statistically significant links between already existing concepts.

Acknowledgement. A part of this work has been performed in the frames of the Master thesis by S.B. in Ukrainian Catholic University (Lviv, Ukraine).

REFERENCES

[1] L. Zhmud, *The origin of the History of Science in Classical Antiquity*. Walter de Gruyter, 2008, vol. 19.

[2] T. Lewens, *The meaning of science: An introduction to the philosophy of science*. Hachette UK, 2016.

[3] O. Mryglod, Yu. Holovatch, R. Kenna, and B. Berche, "Quantifying the evolution of a scientific topic: reaction of the academic community to the chornobyl disaster," *Scientometrics*, vol. 106, no. 3, pp. 1151–1166, 2016.

[4] M. Krenn and A. Zeilinger, "Predicting research trends with semantic and neural networks with an application in quantum physics," *arXiv preprint arXiv:1906.06843v2*, 2020.

[5] L. Leydesdorff and S. Milojević, "Scientometrics" in: Lynch Micheal, editor. *International Encyclopedia of Social and Behavioral Sciences*, 2015.

[6] B. Berche, Yu. Holovatch, R. Kenna, and O. Mryglod, "Academic research groups: evaluation of their quality and quality of their evaluation," in *J. Phys.: Conf. Ser.*, vol. 681, 012004, 2016.

[7] S. Thurner, W. Liu, P. Klimek, and S. A. Cheong, "The role of mainstreamness and interdisciplinarity for the relevance of scientific papers," *arXiv preprint arXiv:1910.03628*, 2019.

[8] A. Zeng, Z. Shen, J. Zhou, J. Wu, Y. Fan, Y. Wang, and H. E. Stanley, "The science of science: From the perspective of complex systems," *Physics Reports*, vol. 714, pp. 1–73, 2017.

[9] I. Iacopini, S. Milojević, and V. Latora, "Network dynamics of innovation processes," *Phys. Rev. Lett.*, vol. 120, no. 4, p. 048301, 2018.

[10] B. Uzzi, S. Mukherjee, M. Stringer, and B. Jones, "Atypical combinations and scientific impact," *Science*, vol. 342, no. 6157, pp. 468–472, 2013.

[11] arXiv gives an open access to 1,640,097 e-prints in Physics, Mathematics, Computer Science, Quantitative Biology, Quantitative Finance, Statistics, Electrical Engineering and Systems Science, and Economics, accessed: 2020-01-03. [Online]. Available: https://arXiv.org

[12] The ScienceWISE project aims to develop a scientist-generated on-line knowledge base fully integrated into the physics ArXiv.org, accessed: 2020-01-06. [Online]. Available: http://ScienceWISE.info

[13] Yu. Holovatch, O. Olemskoi, C. von Ferber, T. Holovatch, O. Mryglod, I. Olemskoi, and V. Palchykov, "Complex networks," *Journ. Phys. Stud.*, vol. 10, pp. 247–289, 2006.

[14] M. Newman, *Networks: An Introduction*. OUP Oxford, 2010. [Online]. Available: https://books.google.com.ua/books?id=LrFaU4XCsUoC

[15] A.-L. Barabási *et al.*, *Network science*. Cambridge university press, 2016.

[16] Yu. Holovatch, M. Dudka, V. Blavatska, V. Palchykov, M. Krasnytska, and O. Mryglod, "Statistical physics of complex systems in the world and in Lviv," *Journ. Phys. Stud.*, vol. 22, no. 2801, p. 21, 2018.

[17] 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*, 2013, pp. 3111–3119.

[18] A. Lerer, L. Wu, J. Shen, T. Lacroix, L. Wehrstedt, A. Bose, and A. Peysakhovich, "Pytorch-biggraph: A large-scale graph embedding system," *arXiv preprint arXiv:1903.12287*, 2019.

[19] A. Martini, A. Lutov, V. Gemmetto, A. Magalich, A. Cardillo, A. Constantin, V. Palchykov, M. Khayati, P. Cudré-Mauroux, A. Boyarsky *et al.*, "Sciencewise: Topic modeling over scientific literature networks," *arXiv preprint arXiv:1612.07636*, 2016.

[20] A. Constantin, "Automatic structure and keyphrase analysis of scientific publications," Ph.D. dissertation, The University of Manchester (United Kingdom), 2014.

[21] github.com/sergibro/concept-graphs

[22] 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, no. 1, p. 28, 2016.

[23] V. Palchykov and Yu. Holovatch, "Bipartite graph analysis as an alternative to reveal clusterization in complex systems," in *2018 IEEE Second International Conference on Data Stream Mining & Processing (DSMP)*. IEEE, 2018, pp. 84–87.

[24] V. Palchykov and Yu. Holovatch, *to be published*, 2020.

[25] M. Á. Serrano, M. Boguná, and A. Vespignani, "Extracting the multiscale backbone of complex weighted networks," *Proc.Nat. Acad. Sci.*, vol. 106, no. 16, pp. 6483–6488, 2009.

[26] S. Brodiuk. *Concept embedding and network analysis of scientific innovations emergence*. Master thesis, Ukrainian Catholic University, Lviv, 2019.

[27] M. E. Newman, "The structure and function of complex networks," *SIAM Review*, vol. 45, no. 2, pp. 167–256, 2003.

[28] R. Albert and A.-L. Barabási, "Statistical mechanics of complex networks," *Rev. Mod. Phys.*, vol. 74, no. 1, p. 47, 2002.

[29] S. N. Dorogovtsev and J. F. Mendes, *Evolution of networks: From biological nets to the Internet and WWW*. OUP Oxford, 2013.

[30] A. Barrat, M. Barthelemy, and A. Vespignani, *Dynamical processes on complex networks*. Cambridge university press, 2008.

IEEE Third International Conference on Data Stream Mining & Processing
August 21-25, 2020, Lviv, Ukraine

Knowledge Novelty Assessment During the Automatic Development of Ontologies

Vasyl Lytvyn
Information Systems and Networks Department
Lviv Polytechnic National University
Lviv, Ukraine
Vasyl.V.Lytvyn@lpnu.ua

Yevgen Burov
Information Systems and Networks Department
Lviv Polytechnic National University
Lviv, Ukraine
yevhen.v.burov@lpnu.ua

Victoria Vysotska
Information Systems and Networks Department
Lviv Polytechnic National University
Lviv, Ukraine
Victoria.A.Vysotska@lpnu.ua

Viktor Hryhorovych
Informatics and Information Systems Department
Drohobych Ivan Franko State Pedagogical University
Drohobych, Ukraine
dspu@dspu.edu.ua

***Abstract*—Evaluation of knowledge in the process of automatic ontology building in this article the estimation of a novelty of knowledge during the automation of ontology construction based on the analysis of text resources is considered. The development of ontology begins with some basic ontology that sets the examined domain area. The algorithm of estimation of novelty knowledge is developed.**

Keywords—ontology; concept; novelty of knowledge

I. INTRODUCTION

Functioning of intelligent decision support systems (IDSS) is a constant decision-making based on the analysis of current situations to achieve a certain goal. A typical IDSS operation scheme consists of the following three steps:

1) Planning for targeted actions and decision-making, i.e. analysis of possible actions and choosing the action that best fits with the system's purpose;

2) Reverse interpretation of the decision made, i.e. forming of the working algorithm for system reaction;

3) Implementation of the system reaction, the consequence of which is changing the external situation and the internal state of the system.

The central subsystem of the IDSS is the Knowledge Base (KB), which deals with the storage, ordering and management of information about the world. The most important parameter of KB is the quality and completeness of knowledge about the software it sets. The quality of KB depends on the structure and format of knowledge, the way they are presented. A widespread implementation of any technology or technique requires a clear and reasoned standard. In the field of KB development, ontology's have become such a standard. Ontology is called an explicit specification of conceptualization. Formally, an ontology consists of terms (notion, concepts) organized into a taxonomy, their definitions and attributes, as well as the associated axioms and rules of derivation [1]. Today there are three types of ontology's: domain-oriented, task-oriented, and top-level. Software Ontology contains taxonomy of concepts, additional relationships, class instances, and various types of constraints (axioms). Axioms impose semantic constraints on the system of relations. The purpose of task ontology is to make knowledge accessible for reuse. Task ontology's determine the extent to which knowledge is used in the process of logical inference. General ontology describes categories - the concept of the upper level. We build a single ontology that contains three types of ontologies at once. Hierarchically, it looks like this: the general ontology is at the top level of the hierarchy, and the software ontology and tasks are connected to it. This approach allows you to comprehensively consider all tasks within the software. All four of the above knowledge models are used to build ontology's: frames are used to define concepts, semantic networks are used to define relations, axioms are assigned 2nd-order logic, and output systems are built for output rules. A semantic network of frames (concepts) is called a concept graph (CG).

II. FORMULATION OF THE PROBLEM IN GENERAL

In order to manually build a fully linked ontology for particular software, it takes a lot of time and resources. The reason for such costs is that such ontology's must contain tens of thousands of elements to be able to solve the wide range of applications that arise in this software. Thus, manual construction of ontology by a human operator is a long routine process, which, moreover, requires a thorough knowledge of software and an understanding of the principles of ontology construction [2, 3]. Therefore, it is necessary to develop methods and algorithms for automatic construction of the ontology. We believe that an expert person should enter the basic terms and the relationship between them manually into the ontology. We will call this initial ontology the basic one and denote it:

$$O_{base} = \langle C_b, R_b, F_b \rangle, \quad (1)$$

where C_b is a finite set of concepts (concepts, terms) of a subject domain, which is set by ontology O_{base}. R_b is a finite set of relationships between concepts (concepts, terms) of a given domain; F_b is finite set of interpretations functions

978-1-7281-3214-3/20/$31.00 ©2020 IEEE

(axiomatization, restriction), set on the concepts or ratios of ontology O_{base}. Thus, the process of constructing a ontology starts from the moment when it already has some data. Therefore, this process will call the development of the basic ontology and mark:

$$\chi : O_{base} \to O \,. \qquad (2)$$

Ontology is a language of science. The language of science, as a structured scientific knowledge, sets a multi-layered hierarchical formation, which highlights blocks: term system; nomenclature; means and rules of forming a conceptual device and timing. So in terms of construction ontology it is necessary to build its term system O_T and nomenclature O_N. In our approach, basic ontology should exactly include a part of the Terminosystem (Fig. 1), that is $O_B \cap O_T \neq \varnothing$.

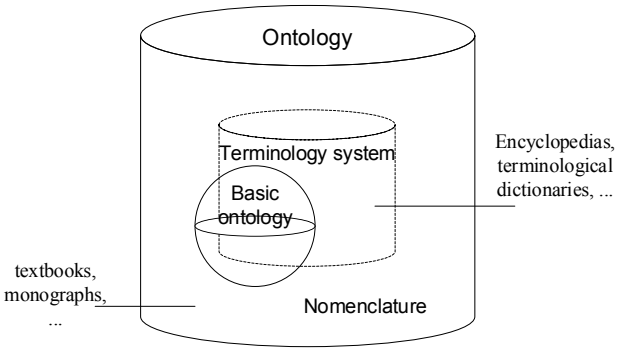


Fig. 1. Architecture of the ontology

Encyclopedias, terminological and model dictionaries, based on which the term system of software is built on, usually have a clear structure and consist of dictionary articles. Therefore, it is necessary to investigate possible structures for recognition of concepts and relations between them. Construction of the nomenclature is harder. If the dictionary terms in some way already allocated, in scientific texts (textbooks, monographs, etc.), they must distinguished, to search for properties of concepts and relations between concepts. We believe that the effectiveness of ontology directly depends on the novelty of knowledge attached to it. Therefore, when building ontology we offer to evaluate its novelty of knowledge. Then base ontology replenished with knowledge of the novelty, which is greater. Back in 1999, the economist Hal Varian speculated that in terms of the economy, "only new information Matters" [4, p. 122].

III. Analysis of Recent Research and Publications

To construct ontology, which adequately describe semantic models of software, it is necessary, first, to solve the problems of extracting knowledge from different sources to identify the plural of concepts and establish hierarchy in this plural. Since much of the information is contained in the natural-language texts (NLT), perspective is the extraction of knowledge from textual information, as well as the intellectual processing of specially selected collections of NLT. There is a number of promising linguistic developments, among which it is advisable to identify a method of lexical-grammatical analysis (Part-of-Speech-tagging), which consists in automatic recognition, to which part of the language belongs to each word in the text. To improve the accuracy of this analysis, two types of algorithms are used: probabilistic-statistical and algorithms based on the product rules that operate in words and codes. As for the latter, they can use rules that are automatically collected from the text body or prepared by skilled linguists. Among the systems developed in Ukraine, it should be noted the development of the department of mathematical informatics of the Taras Shevchenko National University of Kyiv – the system of natural language text processing. The system is designed to solve such problems as analysis and synthesis of natural language texts, automatic generation of abstract text, automatic indexing (theme definition) of the text. The most substantial technical solution in the system is the ability to "weigh" the vertices of the semantic text network. The most important vertices of the network are the vertices, which have the largest number of links with others. This procedure can used in constructing the image of the abstract by weighing vertices and rejecting the easiest-"marginal".

$$\alpha + \beta = \chi \qquad (1)$$

Link Parser Syntax- Semantic analyzer is one of the most effective approaches to automatic text analysis developed and practically implemented at Carnegie Mellon University. The parser uses a priori information about the types of connections that each word can have, with other words positioned in the right sentence and to the left of it, as well as a relatively small number of general grammar rules. The source code of this analyzer is published with the status "open source", which allows you to freely use them to analyze the semantic structure of the text. Semantic analysis is only a preparatory stage for intellectual analysis of the text, the purpose of which is to adopt intellectual system decision on the classification of certain text (recognition) or the existence of some new knowledge that must made in the knowledge base (training).

The assessment of the novelty of knowledge in its broadest sense includes defining any anomalies, deviations from the norm in a given set of data. It has a wide range of applications - from finding neoplasm's (cancer cells) in human body and to detecting of an unfamiliar landscape type by robot. However, in this work we consider only the problem of evaluating the novelty of knowledge in the text fragments. Even with this limitation, the task of finding new information is widely used in modern information services – when creating temporary news reviews, construction of a minimum set of documents in accordance with the request, expert assistance in the analysis of large amounts of text documents that contain duplicates. For example, when the doctor analyzes the medical history of the patient, he examines information about multiple examinations with the most similar results. Usually, the main thing that interests him is the emergence of new symptoms, changes in the patient's health, rejecting the test results from normal or previous status, etc., i.e. "new" data. Let's define the concept of novelty: *Novelty, or new information – these are new answers to potential questions in the context of the user request.* Note that in the definition of the taken phrase "in the context of a user request", that is, the search for new data is carried out taking into account certain scope of user engagement. This is very important because 2 different sentences will often contain some new information relative to each other. Here is an example:

- "The conference was attended by 200 scientists"
- "The conference was attended by 200 young scientists"

If no context is defined, it is clear that the second sentence contains new data concerning the first (the fact that the participants of the Conference are young). However, if a user's request sounds like "Number of conference participants," it becomes clear that both sentences repeat each other in this context. New information may searched at the level of individual events or individual sentences. In the first case, the event we understand the document, article, news, etc. – that is, a certain piece of information defined within the system, which is interpreted as a single entity. New information may searched at the level of individual events or individual sentences. In the first case, the event we understand the document, article, news, etc. – that is, a certain piece of information defined within the system, which is interpreted as a single entity. In this case, the fragment of the data may divided into parts in the analysis process, but the conclusion on novelty is taken for all slices. For example, a news aggregator may analyze individual articles at sentence level, but as a result, it makes a decision on whether to show the entire article to the user or not. In the case of searching novelty at the level of individual sentences, any piece of text splits into smaller indivisible parts bounded by the boundaries of one sentence. The result of search for new knowledge is a set of sentences. This approach is considered in this work.

IV. Formation of Goals

The purpose of the work is to build methods and algorithms for evaluating the novelty of knowledge contained in the linguistic texts when developing an automated ontology. Based on the purpose of the work, define the object of the study – the process of analyzing large text fragments with a high level of redundant (duplicated) information to highlight only unique content in a particular context. Subject of research is flexible algorithms of detection of new knowledge in the text fragments, capable of self-learning based on user's review.

V. Analysis of the Obtained Results

The knowledge evaluation subsystem developed within this work uses the learning algorithms based on the user's feedback. Self-learning as an approach to solving artificial intelligence problems is not in itself fundamentally new, but in the field of definition of novelty, knowledge is almost not used. This may be because system learning often requires experts and is quite costly in economic terms. With consideration of this aspect, we have used a very simple way of user feedback – the subsystem only receives data about the sentences the user has determined. Within an external information system, this can implemented very transparently to the user – fragments of information defined by the subsystem of analysis of the novelty as duplicated can displayed in a summary form with the possibility of deploying a particular block of text. In the case of its deployment, this fact is transmitted back to the novelty detection subsystem to change its parameters. Thus, system education can done in the mode of normal operation, without the need to involve experts. Initial data of the system are obtained based on Acquaint Collection knowledge base analysis. Evaluating the effectiveness of a novelty detection system is a very subjective indicator, depending on the expert using the system. The expert (or system user) determines which sentences, in his opinion, carry "new" information. This estimate will change from one user to another. However, with the averaged evaluation of the developed prototype after the initial training can concluded that the accuracy of detecting the novelty of the developed subsystem is about 70%. This is a fairly high indicator among the results of modern research (for comparison, the algorithm- the winner of the competition within TREC Novelty Track in 2004 year reached of accuracy level 60-65%). This indicates the prospects for the chosen approach. Automatic determination of the semantic similarity of sentences is not an easy task. When a person analyzes the read text, it uses associative thinking, ability to abstract and draw parallels, previous experience, language vocabulary features, etc. – processes that are extremely difficult to simulate on a computer. If a computer program can make a definite conclusion about the similarity of the words "animal" and "animals" (based on the knowledge of lexical rules for the formation of a set of concepts in a particular language), then to estimate the similarity of terms "animal" and "dog" is much more difficult. In addition, for the person it is obvious that the "dog" is a subspecies of "animals", that is, the terms are interconnected by a communication type IS-A. To solve the problem of semantic associations, we will use WordNet-a lexical base of English language [5]. Within this database, each wordform has a set of synonyms (sunsets), word usage frequency and other lexical data. The greatest value of this system for us is the ability to define semantic relationships between words that are not interconnected by a common roots but belong to the same entity class. These ties are characterized by semantic distance. Knowledge evaluation subsystem is focused on the processing of information only in English. This is due to the fact that for English language there is powerful lexical databases, one of which – WordNet – is used in the developed system for definition of semantic distances between sentences. WordNet is a lexical database of the English language, developed within the Princeton University under the direction of George Miller. This database is free to download and after installation the file system, can search. To search using Java tools there is a JAWS library (Java API for WordNet Searching). It provides a simple and intuitive interface to operate lexical data. The distance between the parent and child node in WordNet is calculated as:

$$Dist(c, p) = IC(c) - IC(p), \tag{3}$$

$$IC(c) = -\log P(c) = -\log\{\frac{\sum freq(w)}{N}\}, \tag{4}$$

where $w \in c*$, where $c*$ is set of all hyponyms (subspecies of a word in meaning) c.

Semantic distance between two wordforms in WordNet:

$$Dist(c_1, c_2) = IC(c_1) + IC(c_2) - 2 * IC(LSuper(c_1, c_2)), \tag{5}$$

where $LSuper(c_1, c_2)$ is the closest hyponyms of the two word forms c_1 and c_2.

A word can have different word forms and, accordingly, different sets of synonyms, hyponyms and hyperons. In this case, we choose the most commonly used word form. Based

on this definition of semantic spacing between words, we will formulate a method for calculating the semantic distance between sentences. To do this, we introduce the notion of a minimum distance between a specific word *w* and all other words in a certain sentence *S*: WSSD (Word-Sentence Semantic Distance):

$$WSSD(w,S) = \min\{Dist(w,w_i) \mid w_i \in S\}, \qquad (6)$$

Based on the minimum distance of a certain word and all other words in a sentence, give a definition of the semantic distance between the two sentences (SSD – Sentence Semantic Distance):

$$SSD(S_1,S_2) = \frac{\sum_{w_i \in S_1} WSSD(w_i,S_2) + \sum_{w_j \in S_2} WSSD(w_j,S_1)}{|S_1| + |S_2|}, \qquad (7)$$

where $|S_1|$ and $|S_2|$ number of words in first and second sentences, respectively.

The scheme for calculating the semantic distance between sentences is presented in Fig. 2. Here a_n denotes *n*-th word of the first sentence, b_n - *n*-th a word of the second sentence.

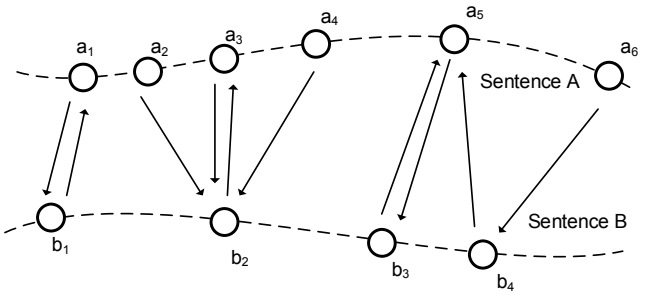


Fig. 2. The scheme of calculating the semantic distance between sentences

The purpose of the work is to develop a flexible subsystem for evaluating novelty knowledge for applying to large volumes of textual information. This subsystem must integrated into the external information system of a certain subject area. Since this subject area is not known, it is important to provide a mechanism for learning, which allows the subsystem to examine the peculiarities of the subject area and to adapt accordingly. In terms of economic efficiency, engaging experts is undesirable. Therefore, another requirement to the developed subsystem is ensuring a simple mechanism for obtaining feedback from the user. System requirements can hierarchically represented as a purpose tree (Fig. 3):

- *Define a custom context*. As noted above, in determining the notion of novelty information, finding new data takes place in a certain context – otherwise, each different of the earlier selected sentences can be considered new (even if it does not provide a user with answers to potential questions). Context can be a user request (if it is supported in the information System). Otherwise, the context is automatically formed based on the most used named entities and terms in the set of text fragments.

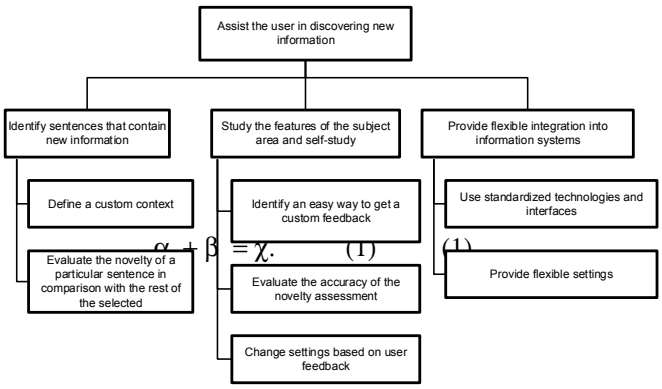


Fig. 3. Targets tree of the subsystem

Consider them more:

- *To assess the novelty of a particular sentence compared to the rest of the selected*. This is the central purpose of the system, which is to determine the level of the novelty of a particular sentence and to form a minimum set of sentences, excluding duplicates.
- *To study the peculiarities of the subject area and self-study*. Since the developed subsystem will designed for integration into various information systems, it is necessary to provide a mechanism for adapting this subsystem to the requirements of a specific area of use.
- *Define an easy way to get a custom review*. Education of the subsystem of estimation of novelty of knowledge with the involvement of experts is economically disadvantageous, and therefore one of the purposes of the system is to provide an easy feedback mechanism, which would allow conducting system training transparently in the process of its use.
- *Assess accuracy of novelty evaluation*. To make a correction of your parameters in the learning process, the subsystem must have a mechanism for evaluating its effectiveness. This allows you to define the coefficients of change parameters during self-learning.
- *Use standardized technologies and interfaces*. To ensure a flexible integration, the subsystem must provide a clearly documented and understandable API (Application Programming Interface), and that in turn should use standardized channels and formats of communication. This will minimize the effort in the integration of the subsystem.

Based on the purposes described above, we will define the main modules of the subsystem:

- Knowledge novelty assessment module.
- The module of training and correction of system parameters.
- Context definition module.
- User interaction module.

Fig. 4 shows DFD-developed subsystem diagrams.

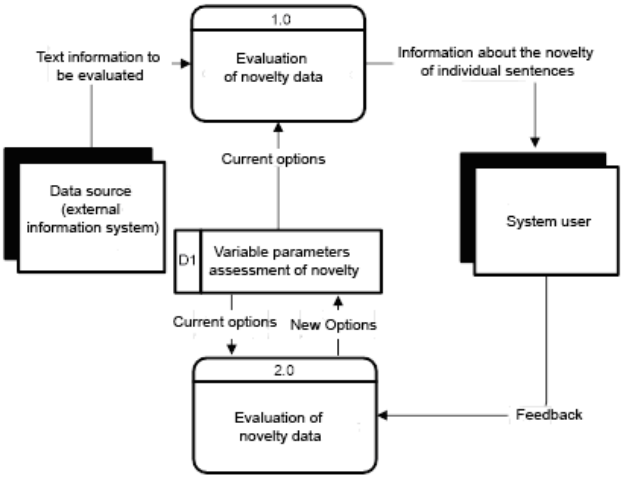


Fig. 4. DFD diagram of the first level

To determine the novelty of the information and to adjust the system parameters it is necessary to analyze the named entities. Named entities – are the answers to questions about the DATE ("When?"), PLACE ("Where?"), PERSON ("Who?"), ORGANIZATION ("Who/What?"), and QUANTITY ("How much?"). Within the developed algorithm, we select 5 types of named entities – terms, which represent one of 5 concepts: the person, place, organization, time or quantity. Identifying these words is necessary for two purposes:

- Markup of duplicated sentences in order to facilitate their analysis for the user. Depending on the specific information system that is embedded in the assessment subsystem of the novelty of knowledge, the information considered duplicate may be displayed to the user as a short summary, so that the user can cast this piece with an eye and evaluate, is he interested in reading it. Because named entities typically carry a much larger meaning than the remaining words in a sentence, they are conveniently used in auto summarizing duplicate pieces of information.
- The number of named entities belonging to the current context in the sentence is as one of the options for defining the novelty of the sentence. The statistical way is shown, that sentences with new ones (those that have not met before) of nominal entities are six times more often [6, p. 247] carry new information in comparison with sentences without named entities.

Because the named entities analysis is only a helper in a developed system, we do not consider the detailed detection algorithm for these terms. The system used the development of BBN, Identifier [7], which is capable of input text:

Jim bought 300 shares of Acme Corp. in 2006.

Convert to the following layout:

<enamex type = "person">Jim</enamex> bought <numex type "quantity">300</numex> shares of < enamex type = "organization">Acme Corp.</ Enamex > in <timex type = "date">2006</timex>

This format is defined by the MUC (Message Understanding Conferences). As was determined in the purpose of this work, considerable attention in the development of the detection subsystem for new data is paid to the flexibility of integration and the ability of the subsystem to self-education in the process [8-12]. Different subject areas require different approaches to defining new information. For example, in the field of medicine threshold of novelty data can be much lower than, say, in the field of news. This is because for a doctor can be an important even slight change in the symptoms of the patient, while in the analysis of news the slight difference in tone coverage of events is not very important. The basic parameters of the system, subject to change based on user feedback, were highlighted [13-24]:

- Impact of the number of named entities in the sentence to the level of novelty
- Weight of each named entity type for the novelty of a sentence
- Threshold of novelty (minimum semantic distance between sentences)
- Context impact on novelty data assessment
- Effect of chronological sequence on evaluation of novelty data

As previously mentioned, the developed subsystem has a very simple way to get feedback from the user – actually, the only information that is fed to the subsystem on the correctness of its operation is the data about the sentences that the user actually scanned. When you receive information that the system mistakenly identified the sentence as a duplicate, it is not clear which settings should changed for work that is more accurate. We use the application accuracy estimate for the current moment (assuming that the user has already read a certain amount of information) to choose the option for change. Thus, for each feedback received from the user, we isolate the change of each parameter and make a reassessment of the accuracy. In addition, the change, which gives the best increase in the accuracy of novelty evaluation, applies. Other options remain unchanged. The algorithm of system parameters correction is shown in Fig. 5. Within the scope of this work, a flexible subsystem for the novelty assessment of knowledge is developed in order to further integrate it into external information systems. This subsystem is developed using a modular approach and has a flexible API (Application Programming Interface). These features provide a convenient integration, but it is clear that the subsystem itself does not have a user interface. In order to assess the efficiency of the selected approach, a prototype of an information system is built, in which the estimation of the novelty of knowledge is embedded. To ensure flexible integration of the subsystem a necessary requirement is to use standardized data formats when interacting with surrounding information system. Therefore, the entire data exchange developed subsystem to the environment takes place in the format XML.

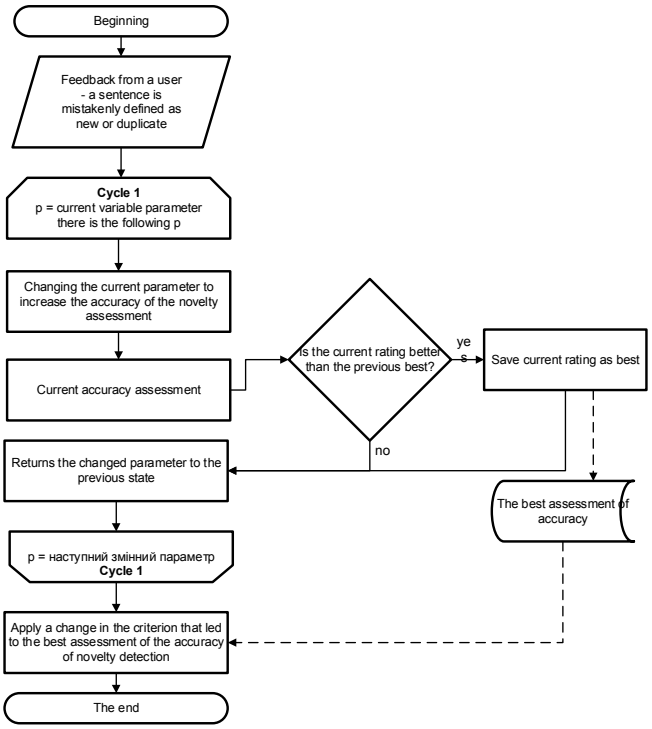


Fig. 5. Algorithm for correction of system parameters

VI. CONCLUSIONS

Evaluating the effectiveness of a novelty detection system is a very subjective indicator, depending on the expert using the system. The expert (or system user) determines which sentences, in his opinion, carry "new" information. This estimate will change from one user to another. However, with the averaged evaluation of the developed prototype after the initial training can concluded that the accuracy of detecting the novelty of the developed subsystem is about 70%. This is a fairly high figure among the results of modern research (for comparison, the algorithm-the winner of the competition within TREC Novelty Track in 2004 reached of accuracy Level 60-65%). This indicates the prospects for the chosen approach. Within the framework of this study, the possibility of introducing the notion of a separate clause in the definition of novelty is analyzed. In our opinion, the use of this approach in conjunction with used algorithms will help increase the accuracy of identifying new knowledge and it is in this direction that research should continue.

REFERENCE

[1] V. Lytvyn, V. Vysotska, D. Dosyn, and Y. Burov, "Method for ontology content and structure optimization, provided by a weighted conceptual graph," Webology, vol. 15(2), 2018, pp. 66-85.

[2] V. Lytvyn, V. Vysotska, D. Dosyn, O. Lozynska, and O. Oborska, "Methods of Building Intelligent Decision Support Systems Based on Adaptive Ontology," IEEE 2nd International Conference on Data Stream Mining and Processing, DSMP, 2018, pp. 145-150.

[3] V. Lytvyn, Y. Burov, P. Kravets, V. Vysotska, A. Demchuk, A. Berko, Y. Ryshkovets, S. Shcherbak, and O. Naum, "Methods and Models of Intellectual Processing of Texts for Building Ontologies of Software for Medical Terms Identification in Content Classification," CEUR Workshop Proceedings, Vol-2362, 2019, pp. 354-368.

[4] H. Varian, "Economics and Search," California: ACM Press, 1999.

[5] Local source annotation with WordNet, www.cogsci.princeton.edu/~wn.

[6] L. Xiaoyan, and W. Croft, "Novelty Detection Based on Sentence Level Patterns", Bremen, 2005.

[7] D.M. Bikel, R.L. Schwartz, and M. Ralph, "An Algorithm that Learns What's in a Name," N.Y.: Columbia University Press, 1999.

[8] Y. Burov, V. Vysotska, and P. Kravets: "Ontological approach to plot analysis and modeling," CEUR Workshop Proceedings, Vol-2362, 2019, pp. 22-31.

[9] S. Sachenko, S. Rippa, and Y. Krupka, "Pre-Conditions of Ontological Approaches Application for Knowledge Management in Accounting," International Workshop on Antelligent Data Acquisition and Advanced Computing Systems: Technology and Applications, 2009, pp. 605-608.

[10] M. Davydov, and O. Lozynska, "Mathematical method of translation into Ukrainian sign language based on ontologies," Advances in Intelligent Systems and Computing, vol. 871, 2018, pp. 89-100.

[11] S. Babichev, "An Evaluation of the Information Technology of Gene Expression Profiles Processing Stability for Different Levels of Noise Components," Data, vol. 3 (4), 2018, art. no. 48.

[12] S. Babichev, B. Durnyak, I. Pikh, and V. Senkivskyy, "An Evaluation of the Objective Clustering Inductive Technology Effectiveness Implemented Using Density-Based and Agglomerative Hierarchical Clustering Algorithms," Advances in Intelligent Systems and Computing, vol. 1020, 2020, pp. 532-553.

[13] I. Khomytska, V. Teslyuk, A. Holovatyy, and O. Morushko, "Development of methods, models, and means for the author attribution of a text," Eastern-European Journal of Enterprise Technologies, vol. 3(2-93), 2018, pp. 41-46.

[14] I. Khomytska, and V. Teslyuk, "Authorship and Style Attribution by Statistical Methods of Style Differentiation on the Phonological Level," Advances in Intelligent Systems and Computing III, vol. 871, Springer, 2019, pp. 105-118.

[15] A.Y. Berko, and K.A. Aliekseyeva, "Quality evaluation of information resources in web-projects," Actual Problems of Economics, vol. 136(10), 2012, pp. 226-234.

[16] O. Cherednichenko, N. Babkova, and O. Kanishcheva, "Complex Term Identification for Ukrainian Medical Texts," CEUR Workshop Proceedings, Vol-2255, 2018, pp. 146-154.

[17] N. Sharonova, A. Doroshenko, and O. Cherednichenko, "Issues of fact-based information analysis," CEUR Workshop Proceedings, Vol-2136, 2018, pp. 11-19.

[18] N. Antonyuk, L. Chyrun, V Andrunyk., A. Vasevych, S. Chyrun, A. Gozhyj, I. Kalinina, and Y. Borzov, "Medical News Aggregation and Ranking of Taking into Account the User Needs," CEUR Workshop Proceedings, Vol-2362, 2019, pp. 369-382.

[19] L. Chyrun, L. Chyrun, Y. Kis, and L. Rybak, "Automated Information System for Connection to the Access Point with Encryption WPA2 Enterprise," Lecture Notes in Computational Intelligence and Decision Making, 1020, 2020, pp. 389-404.

[20] Y. Kis, L. Chyrun, T Tsymbaliak., and L. Chyrun, "Development of System for Managers Relationship Management with Customers," Lecture Notes in Computational Intelligence and Decision Making, 1020, 2020, pp. 405-421.

[21] L Chyrun., A. Kowalska-Styczen, Y. Burov, A. Berko, A. Vasevych, I. Pelekh, and Y. Ryshkovets, "Heterogeneous Data with Agreed Content Aggregation System Development," CEUR Workshop Proceedings, Vol-2386, 2019, pp. 35-54.

[22] L. Chyrun, Y. Burov, B. Rusyn, L. Pohreliuk, O. Oleshek, A. Gozhyj, I. Bobyk, "Web Resource Changes Monitoring System Development," CEUR Workshop Proceedings, Vol-2386, 2019, pp. 255-273.

[23] A. Gozhyj, L. Chyrun, A. Kowalska-Styczen, and O. Lozynska, "Uniform Method of Operative Content Management in Web Systems," CEUR Workshop Proceedings, Vol-2136, 2018, pp. 62-77.

[24] L. Chyrun, A. Gozhyj, I. Yevseyeva, D. Dosyn, V. Tyhonov, and M. Zakharchuk, "Web Content Monitoring System Development," CEUR Workshop Proceedings, Vol-2362, 2019, pp. 126-142.

IEEE Third International Conference on Data Stream Mining & Processing
August 21-25, 2020, Lviv, Ukraine

Information Processing System for Detection Impurity in Technical Oil Based on Laser

Lubomyr Sikora
ACS Department
Lviv Polytechnic National University
Lviv, Ukraine
https://orcid.org/0000-0002-7446-1980

Natalya Lysa
IST Department
Lviv Polytechnic National University
Lviv, Ukraine
https://orcid.org/0000-0001-5513-9614

Roman Martsyshyn
ACS Department
Lviv Polytechnic National University
Lviv, Ukraine
https://orcid.org/0000-0003-2464-8146

Yulia Miyushkovych
PIT Department
Lviv Polytechnic National University
Lviv, Ukraine
https://orcid.org/0000-0001-5890-2208

Rostyslav Tkachuk
CDCMEGP Department
Lviv State University of Life Safety
Lviv, Ukraine
https://orcid.org/0000-0001-9137-1891

***Abstract* – The article describes the method for constructing a laser control system for the impurities level of technical oil (using the example of transformer oil pollution analysis). In the system under consideration, technical oil acts as a transformer cooling medium, and therefore requires impurities control (to ensure the quality of its cooling properties). Determination of indicators of oil impurities is based on the effect of scattering of the laser beam. Literary sources and authors' research were used to construct the structure of the laser measuring system. The effectiveness of the laser meter is verified by a series of experiments. The experiments took place in two stages. The first stage included the calibration of the photometer (to determine the refractive index of the glass of the photometer). The second stage included a series of experiments in which the values of the reference oil were compared with the value of the experimental contaminated samples.**

Keywords – Measurement, Transformer oil, Laser photometry, Expansion of a resonance laser beam.

I. INTRODUCTION

In the process of power plants operation it is important to observe safe operating modes of all its components. Excessive heating of some elements during operation (eg, power transformers) can cause emergency situations (shutdown of individual units, fires). In systems of cooling of transformers are used special technical (transformer) oils (which act as a cooling environment).

Stability of parameters in the operation of power supply systems is one of the main operational tasks. The stable operation of the system depends on the technical state of all its components (which are part of it). An important components of the stable power plant operation are powerful electro-transformers. To ensure the stable operation of transformers, monitoring of its cooling system (to avoid overheating) is required.

II. PROBLEM TASK REVIEW

Nowadays, optical and physical research is very common. This is due to the ability to perform laser research quickly, locally and without direct contact with the object. It is also possible to allocate a high level of the received information when carrying out laser sounding of the environment. This method of research (optical-physical) is a physical experiment. In such an experiment, all the information is transmitted and received by the laser beam [1-7].

There are two components in the process of processing information during laser scanning [4, 5, 7]:

- receiving and processing the image of laser signal and signal structure recognition (as the basis for the formation of informative features) [5];
- obtaining a laser signal, evaluating its characteristics (energy). Discrimination of the laser signal power center is the basis for detecting the structure of the object (energy and geometric). [4, 5].

Radiation sources (optical) that currently exist cannot provide a clear directional diagram. That is why laser sensing of an object (and processing of a scattered signal to reveal additional information about it) is an effective method for obtaining data about the studied environment (its state). [1, 8-12].

In [1-11] only the basic concepts of methods of analysis based on photometry are considered. Such methods can solve a wide range of problems (control in technical systems, chemical and biophysical reactors). In [5, 7, 19] methods of construction of measuring information systems on the basis of laser are considered. Such systems are designed to manage (and control) processes with high operating temperatures. The effectiveness of laser monitoring systems for a number of technologies (eg chemical processes) is also shown.

I In [5, 12, 19] the parameters of chemical solute (optical) were analyzed. Also in [12,19], the substantiation of spectral methods of analysis of media (photoactive) and photometer principles was carried out. In [5,16] the application of spectroscopy (photoelectron) methods in chemistry is considered. This allows you to get information about, adsorption, structure, quantitative analysis and evaluation of chemical shifts. Methods of designing optoelectronic systems are considered in [12, 14]. Unfortunately, the problem of synthesis of laser-based systems has not been considered.

The effect of high voltage on the oil (on the cooling environment of the transformer) affects its molecular

structure, dielectric instability and leads to the appearance of components of electron and ion conductivity. It stimulates breakdown of isolation and emergency situations.

Control of electrotechnical, thermodynamic and physicochemical parameters in the high voltage zone is difficult to implement by direct measurement methods. This, accordingly, contributes to the search for contactless methods for diagnosing oil conditions. One of the possible effective methods for solving this problem is the laser probing of technical oil samples (transformer oil) in cuvettes.

To perform such monitoring, it is necessary to calibrate the parameters at different laser radiation capacities.

The purpose of the work is to substantiate the methods for controlling the dielectric properties of technical (transformer) oils by the method of laser projection sounding in the process of operation of the transformer. Tasks of the research:

1. To substantiate the physical model of quality control of the transformers cooling environment (oils) .
2. To detect the influence of the laser photon flux on the molecular structure of transformer oils.
3. To develop a method and means of creating laser systems for controlling the quality of transformer oils dielectric characteristics.
4. Identify the factors of changing electrical parameters in working mode.

III. Process of Laser Sensing of the Technological Environment

During the operation of the transformer, it becomes necessary to quickly control the quality of the oil. To solve this problem, a laser-based sample sensing system was developed.

Literary sources [1-12,19] and authors' research [13,15,17,18] were used to construct the structure of the laser measuring system. The block diagram of the laser system is presented in Fig. 1.

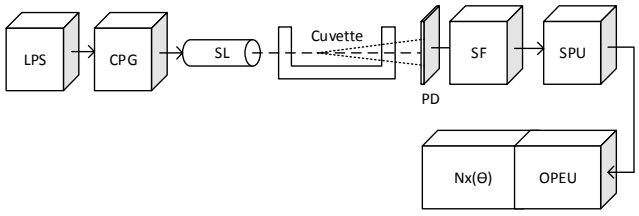


Fig. 2. Scheme of the developed laser system

On Fig. 1: LPS – laser power supply, CPG – clock pulse generator, SL – semiconductor laser, C – cuvette with oil, PD – photodetector, SF – selective filter, SPU – unit for processing signals, OPEU – unit for evaluating the parameters of the oil.

A. Composition of technical (transformer) oil

Transformer oil is a refined oil fraction obtained during distillation, which boils at 300-400 C. Transformer oil has a complex hydrocarbon composition with an average molecular weight. Transformer oil of the brand "Penta TRMS-110" is a heat-resistant silicon-organic compound. Paraffins, cycloparaffins – provide low electrical conductivity and high electrical power [20].

Aromatic carbohydrates increase resistance to partial charges in the volume of oil. Asphalt-resinous compositions are responsible for the appearance of sediment in oil and its color. Sulfur, nitrogen compounds and petroleum acids are responsible for the processes of corrosion of metals in transformer oil.

B. Assessment of the Concentration of Impurities in Technical Oils

During the experimental study of the quality of technical oils, a photometer with two channels was created. The photometer works on the basis of the difference method. Fig.

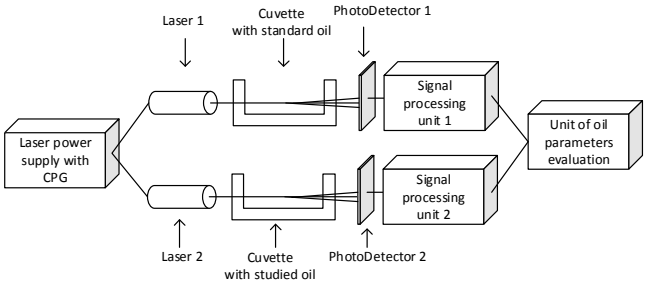


Fig. 1. Scheme of a used laboratory laser system with two channels

2. Represents scheme of a used laboratory laser system with two channels.

Parameters of the experimental cuvette represented on Fig. 3.

The differential comparison procedure (1) was used to calculate the photometer measurement results. This procedure is also called the differential method.

$$\Delta\alpha(C_K,\theta)=\left|\alpha_e(C_K,\theta)-\alpha_i(C_{Ki},\theta_i)\right|\cdot K_F \quad (1)$$

where α_e – scattering coefficient of the reference oil samples, α_i – scattering coefficient control sample oil, KF – photometer ratio.

IV. Results of Experimental Study of Technical Oils Quality by Laser Sensing Method

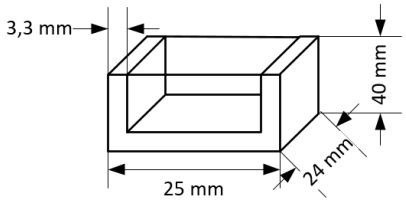


Fig. 3. Experimental unit cuvette dimensions (scheme)

Experimental studies were conducted in several stages. At the first stage, the photometer was calibrated to obtain a photometric coefficient (refractive index of the glass cuvette). The calibration of the system was carried out using a single glass plate and two glass plates.

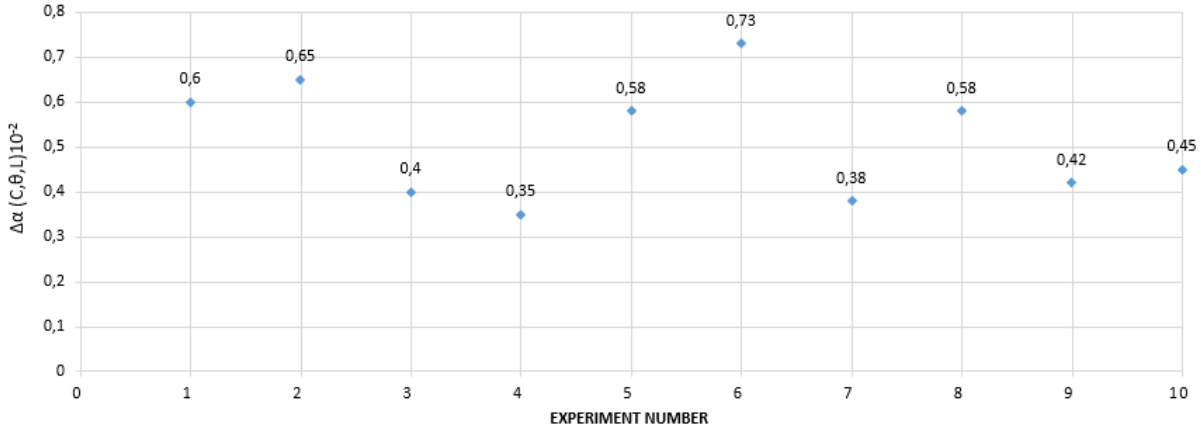


Fig. 4. Photometric calibration results (one glass plate) (X axis represents the number of the experiment; Y axis – represents the value Δα – measured coefficient of scattering)

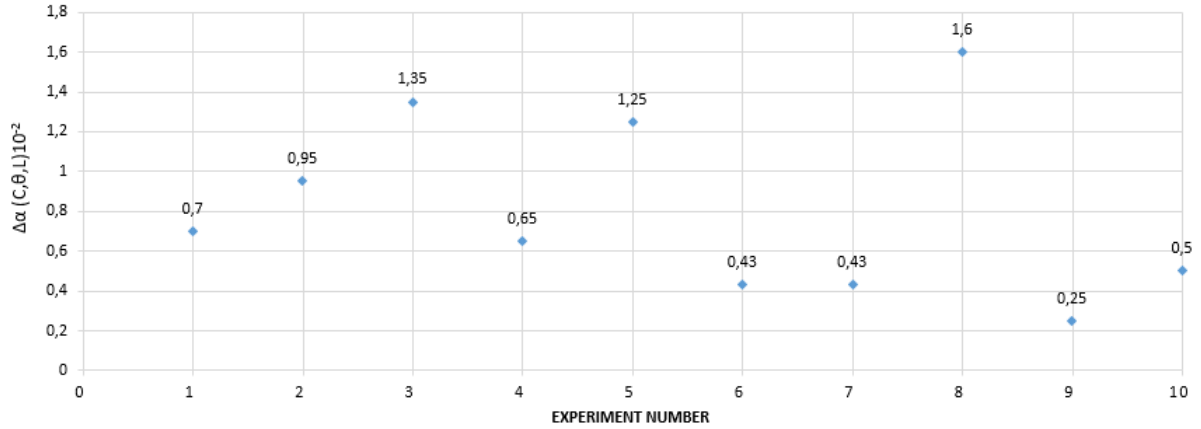


Fig. 5. Photometric calibration results (two glass plates) (X axis represents the number of the experiment; Y axis – represents the value Δα – measured coefficient of scattering)

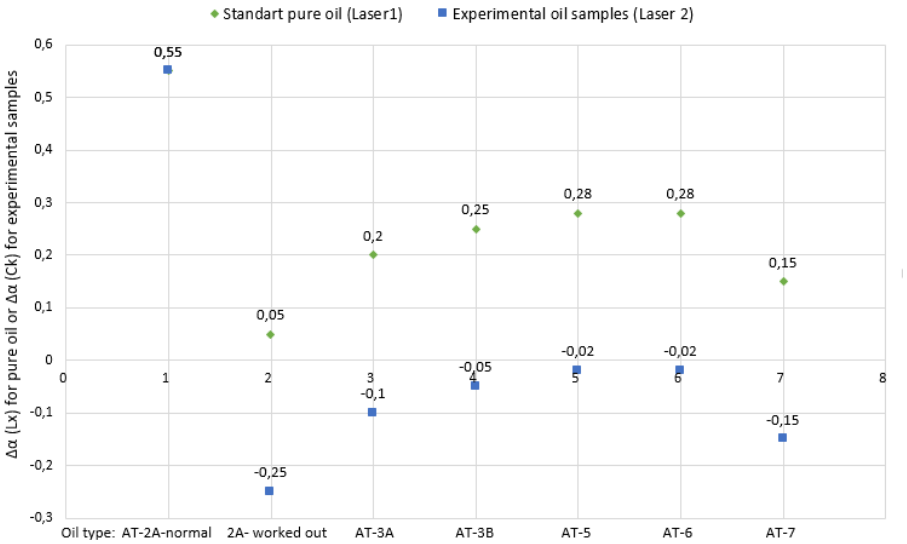


Fig. 6. Graphical representation of the results of experiments №1-7 (X axis represents the number of the experiment (type of oil tested); Y axis – represents the value Δα (Lx) for pure oil or Δα (Ck) for experimental samples)

Fig. 4 shows the results of calibration of the photometer using a single glass plate. Fig. 5 shows the results of calibration of the photometer using a two glass plate. All results were obtained experimentally.

$$\Delta\alpha \cong \alpha_e - \alpha_i \quad (2)$$

where α_e – transmission factor of the standard, α_i – i-sample transmittance.

The main study of experimental oil samples was carried out at the second stage (after the photometer calibration step).

Cuvette №1 was filled with a reference pure model of oil. Various types of oils (AT-2 reference, AT-2 spent, AT-3A, AT-3B, AT-5, AT-6, AT-7) were introduced alternately in the cuvette №2.

In experiment No. 1, both cuvettes contained a reference sample of pure oil.

In experiments №2-7, the cuvette number 1 contained the reference oil, and the cuvette number 2 was the corresponding samples (AT-2 worked out, AT-3A, AT-3B, AT-5, AT-6, AT-7).

In all experiments, the coefficient of refraction of glass (cuvette) Kα0 = + 0.30.

The sign (+) in result determines the active environment that amplifies the laser signal – pure oil; the sign (-) is a loss due to impurities.

Fig. 6 shows a graphical representation of the results of experiments №1-7.

V. CONCLUSION

The considered system of laser monitoring of the technical oil condition enables to carry out research in conditions where standard estimation methods will not be able to carry out the assessment (or the evaluation will be complicated from the technical point of view).

The obtained results show that the system of laser control of technical oils through the method of probing through a cuvette makes it possible to estimate the state of oil (to estimate oil impurities).

The developed system is an important component for monitoring the technological state of aggregates of power-generating objects (in our case - transformers at power plants).

The system of laser control of technical (transformer) oils can help to avoid (prevent) the emergence of an emergency due to excess temperature regime of the transformer during its operation.

REFERENCES

[1] A. Paul, Chemistry of Glasses. New York: Chapman and Hall, 1982.

[2] J. C. Henniker, Infrared Spectrometry of Industrial Polymers. Academic Press, London, New York 1967.

[3] David L. Andrews, Lasers in Chemistry. Springer Science & Business Media, 2012.

[4] Impedance Spectroscopy: Theory, Experiment, and Applications, 2nd ed Edited by Evgenij Barsoukov (Texas Instruments Inc.) and J. Ross Macdonald (University of North Carolina, Chapel Hill). John Wiley & Sons, Inc.: Hoboken, NJ. 2005. xvii + 596 pp. ISBN 0471-64749-7.

[5] A. Donges and R. Noll, Laser Measurement Technology: Fundamentals and Applications, Berlin: Springer, 2015.

[6] A. E. Perkins, P. C. De Oliveira and N. M. Lawandy. "Light propagation in scattering systems with absorption" Technical Digest. Summaries of Papers Presented at the International Quantum Electronics Conference. Conference Edition. 1998 Technical Digest Series, Vol.7 (IEEE Cat. No.98CH36236), San Francisco, CA, USA, 1998, pp. 72-. doi: 10.1109/IQEC.1998.680143

[7] I. Yamaguchi, "Speckle", Optical technological handbook Asakura Publishing Co, pp. 505, 2002.

[8] Akira Ishimaru, Wave Propagation and Scattering in Random Media , IEEE, 1997, doi: 10.1109/9780470547045

[9] R. P. Wayne, Principles and applications of photochemistry. OUP, 1988.

[10] C. X. Wang and Liu Shu-Chang, "Design of embedded laser beam measurement system," in 2009 International Conference on Mechatronics and Automation, Changchun, 2009, pp. 252-256. doi: 10.1109/ICMA.2009.5246349

[11] Instrumentation Reference Book, 4th Edition. by Walt Boyes. Publisher: Butterworth-Heinemann. Release Date: December 2009. 928 pp. ISBN: 9780750683081.

[12] L. Sikora, N. Lysa, B. Fedyna, B. Durnyak, R. Martsyshyn and Y. Miyushkovych, "Technologies of Development Laser Based System for Measuring the Concentration of Contaminants for Ecological Monitoring," 2018 IEEE 13th International Scientific and Technical Conference on Computer Sciences and Information Technologies (CSIT), Lviv, 2018, pp. 93-96, doi: 10.1109/STC-CSIT.2018.8526602.

[13] L. Sikora, N. Lysa, R. Martsyshyn, Y. Miyushkovych, R. Tkachuk and B. Durnyak, "Information Technology of Laser Measurement System Creation for Automated Control Dynamics of Glue Drying in Polygraphy," 2018 IEEE 13th International Scientific and Technical Conference on Computer Sciences and Information Technologies (CSIT), Lviv, 2018, pp. 89-92, doi: 10.1109/STC-CSIT.2018.8526683.

[14] O. Riznyk, N. Kustra, Y. Kynash and R. Vynnychuk, "Method of Constructing Barker-Like Sequence on the Basis of Ideal Ring Bundle Families," 2019 IEEE International Scientific-Practical Conference Problems of Infocommunications, Science and Technology (PIC S&T), Kyiv, Ukraine, 2019, pp. 207-211, doi: 10.1109/PICST47496.2019.9061375.

[15] L. Sikora, R. Martsyshyn, Y. Miyushkovych, N. Lysa and B. Yakymchuk, "Problems of data perception by operators of energy-active objects under stress," The Experience of Designing and Application of CAD Systems in Microelectronics, Lviv, 2015, pp. 475-477, doi: 10.1109/CADSM.2015.7230909.

[16] M. Nazarkevych, Y. Kynash, R. Oliarnyk, I. Klyujnyk and H. Nazarkevych, "Application perfected wave tracing algorithm," 2017 IEEE First Ukraine Conference on Electrical and Computer Engineering (UKRCON), Kiev, 2017, pp. 1011-1014, doi: 10.1109/UKRCON.2017.8100403.

[17] L. Sikora, N. Lysa, R. Martsyshyn, Y. Miyushkovych, Y. Dragan and B. Fedyna, "Technology of Monitoring the Dynamic Displacements of Objects Spatial Structures Using a Laser Surface Sensing System," 2019 IEEE 14th International Conference on Computer Sciences and Information Technologies (CSIT), Lviv, Ukraine, 2019, pp. 9-12, doi: 10.1109/STC-CSIT.2019.8929813.

[18] V. Sheketa, L. Poteriailo, Y. Romanyshyn, V. Pikh, M. Pasyeka and M. Chesanovskyy, "Case-Based Notations for Technological Problems Solving in the Knowledge-Based Environment," 2019 IEEE 14th International Conference on Computer Sciences and Information Technologies (CSIT), Lviv, Ukraine, 2019, pp. 10-14, doi: 10.1109/STC-CSIT.2019.8929784.

[19] E. Golinelli, S. Musazzi, U. Perini and G. Pirovano, "Laser based scanning system for high voltage power lines conductors monitoring," in CIRED 2009 - 20th International Conference and Exhibition on Electricity Distribution - Part 1, Prague, Czech Republic, 2009, pp. 1-3.

[20] S. Missas, M. G. Danikas and I. Liapis, "Factors affecting the ageing of transformer oil in 150 / 20 kV transformers," 2011 IEEE International Conference on Dielectric Liquids, Trondheim, 2011, pp. 1-4, doi: 10.1109/ICDL.2011.6015409.

IEEE Third International Conference on Data Stream Mining & Processing
August 21-25, 2020, Lviv, Ukraine

Investigation on the Processes of Deformation, Heat- and-Moisture Transfer in Media with the Properties of the Effects of "Memory" and Self-Similarity

Yaroslav Sokolovskyy
Department of Information Technologies
Ukrainian National Forestry University, UNFU
Lviv, Ukraine
sokolowskyyyar@yahoo.com

Mariana Levkovych
Department of Information Technologies
Ukrainian National Forestry University, UNFU
Lviv, Ukraine
maryana.levkovych@gmail.com

Olha Mokrytska
Department of Information Technologies
Ukrainian National Forestry University, UNFU
Lviv, Ukraine
mokrytska@nltu.edu.ua

Yaroslav Kaspryshyn
Department of Information Technologies
Ukrainian National Forestry University, UNFU
Lviv, Ukraine
kaspryshyn.ya@gmail.com

Nadiya Yavorska
Department of Ecological Politics and Environment Protection Management
Lviv Polytechnic National University
Lviv, Ukraine
Nadiya.P.Yavorska@lpnu.ua

Abstract. **On the basis of the laws of the mechanics of hereditary media and the properties of the mathematical apparatus of fractional integro-differentiation, the paper investigates the one- and two-dimensional Kelvin rheological model in media with fractal structure. In the two-dimensional Kelvin mathematical model in fractional-differential form, the deformation components in different directions of anisotropy are caused by changes in temperature and moisture content of the material. Thus, one of the objectives of the study in this work is to develop a model of heat-and-moisture transfer, taking into account the inhomogeneity of the material and its effects of "memory". For the mathematical model of heat or moisture transfer, taking into account the fractal structure of the medium, a numerical implementation method, which is based on explicit and implicit finite-difference schemes, has been devised. The difference of implicit finite-difference scheme is shown when using integer differentiation operators and fractional operators. The stability condition for an explicit difference scheme containing fractional derivatives is given. Application software has been developed and the results of its application for the study of deformation processes, heat transfer taking into account the self-similarity and material effects of "memory" have been analyzed.**

Keywords: material inhomogeneity, Kelvin model, fractal structure, self-similarity, deformation.

I. INTRODUCTION

Today, it is relevant to use the mathematical apparatus of fractional integro-differentiation for modeling complex systems and processes. In the works [1], [2], it was shown that the use of non-traditional integro-differentiation is a more effective method. In particular, from the standpoint of the geometric interpretation of fractional operators, it is believed that the structure of materials and the process behaviour cannot be accurately described by known analytical methods. A new approach has emerged - this is fractals, namely, self-similar sets with non-integer dimensions [2]. Examples of such sets are the Koch curve, the sets of Cantor, Julia, Mandelbrot, and others.

Non-integer integro-differentiation apparatus is also widely used for modeling processes of stress-strain state, heat-and-mass transfer [3]-[6]. It is important to take into accoun the effects of "memory" and the self-similarity of a material when modeling deformation or heat transfer processes, or considering the interconnection of deformation processes and heat (moisture) transfer, since this is a step towards improving processes in various fields of science and improving quality of existing materials. Under these conditions, a new class of problems arises for the solution of which it is necessary to apply new modified numerical [7], [8] and analytical methods. Analytical methods include the methods of integral transformation – the Laplace, Mellin, Fourier [9], Sumud methods [10] and others. Relevant numerical methods also include the finite element method [11] and finite difference method [12]-[14].

Given the above, the purpose of this work is to build mathematical models of deformation and heat transfer processes in media with a fractal structure, to adapt numerical methods for development and implementation of software, taking into account the self-similarity and the "memory" effects of the material.

II. MATHEMATICAL MODELS IN MEDIA WITH FRACTAL STRUCTURE

The Kelvin fractional-differential rheological model consists of a series connection of the Voigt body F and the elastic element H (Fig.1). To construct a differential equation that describes the Kelvin model in fractal-structured media, we use Hooke's law

978-1-7281-3214-3/20/$31.00 ©2020 IEEE

$$\sigma(t) = \mathrm{E}\tau^{\alpha}\varepsilon^{(\alpha)}(t), \quad (1)$$

and Newton's law in fractional-differential aspect [1]:

$$\sigma(t) = \mathrm{E}\tau^{\beta}\varepsilon^{(\beta)}(t), \quad (2)$$

where σ is stress, ε is strain, t is time, E is elastic modulus, τ is relaxation time , $\varepsilon^{(\alpha)}, \varepsilon^{(\beta)}$ are fractional derivatives with order α, β in the sense of Riemann-Liouville [2]. If $\alpha = 0$, $\beta = 1$, then relations (1), (2) will describe classical laws of mechanics of hereditary media.

The elements in Kelvin's rheological fractal model are connected in series. And this means that the deformations of these elements are different, and the stresses are accordingly the same. Hence, the total deformation of the Kelvin model is equal to the sum of the deformations of the elastic element and the Voigt body:

$$\varepsilon(t) = \varepsilon_F(t) + \varepsilon_H(t). \quad (3)$$

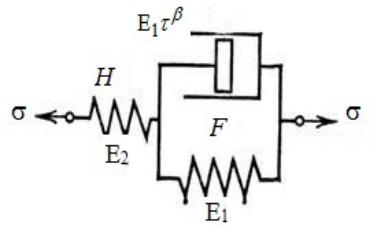


Fig.1. Mechanical diagram of the Kelvin fractional-differential model

We describe the deformation $\varepsilon_H(t)$ through the classical Hooke law (1) and from expression (3) we obtain:

$$\varepsilon_F(t) = \varepsilon(t) - \varepsilon_H(t) = \varepsilon(t) - \sigma(t)/\mathrm{E}_2. \quad (4)$$

We substitute expression (4) into the classical rheological equation of Voigt [15]:

$$\frac{\partial \varepsilon_F(t)}{\partial t} = \varepsilon_F^{(1)}(t) = \sigma(t)/\mathrm{E}_1\tau - \frac{1}{\tau}\left(\varepsilon(t) - \sigma(t)/\mathrm{E}_2\right). \quad (5)$$

After the corresponding mathematical transformations of expression (5) and replacement of the differentiation operators as follows: $\tau\sigma^{(1)}(t) \Rightarrow \tau^{\alpha}\sigma^{(\alpha)}(t)$, $\tau\varepsilon^{(1)}(t) \Rightarrow \tau^{\beta}\varepsilon^{(\beta)}(t)$, $0 < \alpha, \beta < 1$, we obtain a one-dimensional Kelvin model described by the fractional-differential equation:

$$E_1\tau^{\alpha}\sigma^{(\alpha)}(t) + (E_1 + E_2)\sigma(t) = \\ = E_1E_2\left(\varepsilon(t) + \tau^{\beta}\varepsilon^{(\beta)}(t)\right). \quad (6)$$

Taking into account the results of modeling of one-dimensional deformation in media with fractal structure, the bulk stress state for the classical Kelvin model [16], [17] we describe the fractional-differential Kelvin model in the two-dimensional domain by the system of equations:

$$\frac{\sigma_{ii}}{\nu_k} + \frac{\mathrm{E}_1\tau^{\alpha}}{\nu_k(\mathrm{E}_1 + \mathrm{E}_2)}\sigma_{ii}^{(\alpha)} = \frac{\mathrm{E}_{ii}(\varepsilon_{ii} - \varepsilon_{Ti})}{\nu_k(1 - \nu_1\nu_2)} + \frac{\mathrm{E}_{ii}(\varepsilon_{kk} - \varepsilon_{Tk})}{(1 - \nu_1\nu_2)} + \\ + \frac{2\mathrm{E}_1\mathrm{E}_2\tau^{\beta}}{\nu_k(\mathrm{E}_1 + \mathrm{E}_2)}\left(\varepsilon_{11}^{(\beta)} - \varepsilon_{T1}^{(\beta)} + \varepsilon_{22}^{(\beta)} - \varepsilon_{T2}^{(\beta)}\right), \quad (7)$$

$$\sigma_{12} + \frac{\mathrm{E}_1\tau^{\alpha}}{(\mathrm{E}_1 + \mathrm{E}_2)}\sigma_{12}^{(\alpha)} = \mu(\varepsilon_{12} - \varepsilon_{T3}) + \frac{\mathrm{E}_1\mathrm{E}_2\tau^{\beta}}{(\mathrm{E}_1 + \mathrm{E}_2)}\left(\varepsilon_{12}^{(\beta)} - \varepsilon_{T3}^{(\beta)}\right), \quad (8)$$

where $i = 1, 2,$ if $i = 1 \Rightarrow k = 2$ and if $i = 2 \Rightarrow k = 1,$ $\sigma_{11}, \sigma_{22}, \sigma_{12}, \varepsilon_{11}, \varepsilon_{22}, \varepsilon_{12}$ are components of stresses and strains, respectively which depend on time t and spatial coordinates x, y; $\varepsilon_{T1}, \varepsilon_{T2}, \varepsilon_{T3}$ are strains which are caused by a change in temperature ΔT and moisture content ΔU :

$$\varepsilon_{Ti} = \alpha_{ii}\Delta T + \beta_{ii}\Delta U, \ \varepsilon_{T3} = 0, \ i = 1, 2, \quad (9)$$

$\alpha_{11}, \alpha_{22}, \beta_{11}, \beta_{22}$ are coefficients of thermal expansion and moisture-dependent shrinkage; $\mu(T,U)$ is shear modulus, $E_{11}(T,U), E_{22}(T,U)$ are Young's Moduli, $\nu_1(T,U), \nu_2(T,U)$ are Poisson's coefficients, α, β are fractional indices of time derivatives $(0 < \alpha, \beta < 1)$.

The values of the temperature T the moisture content U, which are included in the deformation components and some rheological characteristics can be obtained from the mathematical model of heat (moisture) transfer taking into account the fractal structure of the medium. The heat transfer model is described by the fractional-order partial differential equation in time t and spatial coordinate x :

$$c\rho D_t^{\alpha'}T(t,x) = \lambda D_x^{\beta'}T(t,x) + g(t,x), \quad (10)$$

with the initial condition:

$$T(0,x) = \psi(x), \quad (11)$$

and the boundary conditions of the third kind:

$$\lambda D_x^{\gamma}T(t,0) = \alpha_0^*\left(T(t,0) - t_c\right), \\ \lambda D_x^{\gamma}T(t,a) = \alpha_a^*\left(T(t,a) - t_c\right), \quad (12)$$

The mathematical model of moisture transfer with initial and boundary conditions is written as follows:

$$c\rho D_t^{\alpha'}U(t,x) = aD_x^{\beta'}U(t,x) + f(t,x), \quad (13)$$

$$U(0,x)=\varphi(x), \tag{14}$$

$$\begin{aligned} aD_x^{\gamma}U(t,0)&=\beta_0^*\left(U(t,0)-U_p\right),\\ aD_x^{\gamma}U(t,a)&=\beta_a^*\left(U(t,a)-U_p\right), \end{aligned} \tag{15}$$

where $(t,x)\in D$, $D=\left[0,\tilde{T}\right]\times[0,l]$, c is specific heat capacity, ρ is density, λ is coefficient of thermal conductivity, α_0^*,α_a^* are heat-exchange coefficients, t_c is medium temperature value, a is moisture conductivity coefficient, β_0^*,β_a^* are moisture exchange coefficients, U_p is relative humidity value, $f(t,x), g(t,x)$ is internal heat flux, $\psi(x)$ is initial material temperature value, $\varphi(x)$ is the initial value of the moisture content of the material, $D_t^{\alpha'}, D_x^{\gamma}$ are fractional derivatives by time and spatial coordinate, respectively, in the sense of Riemann-Liouville, $D_x^{\beta'}$ is a fractional derivative by the spatial coordinate in the sense of Grunwald-Letnikov $\left(0<\alpha',\gamma\le 1, 1<\beta'\le 2\right)$.

III. NUMERICAL METHOD OF IMPLEMENTATION OF THE HEAT-AND-MASS TRANSFER MODEL CONSIDERING THE FRACTAL STRUCTURE OF THE MATERIAL

To do this, we will introduce a grid with step h by spatial coordinate x and step $\Delta\tau$ by time t in the region D.

Taking into account the formulas of Grunwald-Letnikov and Riemann-Liouville [1], the resulting finite-difference approximations of fractional derivatives [3] are applied to model (10) - (12) and we obtain:

$$c\rho\frac{T_n^{k+1}-\alpha' T_n^k}{\Gamma(2-\alpha')\Delta\tau^{\alpha'}}=\frac{\lambda}{h^{\beta'}}\sum_{i=0}^{n}q_i T_{n-i+1}^{\omega}+g_n^k \tag{16}$$

$$T_n^0=\psi_n, \tag{17}$$

$$\begin{aligned} \lambda\frac{T_2^k-\gamma T_1^k}{\Gamma(2-\gamma)h^{\gamma}}&=\alpha_0^*\left(T_1^k-t_c\right),\\ \lambda\frac{T_N^k-\gamma T_{N-1}^k}{\Gamma(2-\gamma)h^{\gamma}}&=\alpha_a^*\left(T_N^k-t_c\right), \end{aligned} \tag{18}$$

where $q_0=1, q_i=(-1)^i\frac{\beta(\beta-1)..(\beta-i+1)}{i!}$, if $\omega=k+1$, then we get an implicit difference scheme, at $\omega=k$ - an explicit difference scheme.

One of the features of the obtained mathematical model in the finite-difference approximation (16)-(18) is that in the implementation of the implicit scheme we obtain a matrix of a completely different form than when applying the approximation of integer derivatives:

$$\begin{pmatrix} C_1'+\gamma & -1 & 0 & 0 & ... & 0\\ Wq_2 & (Wq_1-1) & Wq_0 & 0 & ... & 0\\ Wq_3 & Wq_2 & (Wq_1-1) & Wq_0 & ... & 0\\ & & ... & & & \\ Wq_{N-1} & Wq_{N-2} & Wq_{N-3} & ... & (Wq_1-1) & Wq_0\\ 0 & 0 & 0 & ... & \gamma & C_M-1 \end{pmatrix} \tag{19}$$

where $C_1'=\frac{\xi_0}{\varsigma}\Gamma(2-\gamma)h^{\gamma}$, $C_M=\frac{\xi_l}{\varsigma}\Gamma(2-\gamma)h^{\gamma}$,

$W=\frac{\varsigma\Gamma(2-\alpha')\Delta\tau^{\alpha'}}{c\rho h^{\beta'}}$.

Since the explicit finite-difference scheme is unstable, the following convergence condition is obtained, which, unlike the classical one, contains non-integer differentiation operators:

$$\frac{\Gamma(2-\alpha')\Delta\tau^{\alpha'}}{h^{\beta'}}\le\frac{\alpha' c\rho}{\beta'\varsigma}. \tag{20}$$

One important results of the studies is the identification of fractional-differential parameters of mathematical models of strain-stress state [18]. The Prony method and the iterative method were used for this purpose. The application of the iterative method has two stages. At the first stage, assuming that the parameters α,β of the deformation function are integer, based on the least squares method, a search is made for the initial value of stress, relaxation time, and elastic modulus. The identification results obtained at the first stage are taken into account at the next stage, where fractal parameter values are obtained by minimizing expressions describing the creep law for rheological models.The results are refined using the coordinate descent method. For both stages, the objective functions are:

$$\begin{aligned} \sum_{i=1}^{n}\left(\varepsilon_i-\varepsilon_{F,K,M}^{kl}\left(t_i,\tau,\sigma_0,\mathrm{E}\right)\right)^2&\Rightarrow\min,\\ \sum_{i=1}^{n}\left(\varepsilon_i-\varepsilon_{F,K,M}\left(t_i,\alpha,\beta\right)\right)^2&\Rightarrow\min. \end{aligned} \tag{21}$$

IV. APPLICATION SOFTWARE AND RESULTS OF NUMERICAL ANALYSIS

Application software has been developed that allows to take into account different values of fractal parameters when calculating the values of temperature, moisture content, stress and deformation of the material. In the software interface, one part of the rheological, thermophysical internal and external parameters is set in the form of functional dependences, taking into account the initial values of temperature and moisture content, and the second - by specific numerical values. Fig. 2 shows the temperature change taking into account the fractal structure of the material. In particular, the following was established: the influence of the fractal structure on the temperature change is more significant with an increase in the temperature regime of drying; the heterogeneity and self-similarity of the structure of the material slows down the process of its heating; a decrease in the parameters α,β of the model of

heat-and-mass transfer leads to the acceleration of the moisture release process. A numerical experiment was conducted for the Kelvin model and it was found that, regardless of the values of fractional integro-differential parameters of the models, the stresses in the material increase (Fig. 3). It can also be noted that the larger the difference in the deviation of the differentiation parameters from the value - 1, the greater the difference in the values of the deformation (Fig. 4). The Kelvin model consists of only visco-elastic elements, which explains that the deformations under the influence of constant stress will increase.

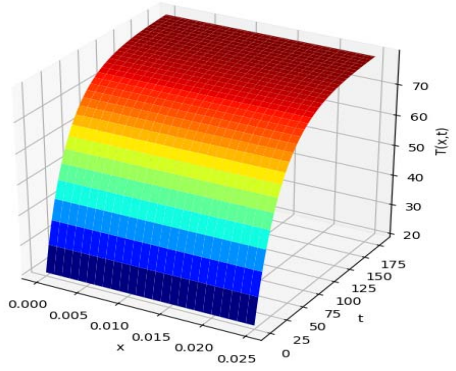


Fig. 2. Temperature dynamics taking into account the fractal structure of the medium

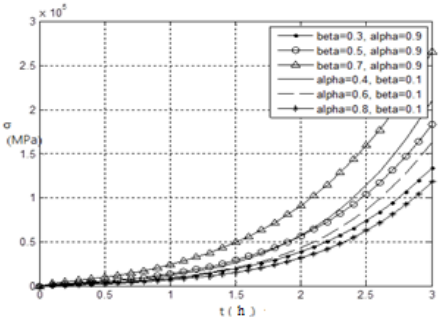


Fig. 3. Stresses at various values of fractional differential parameters

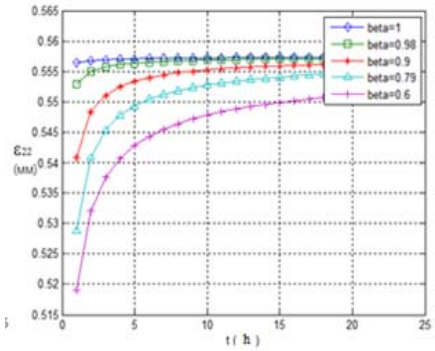


Fig. 4. Deformation in the tangential direction of anisotropy, taking into account the effects of "memory" of the material

CONCLUSIONS

The mathematical Kelvin model in the fractional-differential aspect is investigated and constructed, its mechanical scheme is given. Since it is important to take into account the temperature and moisture state of the material, a mathematical model of heat transfer (moisture transfer) in media with fractal structure is developed. Obtained are finite-difference schemes for a mathematical model of heat (moisture) transfer based on the use of approximations of fractional derivatives. In the work, application software is developed for calculating temperature and moisture content, taking into account the complex structure, heterogeneity, self-similarity and eridarity of the material. Also, analyzed is the stress-strain state of the material with fractal structure.

REFERENCES

[1] I. Podlubny, "Fractional Differential Equations", vol. 198 of Mathematics in Science and Engineering, Academic Press, San Diego, Calif, USA, p. 340, 1999.

[2] V. Uchajkin, "Method of fractional derivatives", Ulyanovsk: Publishing house «Artishok», p. 512, 2008.

[3] Ya.I. Sokolovskyy, M.V. Levkovych, O.V. Mokrytska, and Ya.O. Kaplunskyy, "Mathematical models of biophysical processes taking into account memory effects and self-similarity", Informatics & Data-Driven Medicine, vol. 2255, 2018, pp. 215-228.

[4] V.P. Golub, Yu.V. Pavlyuk, and P.V. Fernati, "On determination of parameters of fractional-exponential kernels of heredity of nonlinear-visco-elastic materials", Applied Mechanics, vol. 49, no. 2, 2013, pp 100-113. [in Russian]

[5] Y. Sokolovskyy, V. Shymanskyi, and M. Levkovych, "Mathematical Modeling of Non-Isothermal Moisture Transfer and Visco-Elastic Deformation in the Materials with Fractal Structure", Proceedings of the 2016 IEEE 11th International Scientific and Technical Conference on Computer Sciences and Information Technologies (CSIT), Lviv, 6-10 September 2016, pp. 91-95.

[6] S. W. J. Welch, R. A. L. Rorrer, and R. G. Duren, "Application of time-based fractional calculus method to viscoelastic creep and stress relaxation of materials", Mech.Time-Dependent Materials, Vol. 3, no. 3, 1999, pp. 279-303.

[7] Q. Yang, F. Liu, and I. Turner, "Numerical methods for fractional partial differential equation with Riesz space fractional derivatives", Applied Mathematical Modelling 34, 2010, pp. 200-218

[8] Y. Sokolovskyy, O. Sinkevych, "Calculation of the drying agent in drying chambers", Proceedings of the 2017 14th International Conference The Experience of Designing and Application of CAD Systems in Microelectronics (CADSM), Lviv, pp. 27-31, 21-25 February 2017.

[9] L. Kexue, P. Jigen, "Laplace transform and fractional differential equations", Appl. Math. Lett, vol. 24, 2011, pp. 2019-2023.

[10] K. Al-Khaled, "Numerical solution of time-fractional partial differential equations using Sumudu decomposition method", Rom. Journ. Phys. Bueharest, vol. 60, no.1-2, 2015, pp. 99-110.

[11] S.V. Boyko, A.M. Eroshenko, "Modeling of physico-mechanical properties of modified wood by finite element method", Technical Sciences and Technologies, no. 2 (4), 2016, pp.184-188. [in Ukrainian]

[12] Y. Sokolovskyy, M. Levkovych, O. Mokrytska, and Y. Kaplunskyy, "Mathematical models of biophysical processes taking into account memory effects and self-similarity", 1st International Workshop on Informatics and Data-Driven Medicine, IDDM-2018, Lviv, 2018, pp. 215-228.

[13] Y. Sokolowskyi, V. Shymanskyi, M. Levkovych, and V. Yarkun: "Mathematical and software providing of research of deformation and relaxation processes in environments with fractal structure", Proceedings of the 12th International Scientific and Technical Conference on Computer Sciences and Information Technologies, CSIT 2017, Lviv, vol. 1, 2017, pp. 24-27

[14] Y. Sokolovskyy, M. Levkovych, O. Mokrytska, and V. Atamanvuk, "Mathematical Modeling of Two-Dimensional Deformation-Relaxation Processes in Environments with Fractal Structure", Proceedings of the 2018 IEEE 2nd International Conference on Data Stream Mining and Processing (DSMP), Lviv, 21-25 August 2018, pp. 375-380.

[15] M. Lavrenyuk "Models of real deformable solids of inhomogeneous media: textbook. tools", Kyiv: Kyiv National University. Taras Shevchenko, 86 p, 2012

[16] J. Bodig, B. A. Jayne, "Mechanics of wood and wood composites", Krieger Publishing Company, 712 p, 1993.

[17] I.A. Birger, R.R. Mavlyutov, "Materials resistance: textbook", M. Nauka, Phys.-Math. Met., p. 560, 1986.[in Russian]

Li D. S, "Identification of Fractal Scale Parameter of Machined Surface Profile", Applied Mechanics and Materials, Vol. 42, 2011, pp. 209-214.

IEEE Third International Conference on Data Stream Mining & Processing
August 21-25, 2020, Lviv, Ukraine

Mathematical Modeling of Non-Isothermal Moisture Transfer Process and Optimization of Geometric Dimensions of the Construction of Composite Materials with Fractal Structure

Volodymyr Shymanskyi
Department of information technologies
National Forestry University of Ukraine
Lviv, Ukraine
vshymanskiy@gmail.com

Ostap Dumanskyi
Department of information technologies
National Forestry University of Ukraine
Lviv, Ukraine
duma.ostap@gmail.com

Yurii Prusak
Department of design
National Forestry University of Ukraine
Lviv, Ukraine
prsk@ukr.net

***Abstract* — The optimal control problem based on the available model of temperature and moisture distribution in the block-block construction was obtained. The optimal wall thickness of the claydite-block construction to provide optimal heat insulating and moisture resistant characteristics was founded. The mathematical model of heat and moisture transfer processes in construction of composite materials with taking into account the fractal structure of them was considered. An implicit and explicit difference schemes for differential equations which descried the mathematical model was obtained. The approximation of the boundary conditions that characterize the interaction of the structure with the external environment was obtained. For numerical solving of consider mathematical model the predictor-corrector finite-difference method was used. Analyzing the obtained data we conclude that selecting a material with an optimal fractal degree can provide the maximum thermal insulation and moisture resistance of the structure. It allows to reduce the wall thickness and the potential cost of the building.**

Keywords — thermal insulation, moisture resistance, claydite, heat and moisture transfer, fractal structure.

I. INTRODUCTION

Construction of industrial and individual buildings often involves the use of expanded clay blocks as the main material. They are characterized by a porous structure and have some advantages over other materials. At the same time, such products are relatively fragile and during production great attention should be paid to respecting the proportions: the lack of binding elements significantly affects the strength of the erected structure.

Therefore, the construction of adequate mathematical models of the heat and moisture conductivity processes in claydite-block constructions with taking into account the complex structure of the materials is relevant today. They will allow you to determine the change of temperature and humidity in the expanded clay unit depending on the characteristics of the environment, inside the building and the physical characteristics of the material. Determining the degree of fractality of the material is closely related to the processing of experimental data, in particular, using artificial neural networks. On the basis of the analysis of the obtained results, it will be possible to give recommendations on the proportions of the ingredients from which the claydite-block is manufactured, in order to increase the operational characteristics of the building.

II. PRODUCTION OF A PROBLEM

Let's consider the dynamic of temperature and moisture in the claydite-block construction of the following form

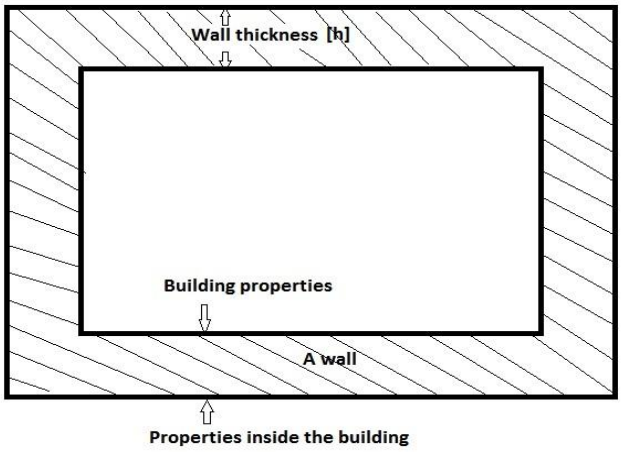


Fig. 1. The scheme of claydite-block construction

The mathematical model of the process of temperature and humidity change in expanded clay-block construction can be written as a system of differential equations of parabolic type with partial derivatives of fractional order. It allows take into account the fractal structure of composite materials [2, 3, 7, 12]:

$$c(T,U)\rho(U)\frac{\partial^{\alpha}T}{\partial\tau^{\alpha}}=\lambda(T,U)\left(\frac{\partial^{2}T}{\partial x_1^{2}}+\frac{\partial^{2}T}{\partial x_2^{2}}\right)+\varepsilon\rho_0 r\frac{\partial^{\alpha}U}{\partial\tau^{\alpha}} \quad (1)$$

$$\frac{\partial^{\alpha}U}{\partial\tau^{\alpha}}=a(T,U)\left(\frac{\partial^2U}{\partial x_1^2}+\frac{\partial^2U}{\partial x_2^2}\right)+a(T,U)\delta\left(\frac{\partial^2T}{\partial x_1^2}+\frac{\partial^2T}{\partial x_2^2}\right) \quad (2)$$

The initial conditions allow us to set the values of temperature and moisture in the structure at the initial time

$$T\big|_{\tau=0}=T_0(x_1,x_2) \quad (3)$$

$$U\big|_{\tau=0}=U_0(x_1,x_2) \quad (4)$$

Boundary conditions of third kind which describing the interaction of the construction with the external and room environment [1-3]

$$-\lambda(T,U)\frac{\partial T}{\partial n}\bigg|_{x_i=l_i}+\rho_0(1-\varepsilon)\beta\left(U\big|_{x_i=l_i}-U^{(R)}_{\text{environment}}\right)=\gamma\left(T_{\text{environment}}-T\big|_{x_i=l_i}\right) \quad (5)$$

$$-\lambda(T,U)\frac{\partial T}{\partial n}\bigg|_{x_i=l_i}+\rho_0(1-\varepsilon)\beta\left(U\big|_{x_i=l_i}-U^{(R)}_{\text{building}}\right)=\beta\left(T_{\text{building}}-T\big|_{x_i=l_i}\right) \quad (6)$$

$$a(T,U)\delta\frac{\partial T}{\partial n}\bigg|_{x_i=l_i}+a(T,U)\frac{\partial U}{\partial n}\bigg|_{x_i=l_i}=\beta\left(U^{(R)}_{\text{environment}}-U\big|_{x_i=l_i}\right) \quad (7)$$

$$a(T,U)\delta\frac{\partial T}{\partial n}\bigg|_{x_i=l_i}+a(T,U)\frac{\partial U}{\partial n}\bigg|_{x_i=l_i}=\beta\left(U^{(R)}_{\text{building}}-U\big|_{x_i=l_i}\right) \quad (8)$$

Where $T(\tau,x_1,x_2)$ - temperature, $U(\tau,x_1,x_2)$ - moisture, τ - time, x_1,x_2 - spatial coordinates, $c(T,U)$ - thermal capacity, $\rho(U)$ - density, $\lambda(T,U)$ - coefficient of thermal conductivity, ε - coefficient of phase transition, ρ_0 - basic density, r - specific heat of steam generation, $a(T,U)$ - coefficient of moisture conductivity, δ - thermogradient coefficient, $T_0(x_1,x_2)$ - initial temperature, $U_0(x_1,x_2)$ - initial moisture value, γ - heat transfer coefficient, β - moisture transfer coefficient, n - outside normal, α - fractional order of derivative (describing the part of the channels open for flow), $T_{\text{environment}}$ - ambient temperature, T_{building} - temperature inside the building, $U^{(R)}_{\text{environment}}$ - equilibrium moisture content on the environment boundary, $U^{(R)}_{\text{building}}$ - equilibrium moisture content on the inside boundary.

Based on the available model of temperature and moisture distribution in the block-block construction, we formulate the optimal control problem, which can be written as follows. Such minimum wall thickness should be found

$$h(T,U,\alpha)\underset{\alpha}{\rightarrow}\min \quad (9)$$

At which the maximum, in terms of modulus, temperature and humidity deviation at the end of modeling time from the initial values at the internal boundary of the construction will not exceed the allowable values, with a given value of the fractal parameter α

$$\max_{(x_1,x_2)\in\Gamma_2}\left|T(\tau,x_1,x_2,h,\alpha)\big|_{\tau=t\_end}-T_0(x_1,x_2)\right|\le\varepsilon_T \quad (10)$$

$$\max_{(x_1,x_2)\in\Gamma_2}\left|U(\tau,x_1,x_2,h,\alpha)\big|_{\tau=t\_end}-U_0(x_1,x_2)\right|\le\varepsilon_U \quad (11)$$

where t_end - the end of modeling time, Γ_2 - internal boundary of the construction, ε_T and ε_U - the maximum deviation of temperature and humidity at the internal boundary of the construction.

III. NUMERICAL METHOD FOR FINDING THE SOLUTION OF THE CONSIDERED MATHEMATICAL MODEL

We define the derivative of fractional order α in the interval $[\tau^n;\tau^{n+1}]$ in the sense of Riemann-Liouville problem [4-6, 8-11].

Using the approximation of derivative of fractional order α [4-6, 13-21] we can obtain an explicit difference scheme for equation (1)–(2) when $s=0$ and implicit difference scheme when $s=1$ [1-3]:

$$c(T^n_{i,j},U^n_{i,j})\rho(U^n_{i,j})\frac{T^{n+1}_{i,j}-\alpha T^n_{i,j}}{\Gamma(1-\alpha)(1-\alpha)\Delta\tau^{\alpha}}=\lambda(T^n_{i,j},U^n_{i,j})\left(\frac{T^{n+s}_{i-1,j}-2T^{n+s}_{i,j}+T^{n+s}_{i+1,j}}{h^2_{x_1}}+\frac{T^{n+s}_{i,j-1}-2T^{n+s}_{i,j}+T^{n+s}_{i,j+1}}{h^2_{x_2}}\right)+\varepsilon\rho_0 r\frac{U^{n+1}_{i,j}-\alpha U^n_{i,j}}{\Gamma(1-\alpha)(1-\alpha)\Delta\tau^{\alpha}} \quad (12)$$

$$\frac{U_{i,j}^{n+1}-\alpha U_{i,j}^{n}}{\Gamma(1-\alpha)(1-\alpha)\Delta\tau^{\alpha}}=$$
$$a\left(T_{i,j}^{n},U_{i,j}^{n}\right)\left(\frac{U_{i-1,j}^{n+s}-2U_{i,j}^{n+s}+U_{i+1,j}^{n+s}}{h_{x_1}^2}+\right.$$
$$\left.\frac{U_{i,j-1}^{n+s}-2U_{i,j}^{n+s}+U_{i,j+1}^{n+s}}{h_{x_2}^2}\right)+$$
$$a\left(T_{i,j}^{n},U_{i,j}^{n}\right)\delta\left(\frac{T_{i-1,j}^{n+s}-2T_{i,j}^{n+s}+T_{i+1,j}^{n+s}}{h_{x_1}^2}+\right.$$
$$\left.\frac{T_{i,j-1}^{n+s}-2T_{i,j}^{n+s}+T_{i,j+1}^{n+s}}{h_{x_2}^2}\right) \quad (13)$$

Let's introduce the following notation:

$$T_{(n+1)}=\begin{cases}\dfrac{T_{i+1,j}^{n+1}-T_{i,j}^{n+1}}{h_{x_1}}, & if \quad n\,\|\,x_1\\ \dfrac{T_{i,j+1}^{n+1}-T_{i,j}^{n+1}}{h_{x_2}} & if \quad n\,\|\,x_2\end{cases} \quad (14)$$

$$U_{(n+1)}=\begin{cases}\dfrac{U_{i+1,j}^{n+1}-U_{i,j}^{n+1}}{h_{x_1}}, & if \quad n\,\|\,x_1\\ \dfrac{U_{i,j+1}^{n+1}-U_{i,j}^{n+1}}{h_{x_2}} & if \quad n\,\|\,x_2\end{cases} \quad (15)$$

The approximation of boundary conditions (5)–(8) will be as follows:

$$-\lambda\left(T_{i,j}^{n+1},U_{i,j}^{n+1}\right)T_{(n+1)}+\rho_0(1-\varepsilon)\beta$$
$$\left(U_{(n+1)}-U_{\text{environment}}^{(R)}\right)=\gamma\left(T_{\text{environment}}-T_{i,j}^{n+1}\right) \quad (16)$$

$$-\lambda\left(T_{i,j}^{n+1},U_{i,j}^{n+1}\right)T_{(n+1)}+\rho_0(1-\varepsilon)\beta$$
$$\left(U_{(n+1)}-U_{\text{building}}^{(R)}\right)=\beta\left(T_{\text{building}}-T_{i,j}^{n+1}\right) \quad (17)$$

$$a\left(T_{i,j}^{n+1},U_{i,j}^{n+1}\right)\delta T_{(n+1)}+a\left(T_{i,j}^{n+1},U_{i,j}^{n+1}\right)U_{(n+1)}=$$
$$\beta\left(U_{\text{environment}}^{(R)}-U_{i,j}^{n+1}\right) \quad (18)$$

$$a\left(T_{i,j}^{n+1},U_{i,j}^{n+1}\right)\delta T_{(n+1)}+a\left(T_{i,j}^{n+1},U_{i,j}^{n+1}\right)U_{(n+1)}=$$
$$\beta\left(U_{\text{building}}^{(R)}-U_{i,j}^{n+1}\right) \quad (19)$$

Using a predictor-corrector finite-difference method we find a numerical solution of a mathematical model (1)–(8). Approximation (12)–(13) will be used as a predictor when $s=0$, and (12)–(13) as a corrector when $s=1$ [1-3].

IV. THE OBTAINED RESULTS

Let's find the optimal wall thickness of the claydite-block construction under such conditions. The temperature inside the building is equal to 21 degrees, outside - 18 degrees. The initial temperature value changes linearly between the specified values. The initial moisture content in the wall is 0.08. After that the ambient temperature dropped from 18 to 12 degrees and ambient moisture has grown to 0.4. The optimal wall thickness of the claydite-block construction with the specified parameters depending on the fractal parameter α is shown in Table 1.

TABLE I. OPTIMAL WALL THICKNESS OF THE CONSTRUCTION

Fractal parameter α	Optimal wall thickness depending on fractal parameter α
1	0.2856
0.98	0.2723
0.96	0.2648
0.94	0.2627
0.93	0.2604
0.92	0.2674
0.9	0.2818

[a] Sample of a Table footnote. (*Table footnote*)

Analyzing the data from Table 1, we can conclude that in order to provide maximum thermal insulation and moisture resistance of the structure, it is necessary to select a material with a fractal degree of 0.3 and to construct a wall with a thickness of not less than 0.2604 meters.

Fig. 2 shows the dependence of temperature on time and wall thickness of the construction.

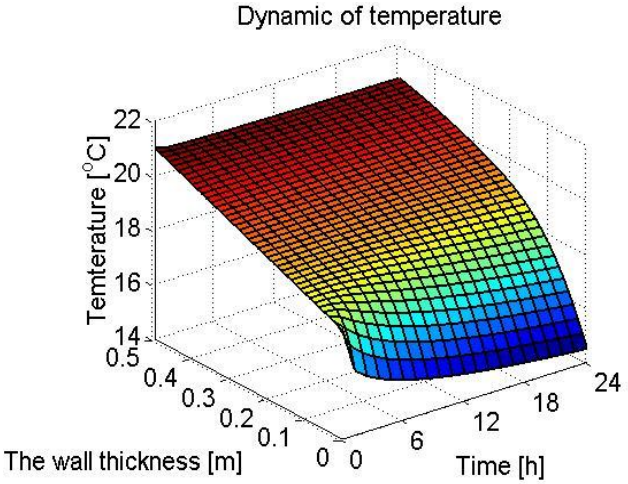


Fig. 2. Temperature depending on time and wall thickness

Analyzing the Fig. 2 we can conclude that the temperature of the wall drops to the ambient temperature from the inside of construction. It remains unchanged inside the structure which indicates a reasonable opportunity to reduce the wall thickness.

Fig. 3 shows the dependence of moisture on time and wall thickness of the structure.

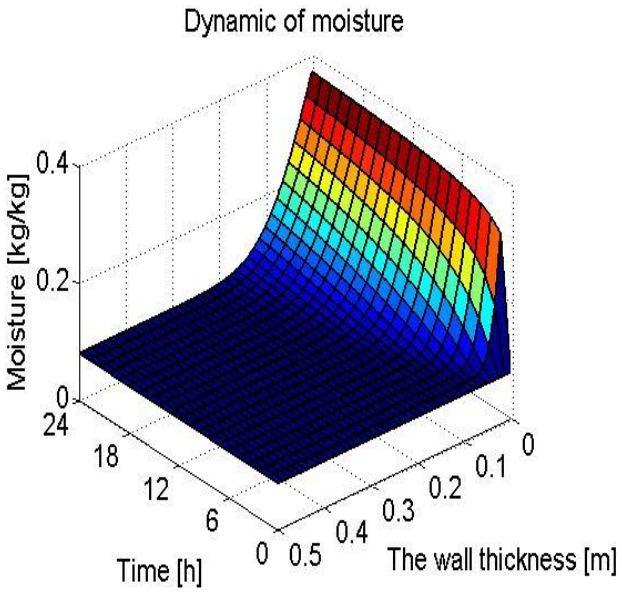


Fig. 3. Moisture depending on time and wall thickness

Having analyzed the Fig. 3 can be concluded that the material gradually absorbs moisture across the border. Over time, humidity spreads deep into the material.

Having analyzed the Fig. 4 and Fig. 5 we conclude that with increasing fractal order of the material it is cooled slowly, that provides better thermal insulation. However, this factor accelerates the process of moisture absorption, which reduces the moisture resistance of the material.

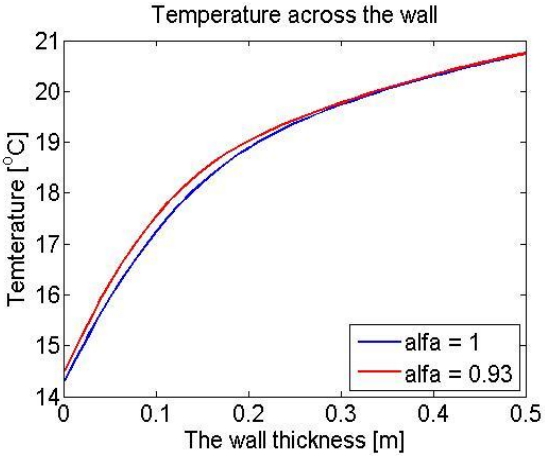


Fig. 4. Temperature depending on wall thickness

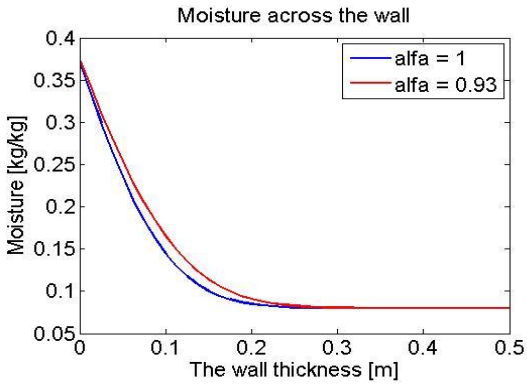


Fig. 5. Moisture depending on wall thickness

Increasing the number of pores, capillaries and level of heterogeneity in the expanded clay block leads to a more pronounced manifestation of the fractal properties of the material.

b. ACKNOWLEDGMENT

Mathematical model and software for optimization of geometrical sizes of constructions of composite materials on condition of providing their heat-insulating and moisture-resistant characteristics are developed. Finite-difference schemes for approximation of differential equations of the mathematical model of nonisothermal moisture transfer in expanded clay-block construction with taking into account the fractal structure of the material are proposed. The fractality parameter was identified by approximating the experimental data using artificial neural networks. Analyzing the obtained data we can conclude that in order to provide maximum thermal insulation and moisture resistance of the structure, it is necessary to select a material with a optimal fractal degree which will also allow to reduce the wall thickness and the potential cost of the building. Increasing the level of fractality of the material leads to improved thermal insulation properties and deterioration of moisture resistance.

c. REFERENCES

[1] Ya. Sokolovskyy, V. Shymanskyi, M. Levkovych and V. Yarkun, “Mathematical and Software Providing of Research of Deformation and Relaxation Processes in Environments with Fractal Structure”, XII-TH International Scientific And Technical Conference "Computer Science And Information Technologies" CSIT-2017, Lviv, Ukraine. pp. 24-27, 2017.

[2] Ya. Sokolowskyi, V. Shymanskyi and M. Levkovych, “Mathematical modeling of non-isotermal moisture transfer and visco-elastic deformation in the materials with fractal structure”, XI-TH International Scientific And Technical Conference "Computer Science And Information Technologies" CSIT-2016, Lviv, Ukraine. pp. 91-95, 2016.

[3] Y. Sokolovskyy, M. Levkovych, O. Mokrytska and Y. Kaplunskyy, “Mathematical models of biophysical processes taking into account memory effects and self-similarity”, 1st International Workshop on Informatics and Data-Driven Medicine, IDDM-2018, Lviv, pp. 215-228, 2018.

[4] A. Kilbas, H.M. Srivastava and J.J. Trujillo, “Theory and Applications of Fractional Differential Equations”, Elsevier, 2006.

[5] H.-Y. Liu, J.-H. He and Z.-B. Li, “Fractional calculus for nanoscale flow and heat transfer”, International Journal of Numerical Methods for Heat & Fluid Flow, vol. 24, no. 6, 2014, pp.1227–1250.

[6] Y. Sokolovskyy, M. Levkovych, O. Mokrytska and V. Atamanvuk, “Mathematical Modeling of Two-Dimensional Deformation-Relaxation Processes in Environments with Fractal Structure”, Proceedings of the 2018 IEEE 2nd International Conference on Data Stream Mining and Processing (DSMP), Lviv, pp. 375-380, 21-25 August 2018.

[7] Y. Sokolovskyy and O. Sinkevych, “Calculation of the drying agent in drying chambers”, Proceedings of the 2017 14th International Conference The Experience of Designing and Application of CAD Systems in Microelectronics (CADSM), Lviv, pp. 27-31, 21-25 February 2017.

[8] A.K. Golmankhaneh and D. Baleanu, “About Maxwell’s equations on fractal subsets of R3”, Cent. Eur. J. Phys., 11, 2013, pp. 863–867.

[9] J. Hristov, “Approximate solutions to fractional subdiffusion equations”, Eur. Phys. J. Spec. Top., 193, 2011, pp. 229–243.

[10] V.E. Tarasov, “Vector calculus in non-integer dimensional space and its applications to fractal media. Commun”, Nonlinear. Sci. Numer. Simul., 20, 2015, pp. 360–374.

[11] M.A.F. Dos Santos, “Non-Gaussian Distributions to Random Walk in the Context of Memory Kernels”, Fractal Fract, 2, 20, 2018.

[12] A. Atangana, “Fractal-fractional differentiation and integration: Connecting fractal calculus and fractional calculus to predict complex system”, Chaos Solitons Fractals, 102, 2017, pp. 396–406.

[13] A. Carpinteriand and F. Mainardi, Fractals and Fractional Calculus in Continuum Mechanics. Springer, New York, NY, USA, 1997.

[14] B.J. Westand, M. Bolognaand and P. Grigolini, Physics of Fractal Operators. Springer, New York, 2003.

[15] I. Podlubny, Fractional Differential Equations. Academic Press, SanDiego(1999)

[16] F. Mainardi, Fractional Calculus and Waves in Linear Viscoelasticity: An Introduction to Mathematical Models. Imperial College Press: London, UK, 2010.

[17] V.V. Uchaikin, Fractional Derivatives for Physicists and Engineers. Springer-Verlag: Berlin,Germany, 2013.

[18] J. Hristov, "An Approximate Analytical (Integral-Balance) Solutions a Non-Linear Heat Diffusion Equation", Thermal Science, 19, 2015, pp.723-733.

[19] C. Cattani et al., Fractional Dynamics. Emerging Science Publishers, Berlin, 2015.

[20] Z.P. Fan et al., "Adomian Decomposition Method for Three-Dimensional Diffusion Model in Fractal Heat Transfer Involving Local Fractional Derivatives", ThermalScience, 19, Suppl. 1, 2015, pp.137-141.

[21] W.J. Krzysztofik, "Fractal geometry in electromagnetics applications—From antenna to metamaterials" Microw. Rev., 19, 2013, pp. 3–14.

IEEE Third International Conference on Data Stream Mining & Processing
August 21-25, 2020, Lviv, Ukraine

Optimization of Parameters of Technological Processes Means of the FlexSim Simulation Simulation Program

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

Oksana Velyka
Department of designing and operation of machines
Lviv Polytechnic National University
Lviv, Ukraine
veloks@ukr.net

***Abstract*— the study of functional processes of mechanical multiparameter systems becomes an urgent task during the fourth industrial revolution, when data becomes the most important aspect in this field. Research of the input data influence, changes in the process of mechanical system modeling, analysis of the interplay of different parameters relationships is an important task to ensure the most efficient production operation. The article presents results of the analysis and practical application of the mechanical systems study results in the area of mechanical engineering, manufacturing of equipment for various functional purposes,as an example, electronics, processing and food industries, etc. The necessity of technological systems creation processes simulation and visualization of actual data for further analysis of their relations, the state of mechanical system operation for the mechanical equipment manufacture, as well as the analysis of production processes for the efficient use of resources and time component of processes are shown. Examples of constructing simulation models of production processes in the FlexSim environment are given to research the impact of data relations that describe the operation of each object in particular and certain affect the overall system process. Research methods of practical use of relations of many independent parameters models of the system are offered.**

Keywords*— *input, output, multi-parameter technical system, information graphic technology, mathematical modeling, FlexSim simulation system.

I. INTRODUCTION

Program **Industry 4.0 (Industry 4.0)** defines the rapid development of information technology, automation of manufacturing, robotics usage in enterprises in Europe. Management of production, packaging, sorting, etc. is done in real time, using data that updates every second. It is important to take into account influence of external factors on the technological process [1, 2, 3]. Creating virtual copies of mechanical engineering objects, modeling the interaction processes of different parameters of the system operation in general, the impact of data on each system object separately describes the general scheme of production. An important role is played by Internet technologies for the modern automated enterprise, since the **Internet of Things** has become an integral component for design optimization, assembly of a product, technological process adjustment of machine-
building industry. An important step in modern production is processing of large amounts of information, data (**BigData**) using cloud and artificial intelligence (**Artificial Intelligence**) technologies. An operator who controls engineering process at any stage must receive the processed data as quickly and conveniently as possible to analyze and decide on their further use in the technological process. Given the above stages of modern production, the urgent task is to use specialized IT platforms, business process management systems, modeling different stages of production, sorting, assembly, analysis of equipment, etc. (Fig. 1).

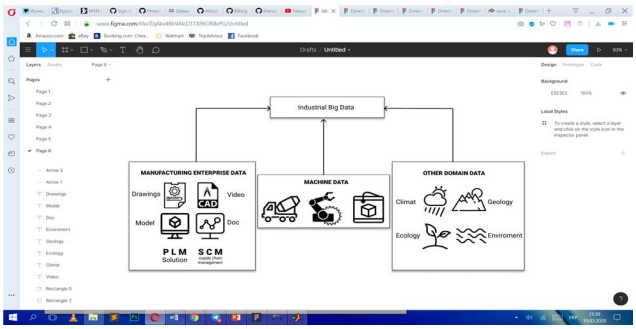


Fig. 1. Structural diagram of the factors that make up **Industrial Big Data**

Data that covers the parameters of the entire process, ie Industrial Big Data include the following components: data on the project **Manufacturing Enterprise data** (drawings of

the production object): drawings of objects **Drawings**, 3D - model, video, documentation, etc.; **PLM** solutions (**Product Lifecycle Management**) [1], product lifecycle management. These are information management systems in which data, processes, business systems and employees
are combined into a single informational space. PLM systems allow you to manage important information throughout the product lifecycle – from idea, design and production stages to maintenance and disposal. Such systems mainly contain mechanical processes of the industry production cycle, managed through the implementation of **PLM** software. Therefore, being an expert in mechanical engineering, it's easy to understand and customize **PLM** software to meet industry needs;

SCM (supply chain management) [1, 2] is a management and organization strategy that approaches management of the entire data flow of raw material, materials, products, services that appear and change during the implementation of logistics, production processes. The purpose of this strategy is to obtain an economic effect (reducing delays, controlling demand for products, etc.).

The next component for **Industrial Big Data** is **Machine Data** [1, 2, 3], which we get directly from production, information about an object or process: time, pressure, speed, temperature, etc. It is **Machine Data** (**Time Series Data**) that is the most voluminous of all data and changes every hour, minute, second.

Other data that affects **Industrial Big Data** are environmental, ecological, climatic, terrain, and more. They determine the main characteristics of the equipment, adjusting the process of processing the object, its manufacturing, etc.

Well-known scientific studies that thoroughly investigate the processes of data collection and processing for industry, model the input and output parameters of technological enterprises. Thus, [1, 2] discusses some aspects and approaches for creating algorithms for complex hierarchical systems of industry, oriented on business analyst, the manager of enterprises, intended for individual choice of models for modeling of various practical problems, taking into account input parameters.

II. Problem Statement

In the structure of industry the highest share is occupied by the branches of ferrous metallurgy, mechanical engineering, electric power, chemical and food industries. Gathering and analyzing data that affects the technological process of these industries is an important element in modeling the steps of creating, processing, sorting parts, or products. Mathematical process modeling allows you to analyze the structure of work and use data to analyze predicted results, change the number of data flows, and more. Geometric modeling provides visualization of the relations between different system parameters. It is effective for multi-parameter systems where the number of system parameters exceeds three. Simulation modeling allows you to combine the two steps and create a logical, mathematical description of the process that will visually present the design results, evaluate the operation of the process and provide a direct impact on the design result in real time. Visual representation using inputs that affect the operation of each object, allows you to better regulate the relations between the objects and to track emergency situations or changes at each stage, that lead to economic losses. Presentation of mechanical systems research results in the **FlexSim** software environment **makes** it possible to optimize the parameters of the system, taking into account its versatility for the study of technological systems at different stages of the product life cycle.

Therefore, the production engineer should have an overall outline of the production line and understand the impact of the data that changes over time for each step. In our opinion, this is effective when creating a simulation model of the process as a whole and separately for each stage of the technological process. The algorithm for constructing a map of processing data creation process can be :

- analysis of the input data taking into account the process features;
- creation of a mathematical model where the relations between different parameters and the interaction between the stages of the technological process are researched;
- creating a simulation model, researching possible contingencies by looking into possible combinations of data that create problem areas in the process;
- prediction of changing parameters, initial data, relations between objects, complication or simplification of technological process model of mechanical engineering to save resources, time;
- possibility to implement changes to model creation.

The development and study of the functionality of simulation models for the analysis of technological systems is an urgent task for modern production.

III. Stages of Research of Production Problem and Creation of Its Model

In order to study the technical system of production, it is first of all necessary to process the data related to the process. Consider a system structure denoted by its ***St*** **.** It is described by many elements of the system

$$Me = \{e1, e2, e3, \ldots\} \quad (1)$$

and the many relations between them

$$Z = \{z1, z2, z3 \ldots\}. \quad (2)$$

Then

$$St = \{Me, Z\}. \quad (3)$$

The behavior of such system (3) is effectively investigated as a sequence in time of its states. Each state of the system is affected by certain external factors, which are also taken into account in the simulation. The use of a simulation model improves the search for rational solutions in the management of an industrial enterprise. We distinguish the main stages of research and analysis of the model:

1. Formulation of goals. Allocation of a goal (multiple goals) that must be reached during management.
2. Definition of the object of management (research). This stage involves a detailed analysis of the part of the process (or the whole process) that interests the engineer and is based on the goal. At this stage, a description of the factors influencing the technological process, the environment, select several variants of the object and choose the appropriate variant on one or more criteria.
3. Structural optimization of the model. This stage consists of the following steps: determination of input and output parameters of the object, selection of structural elements of the model, changes of parameters.
4. Correction of the whole control system, all stages of control: selection of system elements (trimming elements), changing the structure of the system (inputs / outputs of the model).

Analyzing the parameters that affect the model as a whole, we can draw up a model structure that describes the relations between different parameters that affect the flow of the industry process (Fig. 2).

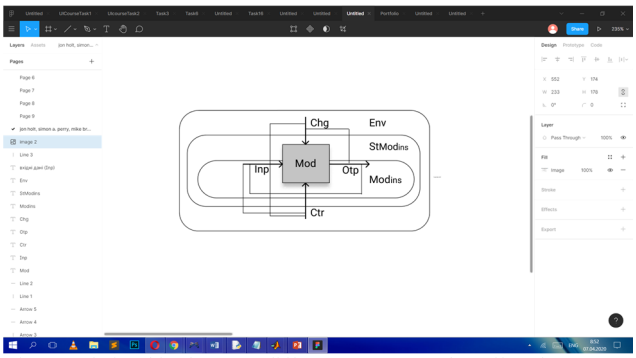


Fig. 2. Relationships between different parameters that affect the process

Consider the process modeling of industrial processes in accordance with Fig. 2.
The study of the **Mod (Model)** of the enterprise takes into account the influence of the following parameters:
Inp - (Input) input: This is information about objects managed by an industrial enterprise, an information stream that comes describing the process;
Chg - (Changes) changes made during the research process, (data entry, variables, relationships);
Ctr - (Control) control rules, their relations and the impact on the process as a whole and on the simulation results. **Chg and Ctr are** interrelated, since the quality control of the modeling processes can be done through changes in the input data that characterize the modeling objects. Output results in **Otp** (**Output**), which are related to the following variables:
Inp (input);
Chg - changes that are made during the study of the modeling process.
The set of data sets and relationships between parameters is investigated in the following spaces:
Mod ins are internal processes that take place within the model and describe the set of relationships between the various parameters that affect the **Mod** process**.**
Env - the influence of the environment on the process of changing external data and their impact on the processes **Mod in;**

StModins is a structure of a modeling process that incorporates **Modins** internal processes and the influence of external factors **Env.**

For the study of industrial lines, we have chosen the **FlexSim** environment, which is effective for simulation. Consider modeling on the example of the process of sorting products in a warehouse with the involvement of the operator [4, 5]. The model can be effective for detecting defects in batch production or checking for product quality (Fig. 3).

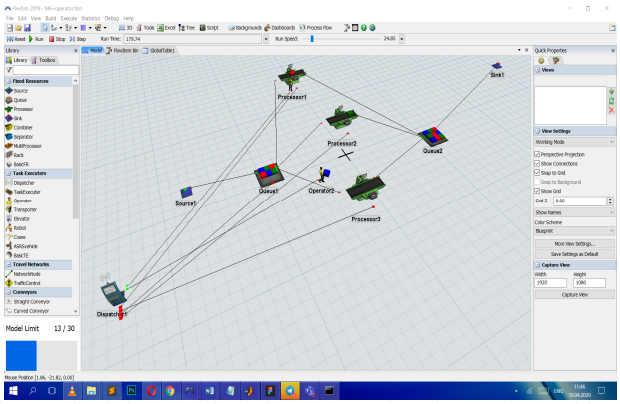


Fig. 3. Simulation model of the process of sorting products with the involvement of operators

The proposed scheme consists of the following objects:

Source1 - object creation,

Queue1, 2 - queue, assembly point,

Operator 1, 2 - operators,

Dispatcher - an object designed to control the operators.

Processor1, 2, 3 - processors,

Tote - sort object,

Slink1 - exit, process stop point.

In the properties of **Source1** (Fig. 4) we enter data on the type of products being investigated and the number of objects at the entrance. For this example, we obtain two types of output as follows:

Source - Triggers - On Creation - Set The Item Type and Color - Item Type - duniform (1, 2, getstream (current)).

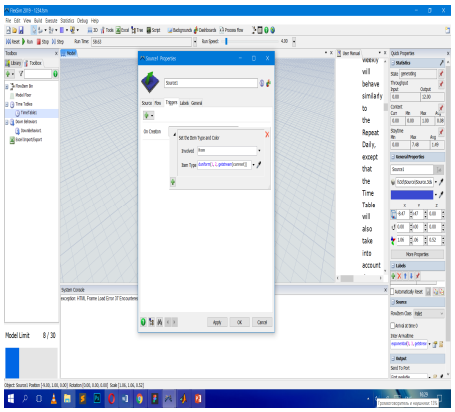


Fig. 4. Enter the initial data to work in the **FlexSim** environment

For each object in the **FlexSim** system, **it is** possible to choose the type of distribution (Fig. 5). The **Source** object specifies the data for the distribution of objects, information about the selected port for the movement of each object. The path to determine the port for object movement is as follows:

Source - Send To Port - First available.

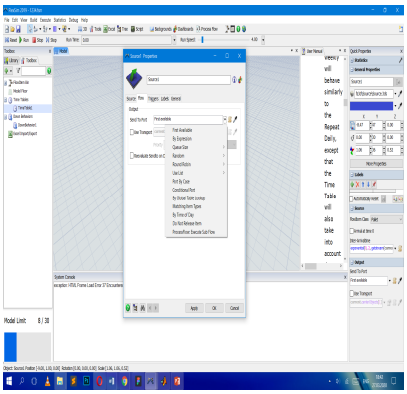


Fig. 5. Select the type of object distribution for **Source1**

During the development of a simulation model of products sorting process involving the operator, the impact of different data on each object in particular and on the system as a whole was analyzed. In Fig. 6 shows a diagram of the relation between model parameters and the display of data that affect each object in particular.

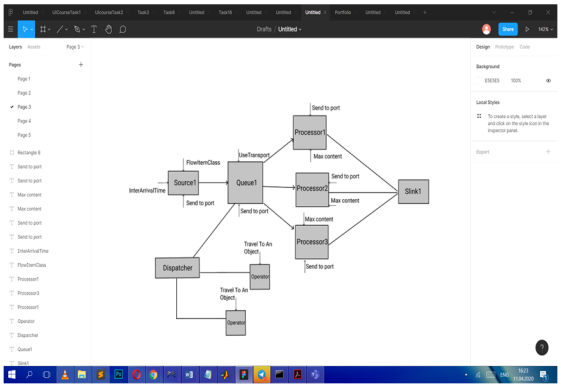


Fig. 6. Scheme of the relationship between the model parameters and the display of data that affect each object in particular

The first step is to set the input for the **Source1** object . Here are the main types:

Inter - ArrivalTime, FlowItemClass, SendToPort.

The **Inter - ArrivalTime option** determines the time of arrival and the type of data distribution, where we select the data to be distributed as follows:

exponential (0, 2, getstream (current)), where

num *exponential* (num *location*, num *scale* [, num stream]).

Value type **num** determine how often flows are created (in seconds) that returns a unique random data stream associated with the object.

The **FlowItemClass** option allows you to select the type of objects, for example, in this case **Pallet is** selected. Among the known options is the ability to select objects of different geometric shapes (**box, cylinder, sphere, circle, pallet, track,** etc.) to study the features of technological processes taking into account the structural shape of models, parts, products, etc. The program also allows you to create your own type of object, to give it the necessary characteristics:

Source1 - Source - FlowItemClass - Go To FlowItemBin.

The data to be entered in **SendToPort 'is** responsible for the number of ports and the principle of allocation of objects or parts, respectively, for each port. The way to specify this option is as follows:

Source1 - Flow - SendToPort –GoToFlowItemBin - FirstAvailable.

FirstAvailable is a type of object allocation that does not require a port number and number of objects, a feature distribution. This type of object distribution allows them to move in one direction only, since only one port and one direction are specified. Otherwise, you must specify a condition for the distribution of objects by the given characteristics, for example: **By Expression**, **Queue Size, Random, Port by Case, By Global Table, By Time of Day**.

Consider the **Queue1** object (Fig. 6) for which the following data is given: **Maximum Content**, for example, 5 units is the number of units of parts or products that will accumulate at a station before moving to the next station. There **are** two ports connected to **Queue1**, that is, two object directions. You must specify a destination for each product type by selecting **Port by Case** or setting an arbitrary direction of travel

Source1 - Flow - Send To Port –

Go To FlowItem Bin - Port by Case.

For **Processor1, Processor2, Processor3,** we enter data corresponding to the number of objects at the specified station and obtain the number of input and output ports corresponding to the number of connections (Fig. 7).

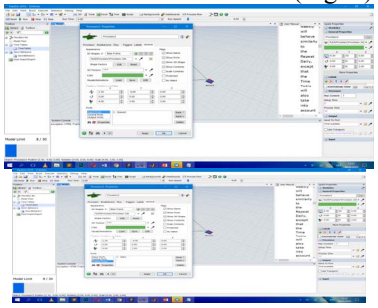


Fig. 7. Input and output ports for **Processor1, Processor 2, Processor3** objects

The **Slink1** object (Fig. 6) is a process stop, a simulation end, an exit.

In order to optimize the processes of the engineer, it is necessary to analyze how it is more efficient to move parts or objects to other stages of the technological process, whether to involve the operator, or to program the types of products sufficiently for further distribution. You need to calculate these questions, create a simulation model and compare the options. Therefore, modeling processes in the **FlexSim** environment is an effective way to explore and analyze different solutions to a single problem.

IV. Substantiation of basic stages of research, analysis of statistics and visualization of simulation data in FlexSim environment.

The analysis shows that in the study of processes using simulation of technological processes in the **FlexSim** system, there are five main stages of the study [8, 9] on (Fig. 8).

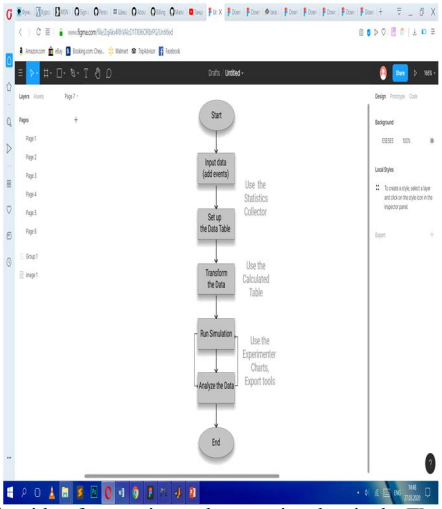


Fig. 8. An algorithm for entering and processing data in the **FlexSim** system

The first step is to set objects, enter data, create events, and build relations between objects. We use **Statistic Collestor** to view standard statistics for each object**.** The standard statistics overview for a 3D object in real time is as follows:

3D model - 3D object - Quick Properties - Expand button.

In particular, we use the processing of these types of data using the **Statistic Collestor** option. These options are contained in the Statistic group tab, which automatically expands when the simulation process starts. In Fig. Fig. 9 presents statistics of the **Queue1** object of the simulation model of the process of sorting products with the involvement of operators, which is shown in Fig.8.

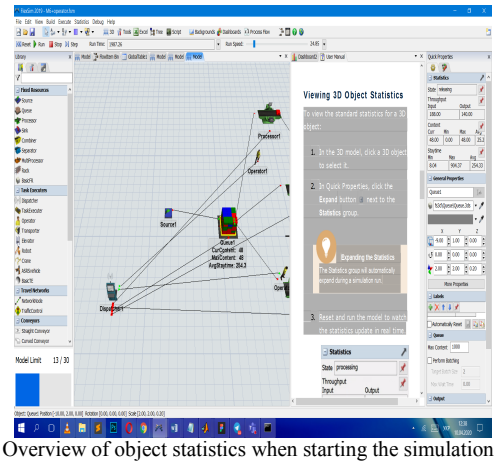


Fig. 9. Overview of object statistics when starting the simulation process

The next step in the **FlexSim** data processing algorithm is **Transform the Data**. We use **Calculated Table** to work with the data**. Calculated Table** can transform data obtained from **Statistics collector in** many ways: filtering, comparing, using advanced calculations, and more.

Once the simulation process is started, it is possible to analyze the data using **Dashboard** and **Use the Experimenter Charts.** For example, we analyze the dependence of the content of objects at a given station on time by creating a **Dashboard (Statistic - Content - Pin to Dashboard - Content vs Time)** (Fig. 10).

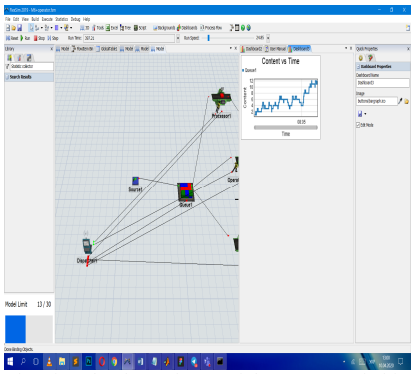


Fig. 10.Graph showing the dependence of the content of objects in a certain time interval

We see that in a certain interval of time (8 seconds) the number of units (4 pieces) that passed through this station was obtained.

V. Substantiation of technological process for the food industry (food industry) on the example of the juice bottling problem. Creation of a simulation model of the problem by means of FlexSim.

Consider the technological stage of production and bottling of juice (Fig. 11). The main production sites for the production of juice are the preparation of raw materials, dosage - mixed section, bottling and finished goods accounting. An important stage is water treatment. The water is purified and softened. The recovery of concentrated juices is that, with the help of special equipment, the moisture is recovered, that is, the prepared water is added. The next stage is the preparation of the container: the receipt of the container on the bottle washer, then on the labeling machine, dosing and filling machine, a special sealing machine. A visual

inspection machine is also required in order to notice a defective product [4].

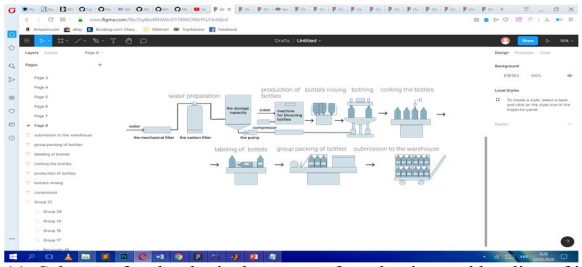


Fig. 11. Scheme of technological process of production and bottling of juice

The process of packing juices is to optimize the entire supply route, taking into account the interests of the manufacturer and the end consumer. Therefore, it is important to calculate the main steps and analyze them. For analysis, we considered the problem of constructing a simulation model of the operation of the juice bottling line. An empty bottle is sent to the line every second, sent to one of three sinks, each of which processes it in 3 ... 5 seconds. The bottles are then fed to a dispensing machine that spends 2 seconds filling each bottle. After that, two packing machines seal the bottles and stick labels. In this case, 10% of bottles are discarded. and optimize the simulation model of the shop [6, 7, 9].

FlexSim environment was used to create a simulation model of this technological process (Fig. 12).

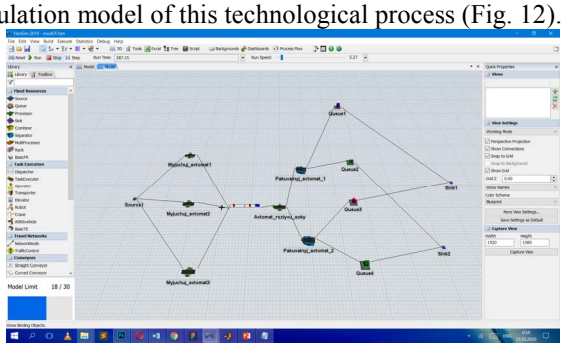


Fig. 12. Construction of a simulation model of the operation of the juice bottling line

From the **Source1** source, imitation models of bottles for three **Myjuchuj_avtomat1, Myjuchuj_avtomat2, Myjuchuj_avtomat3** washing machines are **included every second**, among which there are defective ones (Fig. 13).

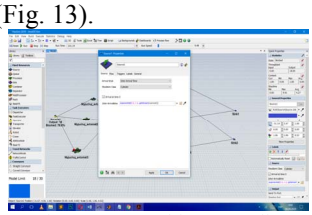


Fig. 13. **Source1** properties to output objects to a line

Assigning three objects to the simulation model of the problem is as follows:

Source1 - Triggers - OnCreation - SetItemType and Color - ItemType - duniform (1,3, getstream (current)) in Fig. 14.

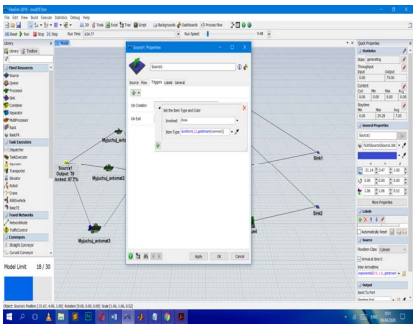


Fig. 14. The way to specify three types of objects

For the demonstration of defective objects is indicated in green in Fig. 12.

By the condition of defective bottles of 10%, so we denote this number of objects at the input, and set the other two types 45% and 45% equally to fulfill the initial condition (Fig. 15): **Source1 - Triggers - OnCreation - SetItemType and Color - Data - Set ItemType by Percentage.**

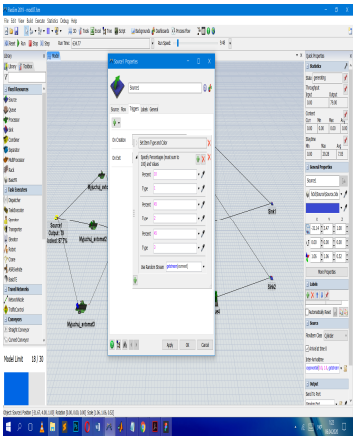


Fig. 15. Specifying object types

From the three washers, the bottles are fed to the conveyor, from where they move to the dispenser (**Processor** 1). After that, they arrive at two **Pakuvalnyj_avtomat_1** packing machines**, Pakuvalnyjavtomat_2**, which seal the bottles and fit the labels. At this stage, the bottles are discarded. To model this stage, we selected a **Separator** object from the **FlexSim 2019** library, named **Pakuvalnyj_avtomat_1** and **Pakuvalnyj_avtomat_2** for both stations, and added the following properties for the sort operation (Fig. 16):

Packing_Automat_1 (Packing_Automat_2) - Separator - Split - Split / Unpack Quantity - By Percentage;
Use Random Stream - getstream (current).

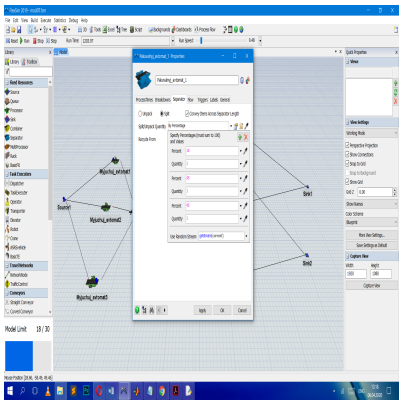


Fig. 16. Bottle rejection data entry

Queue 2 and **Queue 4** queues select discarded products that go to **Slink2** (recycle). **Queue1** and **Queue 3** turn into warehouse products (**Slink1**).

Using statistical analysis, it is possible to observe the frequency of receipt of defective bottles at the **Queue2** and **Queue4 station** for a given period of time in seconds (Fig. 17).

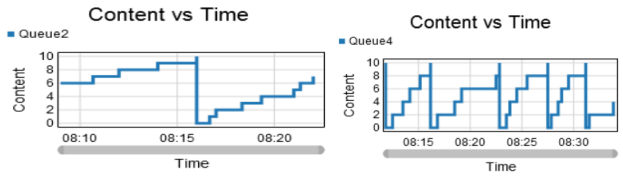


Fig. 17. Graphs showing the number of defective bottles (units) arriving at the station in the allotted time

VI. Conclusions

The result of the development of a structural model of the study of relations between different parameters of technological processes of industry is presented. The steps of data study necessary for simulation in **FlexSim 2019.** The following advantages of using this environment for specific tasks, such as the food industry, are demonstrated:

1) a simulation model of the juice bottling process allows to obtain information on various aspects of the technological process depending on the input factors;

2) the simulation model allows to investigate and analyze the process in cases where analytical calculations and mathematical programming fail;

3) it is much easier to develop an imitation model than an analytical one, since the process of creating an imitation model is step-by-step and modular;

4) the sructure of the simulation model naturally reproduces the structure of the process for which the model is built;

5) a significant advantage of simulation modeling in the process of solving the proposed type of problems is the ability to explore the model in time and create an animation of its behavior, which allows you to quickly find errors.

FlexSim simulation simulation also allows you to meet the challenges of other types of industry: mechanical engineering, electricity, chemical, transportation and transportation logistics etc.

References

[1] Ye. Nong, Data Mining. Theories, Algorithms, and Examples, CRC Press, 2014.

[2] P. Giudici and S. Figini, Applied Data Mining for Business and Industry, Wiley, 2009.

[3] Shu Ing Tay, Lee Te Chuan, AH Nor Aziati, Ahmad Nur Aizat Ahmad, “An Overview of Industry 4.0: Definition, Components and Government Initiatives”, Journal of Advanced Research in Dynamic and Control Systems, 10 (14), 2018, pp. 1379 - 1387.

[4] J. Mager, ILO Encyclopaedia of Occupational Health and Safety. Fourth Edition. Chapter 67, The Food Industry, Geneva, International Labor Office, 1998.

[5] J. Holt, S.A. Perry and M. Brownsword, Model - Based Requirements Engineering, Institution of Engineering and Technology, London, United Kington, 2012.

[6] O.Gumen, S.Ljaskovska and A. Ujma, “Forming the optimal parameters complex for the production process by means of FlexSim” Zeszyty Naukowe Politechniki Częstochowskiej ISSN No. 25, 2019, pp. 55-60.

[7] E. Forcael, M. Gonzalez, J. Soto, F. Ramis and C. Rodriguez, “Simplified Scheduling of a Building Construction Process Using Discrete Event Simulation”, 16th LACCEI International Multi-Conference for Engineering, Education and Technology: " Innovation in Education and Inclusion”, Lima, pp. 1-11, 19-21 July 2018. DOI: 10.18687 / LACCEI2018.1.1.194.

[1] S. Liaskovska, “Data processing of technological processes in mechanical engineering”, Scientific Bulletin of the Tavria Agrotechnological State University, Melitopol: TSATU, Is. 9, vol.1, 2019. [URL: http://oj.tsatu.edu.ua/index.php/visnik. DOI: 10.31388 / 2220-8674-2019-1ISSN 2220-8674.9]

IEEE Third International Conference on Data Stream Mining & Processing
August 21-25, 2020, Lviv, Ukraine

Informational Graphic Technologies for Fire Safety Level Determination in Special Purpose Buildings

Yevgen Martyn
Department of Project Management, Information Technologies and Telecommunications
Lviv State University of Life Safety
Lviv, Ukraine
evmartyn@gmail.com

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

Oleksandr Prydatko
Department of Project Management, Information Technologies and Telecommunications
Lviv State University of Life Safety
Lviv, Ukraine
o_prydatko@ukr.net

Igor Malets
Department of Project Management, Information Technologies and Telecommunications
Lviv State University of Life Safety
Lviv, Ukraine
igor.malets@gmail.com

***Abstract* — The article deals with the problem of public awareness about existing special purpose buildings such as protective buildings and shelters that can be used to protect against emergencies. The status update on the problem is considered. During research were analyzed recent scientific papers in the field of modern information technologies integration into civil protection system. The necessity of using modern informational graphic technologies for fire safety level determination in such type of shelters was substantiated. The features of actual data visualization with the involvement of mathematical transformations and methods of visual information graphing are investigated. Based on the offered methods of visualization graphics information has been developed software emulator "Fireware Emulator". It can be used for fire safety level determination in buildings of such type.**

Keywords— informational graphic technologies, software emulator, shelter, civil protection, room plan.

I. INTRODUCTION

The current integration state of modern information technologies to human life environment initiate appearance of new natural threats which can be serious risks to its normal and safety life. Military conflicts and the risks to use mass destruction weapon forced us to make protective purpose buildings - safety places such as housing units and shelters where peoples could stay and live during emergency situation.

Based on the daily facts of human life process both society and its individual members, we observe a progressive tendency towards increasing potential threats to humanity. The amount of the increasing anthropogenic and natural character risks is proportional to the number of tasks to be undertaken to ensure the safety of human life. Among these problems we could highlight the problems of peoples safe staying in protective purpose buildings: housing units and shelters.

The process of necessary building choosing is based on own human's preferences, its geographic location, type of emergency situation and other additional factors. But the most important among them is security (Fig.1). First, in such protective building, user should to determine the state of fire safety, the possibility of safe evacuation from the rooms and the comfortableness level of staying. After it, on the basis of visualized data about protective building, human should to analyze the possible deterioration causes of the situation in the shelter and to identify ways to better its planning. All upgrades must be produced by taking into account the position of ensuring proper level of fire safety.

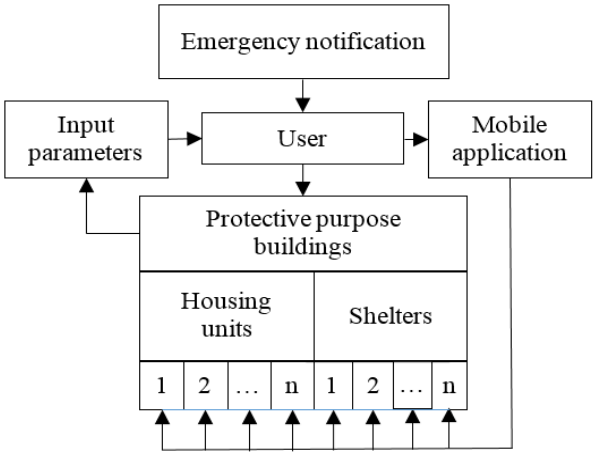


Fig. 1. Flowchart of necessary protective purpose buildings choosing process

If emergency situation will appear and it will be necessary to use protective building or air raid shelter, user should know in advance all important data, including comfortableness level and safety stay in it.

The main requirements that applies to protective buildings and shelters are regulated by the relevant state documents and the Civil Protection Code of Ukraine. These documents specify the external and internal parameters of the protective buildings. External includes geographical location and distance from neighboring homes, and internal includes

buildings configuration and planning of elements for safe staying users indoors. However, it is also important for users to have actual information about such buildings.

II. Problem Statemen

Review of recent international research shows numbers of existing developments in the fire safety field, particularly about uses of information technologies to improve human's life safety. The whole process of modeling protective environment in buildings for different purpose, including shelters has been describing in detail by scientists. In [1] were explored some aspects and approaches for creation of user-oriented mobile information technologies. This software is developing for individual use to help people in choosing suitable fire protective building where they could hide during emergency situation. According to [2], authors had analyzed fire safety requirements to information technology for implementation in selected commercial building. The main practical requirements to information technologies for fire safety providing in different types of buildings are highlighted and analyzed in [3, 4, 5].

Researches [6 - 10] refer to evaluation system of building fire emergency response capability maturity (FE-CMM) that based on the capability maturity model (CMM). Mostly they are considering to acceptable ways for emergency situations response. Authors declared that the proposed module could preliminarily realize the intelligent evaluation of building fire emergency response capability. Also, it able to improve the practice and intelligence of the fire emergency response capability evaluation, especially in smart cities. Studies [10-12] analyze the main features of developing, improvement and uses of modern information technologies in the field of civil protection, particularly during implementation of safety-oriented emergency response projects.

However, there are still little number of studies that related to software development for protective buildings safety evaluation, as well as for situation comparing capability in different buildings during best option choosing process. Today, there are also not enough information about qualitative evaluation of information technology impact to decision making process of choosing particular building type.

Current state regulatory documents determine all requirements to protective structures, buildings and shelters arrangement, particularly, from the point of safe stay in them. But, due to requirements universality and generality, shelters are objectively limited in ability of getting attention to each ones. The reason of it – lack of a real possibility to create completely safe building for a long stay. As a result, people must make choice of protective building and carry out its arrangement by themselves. This will be possible only if the following conditions will meet:

- the choice of each protective building or shelter is based both on its level of comfortableness and fire protection;
- all possible ways of getting to the protective building or shelter are analyzed depending on humans' location during an emergency;
- information technologies are used as a helping tool to make choice of protective building;
- preliminary prediction is made to determine unforeseen occasions appearance possibility during humans' stay in shelter;
- it is possible to make own additions to the contingency plan.

Therefore, according to previously outlined problems, developing of graphic information technology for fire safety evaluation of protective purpose buildings is one of the most important and urgent task in nowadays moving world.

III. Substantiation of Data Visualization Methods Use in Protective Buildings Fire Safety Evaluation Process

In general, users informational content I about protective buildings or shelters include two components: visual (presence component) and virtual (technological component) object review. Some users can get necessary information directly by observing selected object without using any information technology (IT), while others can use special software. Bases on this and according to [10], informational content I could be determine by the following equation:

$$I=u+iv=(x+iy)^3 \quad (1)$$

where u, v – components of information content, respectively, without and with the use of information technology; x – presence component parameter that indicate the level of knowledge about object without using IT; y – technological component parameter that indicate the level of knowledge about object with using IT.

Parameter y also takes into account both the level x' of knowledge about the object(building) and the use of developed application f:

$$y=x'+f \quad (2)$$

In most cases, x' does not always equate to x, but for simplifying research, we accept that $x = x'$. According to this, (1) takes the following form:

$$I=u+iv=x((4x^2+6xf+3f^2)+i(x^2-3f^2)) \quad (3)$$

The working range of variables x and f in information content I level determining process is in diapason between 0 and 1. Equation (3) determine information contents value I as a direct numerical dependence to the parameter x: for each level of information support f, the value of information content is equal to zero if $x = 0$. As a result, information content components u and v defines function of x and f. Thus, if we take these conditions into account, components values will determine like (4) and (5):

$$u=4x^2+6xf+3xf^2 \quad (4)$$

$$v=x^3-3xf^2 \quad (5)$$

If user does not use software for help in choosing suitable shelter and condition (2) is fulfilled, the value of components u and v will determine (6) and (7):

$$u=4x^3 \quad (6)$$

$$v=x^3 \quad (7)$$

Component v has value change restriction (8) of x and f. In our research we use only positive values that fits condition $x>3f = a$ and belongs to zone A (Fig. 2).

$$v=x(x^2-3f^2)= x(x-\sqrt{3}f)(x+\sqrt{3}f) \quad (8)$$

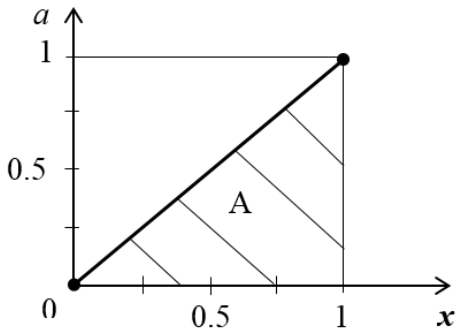


Fig. 2. Restriction zones of parameters

Now, let we determine the influence of parameters x and f to component u (4) values change. Even if x gets small values, such as $x = 0.1$, the values of component u increases extremely fast (Fig. 3 a). This proves the efficiency of using information technologies f.

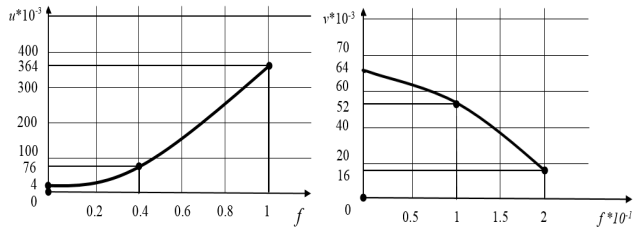


Fig. 3. Dynamics of components u (a) and v (b) values change

The information contents' component v is determined by the amount of knowledge set ($x'= x$) about buildings fire safety state according to which IT is used for help. The main condition that applies to v it is $x'= x >\sqrt{3f}$ (Fig. 3 b).

IV. Development Process of Information Graphic Technology for Protective Purpose Buildings Safety Level Determining.

Information technologies and its use in the field of Civil Defense can solve problem of visualizing data about protective purpose buildings. Basically, proposed emulator will work with graphical parameters of the buildings, which are used during its floors plan creation. Also, there will be possibility add another data, such as temperature of the shelter environment, the presence of fire extinguishers, etc. (Fig. 4).

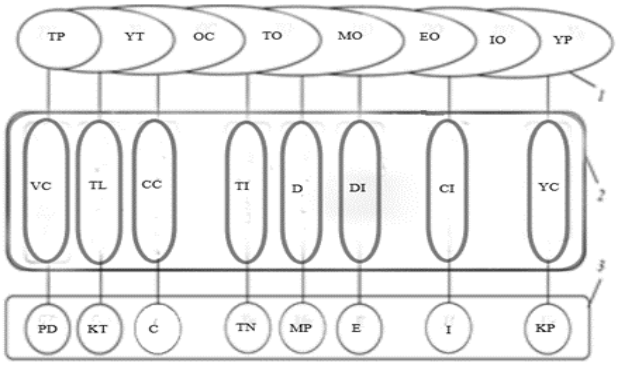


Fig. 4. Information model of data visualization process about protective purpose buildings with use of emulator

where 1 – emulator software operation block: TP - the technological process of emulator creation; YT - control units; OC - provision by programmers; TO - the level of equipment provision; MO - material resources; EO - energy resources; IO - the financial value of technology; YP - software resources; 2 - functional parameters of providing data visualization processes: VC - emulator implementation environment; TL - involved information technologies to the emulator creation process ; CC - the level of performers provision; TI - the level of information provision; D - the level of GPS monitoring technologies involvement of ; DI - the level of easy getting to the building; CI - transport infrastructure near buildings location; YC - basic standards for program creation; 3 - visualization resources: PD - the output product of emulator ; KT - programmers team; C - the product of created program; TN - technical means; MP - material resources; E - energy resources; I - information resources; KP - IT project team.

Information graphic technology is designed as an emulator program “Fireware Emulator”. For its development were chosen Java. “Fireware Emulator” allows users to evaluate protective buildings or shelters safety level.

Java is an object oriented language which gives a clear structure to programs. There are some major advantages of this language. Java is straightforward to use, write, compile, debug, and learn than alternative programming languages. Object oriented programming is associated with concepts like class, object, inheritance, etc. which allows you to create modular programs and reusable code. Java code runs on any machine that doesn’t need any special software to be installed. Those advantages are essential in developing secure, powerful device or computer system that gives you possibility for quick use of the right class and method. Software emulator “Fireware Emulator” was developed using Intellij IDEA 2019.1 Community Edition.

A flowchart of proposed software development stages was designed to facilitate the process of its programming (Fig. 5). The emulator has friendly user interface, is quite reliable and provides high speed of safety evaluate algorithm calculation.

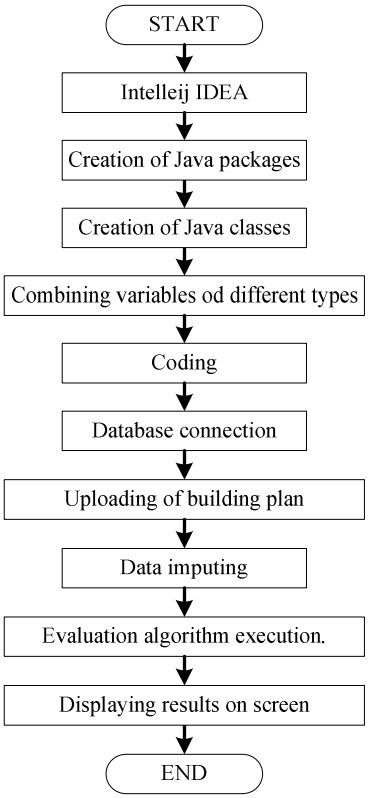


Fig. 5. The flowchart of the "Fireware Emulator" development process

Designed information graphic technology as an emulator program allows users evaluate the security level of shelters independently. The software supports different types of protective buildings and their plans. An algorithm for the process of choosing a safe building is presented in Fig. 6.

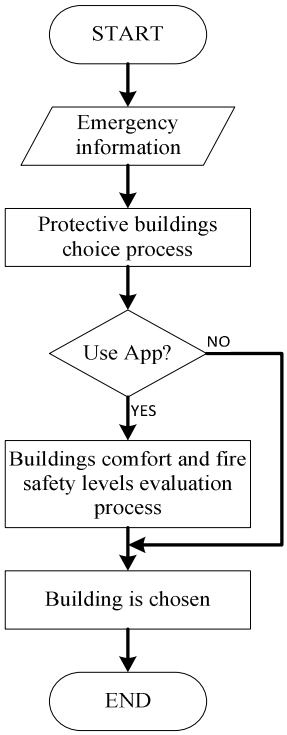


Fig. 6. The flowchart of protective building choosing process

User obtain actual information about buildings fire safety level and can check its reliability and evaluate his own protection in emergencies. This is possible by integrating buildings plan into emulator's database.

When buildings or shelters plans are importing into software database, the following parameters must be specified: inner temperature, height and width of the room, a wall material, the number of persons who may present at the same time. Lighting and air filtration settings are also available. Based on all this data, the program makes analyze and outputs the result of security level evaluation.

Software adaptation to the real conditions of the shelters and protective buildings is provided by changes of their plans. They are made based on additional perimeter measurements. Emulator "Fireware Emulator" stores visualized data of plan in its database, that makes them acceptable for changes up to different conditions.

The emulator interface for the school building is shown in Figure 7.

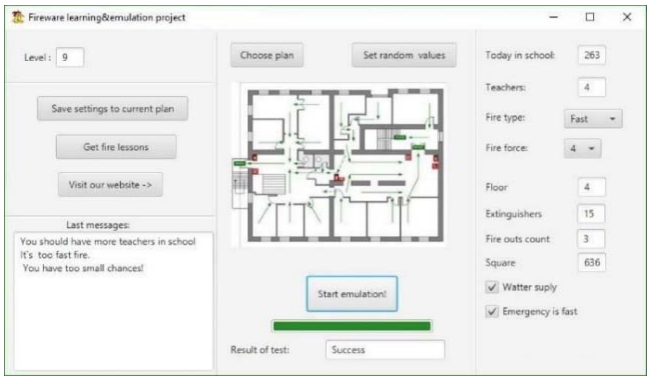


Fig. 7. Initial parameters values setting in software"Fireware Emulator"

Input parameters: two hundred sixty-three students and four teachers will be staying in building at the same time and each of them had passed a fire safety courses; there are four floors, three emergency exits and fifteen fire extinguishers; all-day water supply.

Developed software is open for updates. There is an opportunity to quickly add plans of buildings, functionality, providing reliability, accessibility and multiplatform. A high level of protection against unauthorized changes to the code was used.

V. USES OF DEVELOPED SOFTWARE IN FIRE SAFETY EVALUATION PROCESS

Next stage of research is analyzing correctness of the software algorithms. There were 4 practical experiments conducted using different input parameters, protective building plans and squares.

Experiment 1. Input parameters used in experiment: storage temperature - 21 °C, height - 2.45 m, width - 4.35 m, wall material - metal construction, number of persons - 5. The result of emulation (Fig. 8) demonstrates that such protective building complies to all standards and regulations and is completely safe to stay in an emergency.

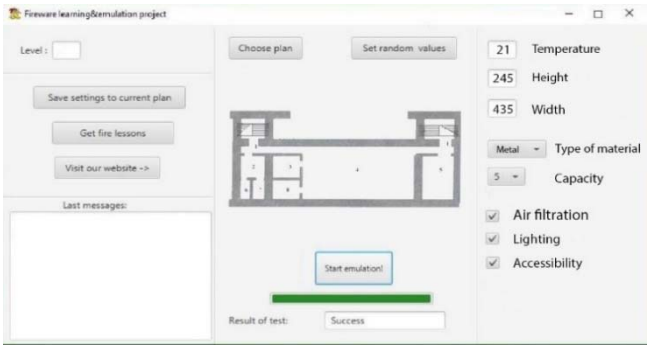


Fig. 8. Experiment 1: result of "Fireware Emulator" running with imput paramemters

The program provides a large number of parameter combinations. The algorithm selects the most optimal by analysing of given data [10]. In case of inconsistency, the program produces a negative result.

Experiment 2. Were used parameters which did not complies to standards or regulatory documents. Due to input parameters this shelter is uncomfortable for staying and the result of fire safety evaluation is negative (Fig. 9).

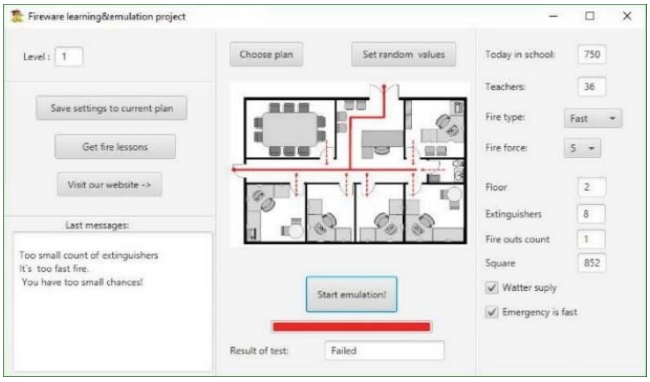


Fig. 9. Experiment 2: the result of buildings fire safety evaluation with many people inside

This is because there is no ability to avoid the danger: the building has a large square and only one spare exit, the number of people is too big, the fire ignites quickly and there are only eight fire extinguishers.

Experiment 3. A shelter with small square were analysed. Also there were a low wall height and inside temperature (8 ° C), the number of staying persons - 6. According to the normative documents, such protective building is not suitable for a comfortable staying despite positive level of fire safety evaluation. The result of the program demonstrates a similar conclusion (Fig. 10).

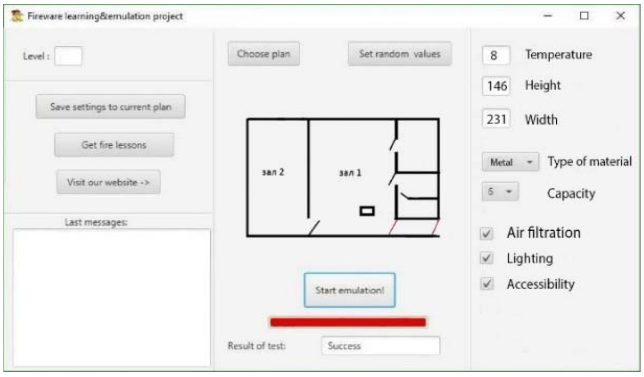


Fig. 10. Experiment 3: negative result of shelter staying comfortability

Experiment 4. In developed software "Fireware Emulator" is predicted a graphic information window for displaying problems which must be solved to increase both fire safety and comfortability levels. In Fig. 11. is presented the result of evaluation process, according to the input parameters. In information window is displayed text message with recommendations.

Analyses of input conditions and received recommendations (Fig. 11) shows that the algorithm of message output is correct. The inside temperature of 4 degrees above zero storage does not correspond to the norm, the height - 1.68 m and the width - 4.61 m. In this experiment, there is lighting, but no air filtration. All this conditions are dangerous and not comfortable for staying of 52 people in shelter.

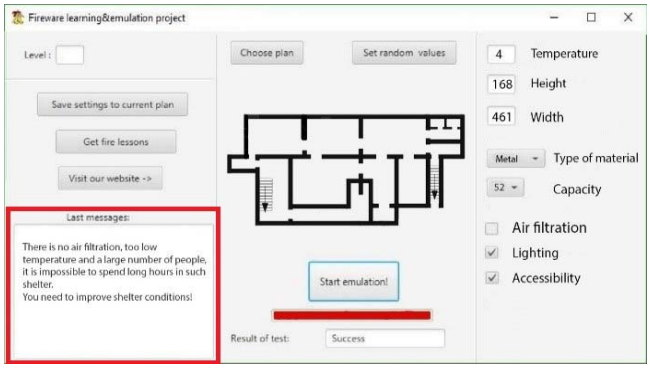


Fig. 11. The result of evaluation process with generated recommendations

The errors and deviations of the program are minimal, and their analysis indicates correctively of algorithms calculations.

VI. CONCLUSION

Current rates of science and technology development, integration of information technologies into everyday life of society, state of ecology and influence of anthropogenic factor on the environment lead to appearance of new emergencies. The ability to make the right decisions in such circumstances is a guarantee of safety human's life. That is why it is so important for human today to have as much as possible information about around environment.

Informational graphic technology was developed as an emulator program to help people in process of protective buildings fire safety level evaluation. The emulator program "Fireware Emulator" is a product of universal purpose. It helps to determine the level of person's security staying in protective purpose buildings. Developed software is designed for personal use. Performance of the program can be enhanced by the combination of prior knowledge about object and the information component comparing. Also were made mathematical substantiation of evaluation algorithms execution correctness.

Integration of the latest information technologies in the field of civil protection will provide to society an effective tool for analyzing and using data to explore possible ways of avoiding or protecting against emergencies.

REFERENCES

[1] E. Martyn, S. Ljaskovska and N. Tarapata, "Emulator of analysis of bombshelters", Scientific bulletin of the Tavria agrotechnological state university, Melitopol: TSATU, Is. 9, vol.1, 2019. DOI: 10.31388/2220-8674-2019-1-62

[2] M. Anjay and S. Anjuli, "Assessment of Exit Requirements for Fire Safety of Commercial Buildings, Kathmandu, Nepal", International Journal of Emerging Technologies and Innovative Research, vol.4, Issue 10, 2017, pp. 248-255. DOI: 10.1717/JETIR.17074

[3] Building Department, Code of Practice for Fire Safety in Buildings, 2011, Available at: [www.bd.gov.hk/english/documents/code/fs_code2011.pdf].

[4] OSHA, Available at: [https://www.osha.gov/OshDoc/data_General_Facts/emergency-exit-routes-factsheet.pdf].

[5] G. Ma, S. Tan and S. Shang, "The Evaluation of Building Fire Emergency Response Capability Based on the CMM", International Journal of Environmental Research and Public Health, 16(11), 2019, 1962. DOI: 10.3390/ijerph16111962.

[6] A. Rego, L. Garcia, S. Sendra and J. Lloret, "Software Defined Network-based control system for an efficient traffic management for emergency situations in smart cities", Future Generation Computer Systems, 88, 2018, pp. 243-253. DOI: 10.1016/j.future.2018.05.054

[7] S. Granda and T. Ferreira, "Assessing Vulnerability and Fire Risk in Old Urban Areas: Application to the Historical Centre of Guimar√£es", Fire Technology, 55, 2019, pp. 105-127. DOI: 10.1007/s10694-018-0778-z.

[8] N. Kwok, C. Bratiotis, M. Luu, N. Lauster, K. Kysow and S. Woody, "Examining the Role of Fire Prevention on Hoarding Response Teams: Vancouver Fire and Rescue Services as a Case Study", Fire Technology, 2017. DOI: 10.1007/s10694-017-0672-0.

[9] J.C. Cheng, Y. Tan, Y. Song, Z. Mei, V.J. Gan and X. Wang, "Developing an evacuation evaluation model for offshore oil and gas platforms using BIM and agent-based model", Autom. Constr., 89, 2018,pp. 214-224. DOI: 10.1016/j.autcon.2018.02.011.

[10] R. Ratushnyi, P. Khmel, A. Tryhuba, E. Martyn and O. Prydatko, "Substantiating the effectiveness of projects for the construction of dual systems of fire suppression", Eastern-European Journal of Enterprise Technologies, 4, 2019, pp. 46-53. DOI: 10.15587/1729-4061.2019.175275.

[11] O. Smotr, N. Burak, Yu. Borzov and S. Ljaskovska, "Implementation of Information Technologies in the organization of Forest Fire Suppression Process", in Proceedings of the 2018 IEEE Second International Conference on Data Stream Mining & Processing (DSMP), Lviv, Ukraine, pp. 157-161, August 21-25, 2018. DOI: 10.1109/DSMP.2018.8478416

[12] I. Malets, O. Prydatko, V. Popovych and A. Dominik, "Interactive Computer Simulators in Rescuer Training and Research of their Optimal Use Indicator", in Proceedings of the 2018 IEEE Second International Conference on Data Stream Mining & Processing (DSMP), Lviv, Ukraine, pp. 558-562, August 21-25, 2018. DOI: 10.1109/DSMP.2018.8478486

IEEE Third International Conference on Data Stream Mining & Processing
August 21-25, 2020, Lviv, Ukraine

Increasing the Animation Study Management Services Functioning Efficiency

Olga Smotr
Department of Project Management, Information Technologies and Telecommunications
Lviv State University of Life Safety
Lviv, Ukraine
olgasmotr@gmail.com

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

Oksana Karabyn
Department of of Applied Mathematics and Mechanics
Lviv State University of Life Safety
Lviv, Ukraine
karabynoks@gmail.com

***Abstract*—the research deals with the issues of animation of the animation studio services market, and the advantages and disadvantages of their functioning, both for the ordinary user and from the developer. It made clear that it is appropriate to provide information to the client about two areas: pricing policy and content of services. The services automation methods within animation studios are researched using modern methods of Data Mining.**

It is proposed to build the work logic using data mining methods, Decision Trees methods, for the animation studio management services effective functioning. It is determined that in case of client priority search criteria choice is related to pricing policy, it makes sense to organize the operation of the animation studio management service based on the backtracking algorithm. In the case of client's search criteria choice priority are content-related topics of services, apply algorithms for constructing decision-making tree.

Keywords — information technology, Data Mining, Decision Trees, animation studio, entertainment

I. INTRODUCTION

Recently, there has been a rapid development of the entertainment industry worldwide. This is the entry into a post-industrial stage of development most the developed countries in the world, a characteristic feature is the significant growth of services and service industries. Despite the fact that the entertainment industry is one of the youngest sectors of the socio-cultural sphere, it accounts for about 6 % of the world's capital [1].

In today's conditions of informatization and computerization of the society [2] and the rapid increase in demand for services of the entertainment industry [3-4]. Business development in this area requires new approaches to information processing and decision-making.

The most notable change is in the travel companies that actively involve organizations for process activities advanced digital technology [5-9]. These include global computer reservation / reservation systems, integrated communications networks, multimedia systems, smart cards, management information systems, and more.

However, they are mostly dedicated to addressing the digitization of tourism animation services and, unfortunately, very little attention has been paid to digitizing the children's animation studio segment. Modern animation is the activity of developing and implementing special leisure programs [10]. Although the demand for services in this particular segment of the entertainment industry is huge. With a data web - site for parents Britain Netmum, parents spend on children's holiday from $ 200 to several thousands of dollars, and in general, the market value of children's parties is over $ 1.5 billion [11]. Similarly, in Ukraine, on average, the organization of children's holidays costs about 5 thousand UAH [4].

II. PROBLEM STATEMENT

Of course, today most animation studio, event -agency etc. have their own web services, where the highlight information about their services and show potential customers their own successful experiences conducted their activities, etc. However, this is not enough. The client (animation customer) needs the ability to quickly and conveniently view the offerings of different animation studios with the ability to compare their services, pricing, quality of service and the ability to provide these services in the right location and time range, etc.

Analyzing typical for the market of Ukraine, web services animation studio "Children's Planet" [12] "Igroland" [13], "Papashon" [14], event - agency "Empire holidays" [14] it is difficult not to notice, that the data the services are not convenient and do not meet the above requirements.

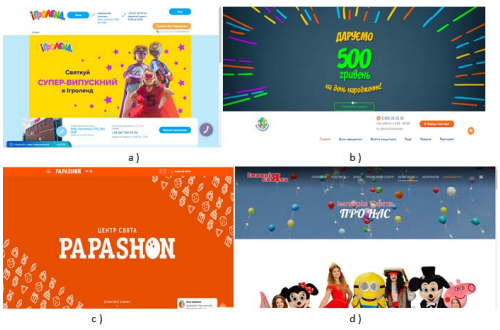


Fig. 1. View of the main pages of services of children's animation studios: a) children's animation studio "Igroland"; b) children's animation studio "Children's Planet"; c) the PAPASHON entertainment complex; d) "Empire of the Holiday" event-agency

Obviously, all four main pages of web services (Fig. 1) are not radically different and do not carry information regarding the choice of possible services. They do not have an intuitive or adapted interface. After pressing certain buttons and selecting menus, which, by the way, are not easy to find, the information becomes not much more. So, for example, to find the list of services of the children's animation studio "Children's Planet", you must use the page scrolling, on the website of the children's animation studio " Igroland " you need to find the "Menu" button, and click on the "Details" button on the main page of the entertainment complex PAPASHON. However, even after such manipulation of information comes not much more. None of these services is immediately accessible to the complex of services that can be obtained based on a certain budget. Also, none of these services has the opportunity to introduce, for example, topics that are interesting for the child to get information about the range of services that the company can provide and the price range of these services, it is difficult to find differentiation of services by age categories and so on.

In addition to imperfect client-side work, these services are not functional enough to meet the needs of animation studio staff. Here are some of the main unsolved problems:

- does not have a cumulative database of customer preferences and needs;
- there is no service of forming comparative characteristics (cost, duration in time, age restrictions and number of participants, etc.) of different offers;
- when ordering a client an offer to the administrator comes just a message about this event, which is not processed by the service itself, respectively, with a large influx of people who want to order the same service, the service will not help the administrator with this problem;
- there is no dynamic selection of the set of possible offers for the client according to the criteria he / she has set;
- there is no dynamic selection of possible offers for the client, taking into account time and quantitative parameters, the possibilities of the studios themselves, etc.

III. THE METHOD OF IMPROVING THE WORK OF CHILDREN'S ANIMATION STUDIO SERVICES WITH THE USE OF DATA MINING METHODS IS PROPOSED

For fast and qualitative sampling of data, modern methods of analysis of intellectual data must be applied by certain criteria. We believe that it is most appropriate for the organization of the effective operation of a service for managing the work of animation studios to use the Decision Trees methods, such as to enable, step by step , based on the client's answers, to form from the existing set of services, the set of the most client-friendly solutions.

To construct Decision Trees most widely used in practice, algorithms binary search tree (BST) ; returning search algorithm (backtracking algorithm) and algorithms decision tree.

In our opinion, to sample a variety of possible services, it is most advisable to use the backtracking algorithm when choosing a client as a priority, the direction of pricing.

The operation of this algorithm can be interpreted as a process of bypassing a tree. Each peak corresponds to it a sequence $(x_1, \ldots, x_i)$, with peaks that correspond to sequences of the form $(x_1, \ldots, x_i, y)$, sons of the summit. The root of the tree corresponds to an empty sequence.

This tree is being traversed by searching deep. In addition, a predicate P is specified on all tree vertices. If $P(v) = False$, then the subtree vertices with root at vertex v are not considered, and the volume of the bust decreases. The predicate $P(v)$ acquires the value False when it becomes clear that the sequence $(x_1, \ldots, x_i)$, corresponding to vertex v cannot in any way be added to the complete solution. To apply this method, the solution of the problem must look like a finite sequence of elements $(x_1, \ldots, x_n)$.

In our case, the elements $x_1, \ldots, x_n$ of this sequence are the cost of the services that can be provided to the client within the budget that he or she determines (predicate P). This is exactly what a client wants, based on a specific budget.

Initially, the client is offered a set of services that can be provided within a given budget, considering a possible set of services, the client is able to remove / add certain of them, then the budget for other services increases / decreases. The algorithm returns the customer to the previous selection step, with the other set of service elements until the desired set of services within the client's budget is formed.

Description of the backtracking algorithm

After entering in the search window the "budget" of the amount that the client focuses on, the logic of the system employs a backtracking algorithm , which allows to form from the existing set of services, the set of services that can be provided within the specified budget.

Let us illustrate the algorithm for generating multiple offers (search with returns) in case the client chooses as a priority, the direction of the pricing policy, that is, according to the criterion "budget", in a specific example.

Suppose that table 1 sets the sets of services $\{p_1, \ldots, p_n\}$.

TABLE I. SET OF EXISTING SERVICES

Code	Service	Cost	Duration
p1	Costume photoshoot	1000 UAH	1 hour
p2	Rolledrome	600 UAH	1 hour
p3	Birthday greetings with cake	550 UAH	30 minutes
p4	Weaving of African pigtails	500 UAH	1 hour
p5	Soap bubbles show	500 UAH	1 hour
p6	Trampoline arena	120 UAH	1 hour
p7	Aqua makeup	90 UAH	30 minutes
p8	Trampoline «Treasure Island»	80 UAH	30 minutes
p9	Maze	60 UAH	unlimited

Each service is matched by its cost. You need to find a subset of which the cost of the elements does not exceed the client's criterion "budget" (CB). If the sum is so large that the addition of any new number exceeds CB then we go back and change the last addition of the sum. Let $CB = 1$ *thousand UAH*. Figure 2. illustrates part of the backtracking algorithm for the problem of finding a subset of a set of services $\{p_1, \ldots, p_n\}$ with the criterion "budget" equal to 1 thousand UAH.

In general, when a subset of the set of services $\{p_1, \ldots, p_n\}$, given in Table 1. from the criterion "budget" is equal to 1 thousand UAH. The client may be offered 6 possible service options. Formed variants of subsets of possible services are given in Tables 2-7, respectively. And it should be noted that these options are provided, with information on the possible duration of the organized holiday.

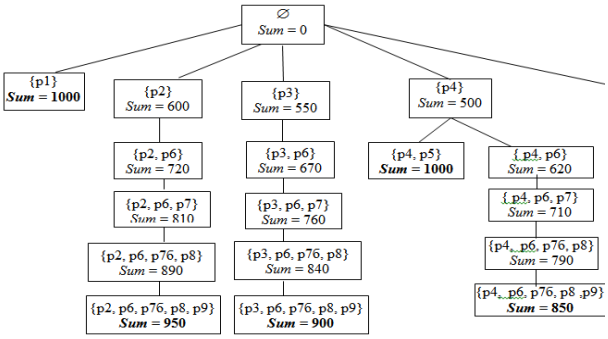


Fig. 2. Part of the backtracking algorithm for problem of finding the subset of a set of services $\{p_1, \ldots, p_n\}$ with the criterion "budget" equal to 1 thousand UAH.

TABLE II. SET OF OFFERED SERVICES

Code	Service	Cost	Duration
p1	Costume photoshoot	1000 UAH	1 hour
	Sum	**1000 UAH**	**1 hour**

TABLE III. SET OF OFFERED SERVICES

Code	Service	Cost	Duration
p2	Rolledrome	600 UAH	1 hour
p6	Trampoline arena	120 UAH	1 hour
p7	Aqua makeup	90 UAH	30 minutes
p8	Trampoline «Treasure Island»	80 UAH	30 minutes
p9	Maze	60 UAH	unlimited
	Sum	**950 UAH**	**unlimited**

TABLE IV. SET OF OFFERED SERVICES

Code	Service	Cost	Duration
p3	Birthday greetings with cake	550 UAH	30 minutes
p6	Trampoline arena	120 UAH	1 hour
p7	Aqua makeup	90 UAH	30 minutes
p8	Trampoline «Treasure Island»	80 UAH	30 minutes
p9	Maze	60 UAH	unlimited
	Sum	**870 UAH**	**unlimited**

TABLE V. SET OF OFFERED SERVICES

Code	Service	Cost	Duration
p4	Weaving of African pigtails	500 UAH	1 hour
p5	Soap bubbles show	500 UAH	1 hour
	Sum	**1000 UAH**	**2 hour**

TABLE VI. SET OF OFFERED SERVICES

Code	Service	Cost	Duration
p4	Weaving of African pigtails	500 UAH	1 hour
p6	Trampoline arena	120 UAH	1 hour
p7	Aqua makeup	90 UAH	30 minutes
p8	Trampoline «Treasure Island»	80 UAH	30 minutes
p9	Maze	60 UAH	unlimited
	Sum	**820 UAH**	**unlimited**

TABLE VII. SET OF OFFERED SERVICES

Code	Service	Cost	Duration
p5	Soap bubbles show	500 UAH	1 hour
p6	Trampoline arena	120 UAH	1 hour
p7	Aqua makeup	90 UAH	30 minutes
p8	Trampoline «Treasure Island»	80 UAH	30 minutes
p9	Maze	60 UAH	unlimited
	Sum	**820 UAH**	**unlimited**

Once a client is selected, a subset of the solutions that are most appropriate for him / her, he / she can make changes by removing / adding certain services (this will result in a certain budget change) and make an order.

If the client chooses as a priority, the content area of services, the most appropriate is to use the algorithm of building a decision tree. In this case, we need to solve the typical classification problem.

After all, when we consider a client's request in terms of the content topics of animation studio services , we need to make decisions about the set of existing objects (animation studio services), assigning them to certain thematic classes, that is, providing these objects with classification features. So we need to solve a typical classification problem whose set of conditional attributes ***A*** will be made up by the client's requirements. The set ***W*** is an active animation studio service; the set ***d*** is a decision attribute - two elements {"good luck", "bad luck"}.

Description of the algorithm for building a decision tree

After selecting a search box of your request priority area "subject", the customer will be asked to answer a few questions that are conditional attributes and accordingly help shape due to the algorithm for constructing Decision Tree , the set of possible proposals for a client under given his attributes.

Let's use one of the algorithms for building a Decision Tree, namely the algorithm ID3 (Iterative Dichotomizer-3 algorithm) [15]. To illustrate the algorithm for generating set of proposals using the algorithm for constructing Decision Tree to a specific example.

Let Table 8 provide information on options for a children's holiday. We construct Decision Tree for given Table 8.

TABLE VIII. SET OF OFFERED SERVICES

Version	Age	Gender	Number of participants	Subject	Type of entertainment	Duration of entertainment	Budget (UAH)	Holiday
1.	up to 5 years	boy	up to 4	not thematic	are active	unlimited		Yes
2.	up to 5 years	boy	up to 4	ninja	science	30 minutes – 1 hour	5000-1000	No
3.	5-10 years	boy	4-8	ninja	science	30 minutes – 1 hour	5000-1000	Yes
4.	up to 5 years	boy	more than 8	Peppa	art	до 2 hour	unlimited	No
5.	up to 5 years	girl	more than 8	Peppa	art	unlimited	5000-1000	Yes
6.	5-10 years	girl	4-8	ninja	team entertainment	unlimited	5000-1000	No
7.	5-10 years	girl	4-8	Lady bug	team entertainment	unlimited	5000-1000	Yes
8.	5-10 years	boy	more than 8	Lady bug	team entertainment	unlimited	5000-1000	No
9.	5-10 years	boy	more than 8	pirates	team entertainment	unlimited	up to 500	No
10.	more than 10 years	girl	more than 8	Peppa	art	до 2 hour	5000-1000	No
11.	more than 10 years	girl	up to 4	Peppa	art	до 2 hour	5000-1000	Yes

12.	more than 10 years	girl	more than 8	Peppa	are active	unlimited	5000-1000	No
13.	more than 10 years	girl	more than 8	not thematic	beauty and fashion	до 2 hour	unlimited	Yes
14.	more than 10 years	boy	more than 8	not thematic	beauty and fashion	до 2 hour	unlimited	No

A holiday is a decision-making attribute. The set of all conditional attributes $\boldsymbol{A}$ = {"age", "gender", "number of participants", "subject matter", "type of entertainment", "duration of entertainment", "budget"} corresponds to the root node. Select the attribute "age" and mark it the root vertex. The set of values of this attribute consists of three elements: up to 5 years, 5-10 years, more than 10 years. Put the root vertex in correspondence with three edges, each of which is attributed to the value of the attribute "age". Set examples will be divided into three subsets that correspond to the values of the attribute "age"; these subsets correspond to each of the vertices 2, 3, 4 of the tree shown in Figs. 3 We remove the attribute "age" from the set A and get the set $\boldsymbol{A}$ = {"gender", "number of participants", "subject matter", "type of entertainment", "duration of entertainment", "budget"}.

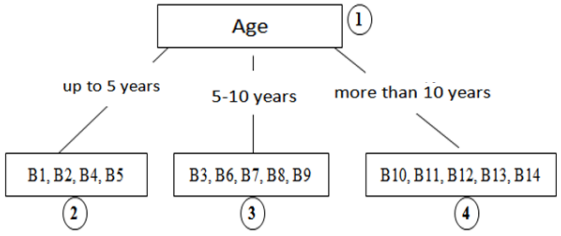


Fig. 3. The first step of the ID3 algorithm (removing the "age" attribute).

Consider the vertex number 3. It is matched by the subset of examples {B3, B7} that have the value of the decision attribute "*yes*" and the subset of examples {B6, B8, B9} that have the value of the attribute of the decision "*no*". We select the following attribute from the set $\boldsymbol{A}$; let it "gender". Denote by vertex 3, construct two edges with the values of this attribute, and divide the set of examples in vertex 3 into two subsets, in each of which the values of gender are the same.

Consider the vertex number 6. It corresponds to the subset {B3}, which has the value of the decision attribute "yes", and the subset of examples {B8, B9}, which have the value of the decision attribute "no". We select the following attribute from the set $\boldsymbol{A}$; let it be the "number of participants". Denote by vertex 6, construct three edges with values of this attribute and divide the set of examples in vertex 3 into two subsets, in each of which the values of the number of participants are the same.

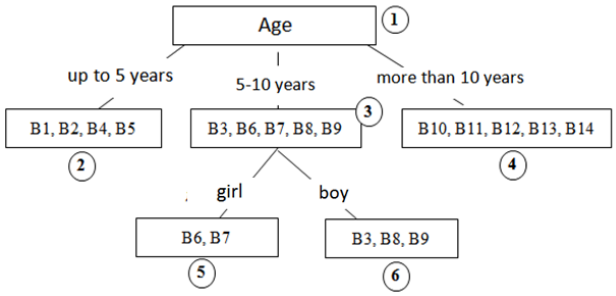


Fig. 4. The second step of the ID3 algorithm (removing the "gender" attribute)

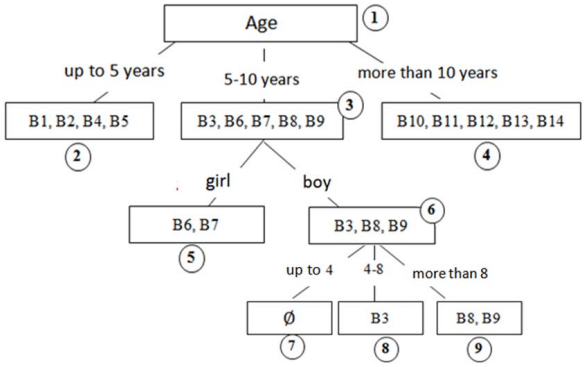


Fig. 5. The third step of the ID3 algorithm (removing the "number of participants" attribute)

In Fig. 5 in verse 7 we have an empty set, which indicates that under such criteria given by the client, we will not be able to offer him anything, that is, a holiday cannot be organized, so we will mark this vertex "no" and it will become a leaf. In verse 9, examples B8 and B9 have the same attributes of the game attribute - no. Therefore, we denote this vertex by "no" and it will become a leaf. Similarly, we denote vertex 8 as "yes" and it will also become a leaf

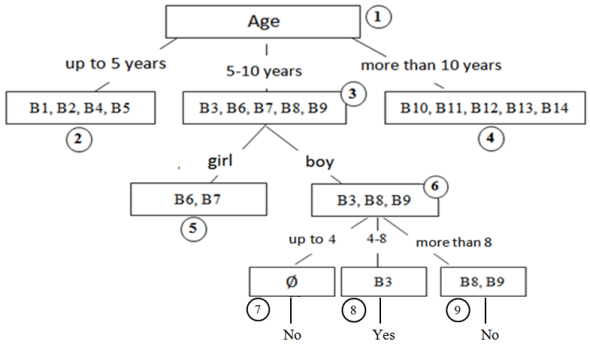


Fig. 6. The fourth step of the ID3 algorithm.

IV. CONSIDERATION OF THE DEVELOPED FUNCTIONAL OF THE SERVICE

The main page (see Fig. 7) of the service for managing the work of children's animation studios, includes buttons: "Offers" - for viewing, all available offers, "Order" - go to the window where you can place an order, "Register" - to register for service, "Login" - to enter the service and also 3 offers that can be accessed by clicking on the "Details" button.

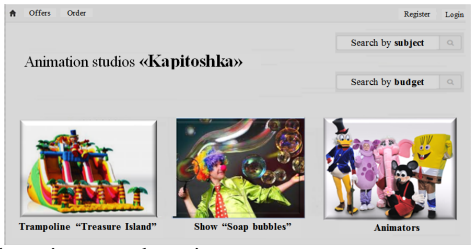


Fig. 7. The main page web service

After the user, when logging in to the site, chooses a budget search option and enters the amount for which he wants to organize a holiday, he will be offered as many services as possible within the selected budget, as shown in the figure. 8.

Fig. 8. A list of suggested holiday options

If someone at the entrance to the site will select the search option offers the direction of "subjects" and introduce subjects on which it wants to organize a holiday, he will be asked to respond to several questions, such as: the age of the child; to become a child; budget, the number of children on holiday, he will be offered as many services within the chosen subject. One of the variants of the proposals when choosing the theme "Active entertainment" and given attributes: "age" - 11 years; "gender" - boy ", budget - 1000 UAH, shown in the figure. 9.

Having determined the optimum number of services, the user can order the selected offers.

V. CONCLUSIONS

To organize the effective functioning of animation studio management services, taking into account the specifics of their work, the logic of query processing systems must be built using modern methods of Data Mining, namely the "Decision Tree" methods. In order to provide quality services, in a user-friendly format, it is most appropriate to provide information to the client in the context of two areas: price policy and content topics of services. If the client chooses as a priority , the direction - pricing policy, it is advisable to use the backtracking algorithm when choosing the client, as the priority , direction - the content of services - algorithms for building a decision tree.

Fig. 9. A list of the proposed "Active Entertainment" holiday option

REFERENCES

[1] Why tourism? World Tourism Organisation UNWTO, access mode: [http://www2.unwto.org/ content/why-tourism].

[2] Top 10 Digital Trends in 2018 year, Idea Digital Agency. 2018. 19 марта. [URL: https://ideadigital.agency/10-golovnih-digital-trendiv-u-2018-rotsi/].

[3] The Business of Theme Parks (Part II): How Much Do They Cost? And Earn? Retrieved from : [http://www.theparkdb.com/blog/the-business-of-theme-parks-part-ii-how-much-do-they-cost-and-earn/].

[4] Official web-page of State Statistics Service of Ukraine, 2019. Retrieved from : [http://www.ukrstat.gov.ua/].

[5] H. Gelter, "Digital Tourism – An analysis of digital trends in tourism and customer digital mobile behavior for the Visit Arctic Europe project", Summary. Interreg Nord, 2017.

[6] TheParkDatabase, World Amusement Parks statistics. Retrieved from: [http://www.theparkdb.com].

[7] S. Babushko, "Digitizing the Activity of Travel Agencies", Topical issues of tourism and tourism practice: Collection of materials of the scientific-practical conference on the 25th anniversary of the Institute of Tourism of the Federation of Trade Unions of Ukraine, Kiev: APSVT, pp. 13-15, April 18, 2019.

[8] O.I Artemenko, V.V Pasichnyk and V.V Egorova. "Information technologies in the field of tourism. Analysis of research applications and results" Bulletin of the National university "Lviv Polytechnic". Information systems and networks, № 814, 2015, pp. 3-22.

[9] N.I. Dumich and O.O. Smotr "Construction composite web application animation studio", in the Third All-Ukrainian Scientific and Practical Conference of Young Scientists, Students and Cadets "Information Protection in Information and Communication Systems", Lviv, pp. 268-269, November 28, 2019.

[10] S.M. Kylymystyj, "Classification of animation activities", International Bulletin: cultural studies, philology, musicology, Issue II (5), 2015, pp. 77-83.

[11] The Adult Price for a Children's Holiday: How Much Do Children's Parties Have? BBC News. Retrieved from: [https://www.bbc.com/ukrainian/entertainment/2015/04/150407children_party_dt].

[12] The official website of the children's animation studio "Children's Planet". Retrieved from: [https://dityacha-planeta.com/].

[13] The official website of the children's animation studio "Igroland". Retrieved from: [http://igroland.com.ua/ua].

[14] The official website of PAPASHON entertainment complex. Retrieved from: [https://papashon.com/].

[15] The official website of the event-agency "Empire of the Holiday" Retrieved from: [http://isvyata.org/].

T. Corman, Algorithms: Construction and Analysis, 3rd ed, Trans. with English, M.: LLC "I. D. Williams", 2013.

IEEE Third International Conference on Data Stream Mining & Processing
August 21-25, 2020, Lviv, Ukraine

Method of Ontology Use in OODA

Vasyl Lytvyn, Dmytro Dosyn, Victoria Vysotska
Information Systems and Networks Department
Lviv Polytechnic National University
Lviv, Ukraine
Vasyl.V.Lytvyn@lpnu.ua

Andrii Hryhorovych
Informatics and Information Systems Department
Drohobych Ivan Franko State Pedagogical University
Drohobych, Ukraine
a.hryhorovych@gmail.com

***Abstract*—The article examines the behavior of an intellectual agent in a competitive environment. An OODA loop is selected to model the behavior. The interaction of the stages of the OODA loop (observation, orientation, decision-making, action) with the ontology of the tasks and the domain within which this agent functions is considered.**

Keywords—OODA loop; observation; orientation; analysis; synthesis; decision making; action, ontology.

I. INTRODUCTION

According to the ideas of John Boyd and his followers, any activity in a competitive environment (for example, in the military field) with some degree of approximation can represented as an OODA cybernetic model [1]. This model involves repeated repetition of a loop of actions, consisting of four consecutive interacting processes (Fig. 1): orientation, decision, action. This model has successfully used to model activities and decision-making in business, politics, sociology, etc., in those areas where there is a competing party. According to Boyd's theory, every person or organization has its own decision-making and activity loop in solving tasks.

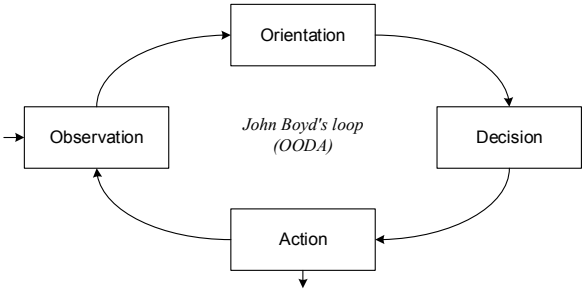


Fig. 1. OODA loop processes

Develop an approach to improve the performance of a management entity in a competitive environment. The purpose of the study is to develop and validate a method for modeling the OODA loop using the knowledge base ontologies of the environment in which the control object operates.

II. ANALYSIS OF RECENT RESEARCH AND PUBLICATIONS

According to the literature on the theory of J. Boyd, summarized in [1], the purpose of the loop stages and their functional purpose is as follows. Observation is the process of gathering the information needed to make a decision in a particular case. The necessary information can obtained from both external and internal sources. Internal sources of information are understood as feedback elements of the loop. As external sensors and other channels of information are used. In order for the observation to be scientific, it must:

- Systematic, not accidental;
- Carried out consistently and systematically;
- Sufficiently broad information was provided about the phenomenon being monitored (as many facts as possible should be operated on);
- provided accurate recording of observation results.

Data can collected:

- Mechanically. Mechanical registration of data is that the source of information, i.e. «event» or «phenomenon», manifests itself as a change of some physical state, and this new state is recorded mechanically.
- Expert observation, followed by the recovery of memory results, which is referred to as «recording».
- Through experimental research, the peculiarity of which is that the phenomenon (the subject of the study) is studied under different conditions and circumstances; the use of this method of research contributes to a deep and very accurate study of a certain psychological patterns.

Orientation is the most responsible and cognitively complex stage in the entire OODA cycle. The orientation stage consists of two sub-stages: destruction and creation. Destruction involves breaking the situation into smaller elemental parts that are easier to understand. The decision-maker or organization will seek to decompose the task to such a level until the newly created components of the task are close to the standard or typical situations for which the decision-maker (ODA) has an action plan. Being acquainted with these elementary typical sub-tasks is achieved through training, coaching, experience and coaching. ODA identifies the current situation, that is, assigns it to a specific sub-task, and applies a pre-prepared action plan for that sub-task. Then these constituent elemental subplots are combined into a general plan of action, which corresponds to the sub-stage of «creation». If there are no plans from which to choose a solution, then the process remains at the orientation stage and further decomposition of the task is carried out. Failure to develop a plan with a real chance of success can further grind to the last cycle. Orientation uses methods of analysis and synthesis, which are closely linked. They are designed to process information obtained through the application of research methods. Analysis is the study of the qualities, properties and characteristics of the object under study by its conditional separation into separate components. In turn, the

synthesis is to summarize information about individual components and to form a set of information data on the object of study as a whole. The results of the analysis and synthesis process are the basis for making various forecasts for the near and far future. Prediction can done by calculation and extrapolation methods.

At the stage of decomposition of the system is carried out:

- Defining and decomposing the overall purpose of the study and the main function of the system as a limitation of the trajectory in the space of states of the system or in the field of acceptable situations. Most often, decomposition is performed by building a goal tree and a function tree:

- Isolation of the system from the environment (division into «system» and «non-system»);
- Description of influential factors;
- Description of development trends;
- Description of the system as a «black box»;
- Functional (by function), component (by type of elements), structural (by type of relations between elements) decomposition of the system.

Decision-making is the third stage of the OODA cycle. If, at this stage, the ODA has been able to form only one real plan, then the decision is made whether to implement the plan or not. If several alternative options are formed, the ODA at this stage selects the best one for further implementation. Choosing the best plan can made on the criterion of value for money. With a time limit, the plan that meets the requirements of fast reliability is preferable. The following methods are used to make a decision:

- The cost-effectiveness method takes into account three stages: construction of the model of efficiency, construction of the model of cost, synthesis of cost and efficiency;
- Methods of theory and practice of reliability are based on the application of the apparatus of probability theory and random processes, mathematical statistics and modeling.

Action is the final stage of the cycle, which involves the practical implementation of the chosen course of action or plan. There are two main ways to achieve competitive advantage in pursuing different types of professional activity. The first way is to quantify your action cycles faster. This enables us to be the first to make decisions and make competitors react to our actions. The second way is to improve the quality of the decisions made, that is, to make decisions that are more relevant to the situation than the competitors are. Improving the quality of one's own decisions can achieved in various ways, which include the use of modern formal mathematical methods, the use of automated control systems, decision support systems, and expert systems. If the latter are used, the modern approach to their construction is used as the nucleus of the ontology knowledge bases [2]. Therefore, the task of developing ontology usage methods in the OODA loop arises.

III. BASIC MATERIAL

The distinguishing feature of the OODA cycle from other cyclical models is that in any situation, it is always assumed that there is a competitive side. In Fig. 2 lists three controls that are in some of their initial states and have their own states of purpose. The fact that these objects operate in a competitive environment, during the passage of the OODA loop, to analyze the states in which the competitors are and their actions, as shown in the figure by the appropriate arrows. It is proposed to use the OODA loop system to use an intelligent system whose core knowledge base is an ontology consisting of domain ontology and task ontology that may arise in this environment. In our opinion, the content of the ontology directly affects the 2nd and 3rd stages of the cycle, and the structure and content of the ontology depends on the 1st and 2nd stages (Fig. 3). Let's take a closer look at each stage of the OODA loop as it interacts with the domain ontology and the tasks that arise in this area. The observation phase enables the ontology development process to analyzed and analyzed to select the relevant information that is required in the next stages of the OODA loop. Formally, an ontology consists of terms (concepts, concepts) organized into a taxonomy, their definitions and attributes, as well as the associated axioms and rules of inference. Therefore, the model of ontology O is understood as three types [2]

$$O = \langle C, R, F \rangle, \tag{1}$$

where C is a finite set of concepts (concepts, terms) of software, R is finite set of relations between concepts (concepts, terms) of a given software, F is interpretation of concepts and relations (axioms). Axioms impose semantic constraints on the system of concepts and relations [3].

In order to build an ontology that adequately describes the semantic software model, it is necessary, first, to solve the tasks of obtaining knowledge from different sources to identify many concepts and establish a hierarchy on that set. Since much of the information is contained in natural-language texts (NLTs), it is promising to gain knowledge of textual information as well as intelligent processing of specially selected NLT collections. One of the most effective approaches to complete ontology is its automated teaching of natural language texts. Automated filling can implemented by analyzing text documents using a knowledge processor (Fig. 4). This approach is discussed in more detail in a monograph [4]. In the given scheme, the task of the linguistic processor is to perform its lexical, grammatical, syntactic and semantic analysis. As a result, the ontology is replenished with concepts, TLDs (subject - action - object) and cause and effect relationships between TLDs. The other part of the important relationship between concepts and their properties is established by the ontology processor, which builds an ontological structure for each concept obtained after analyzing the text.

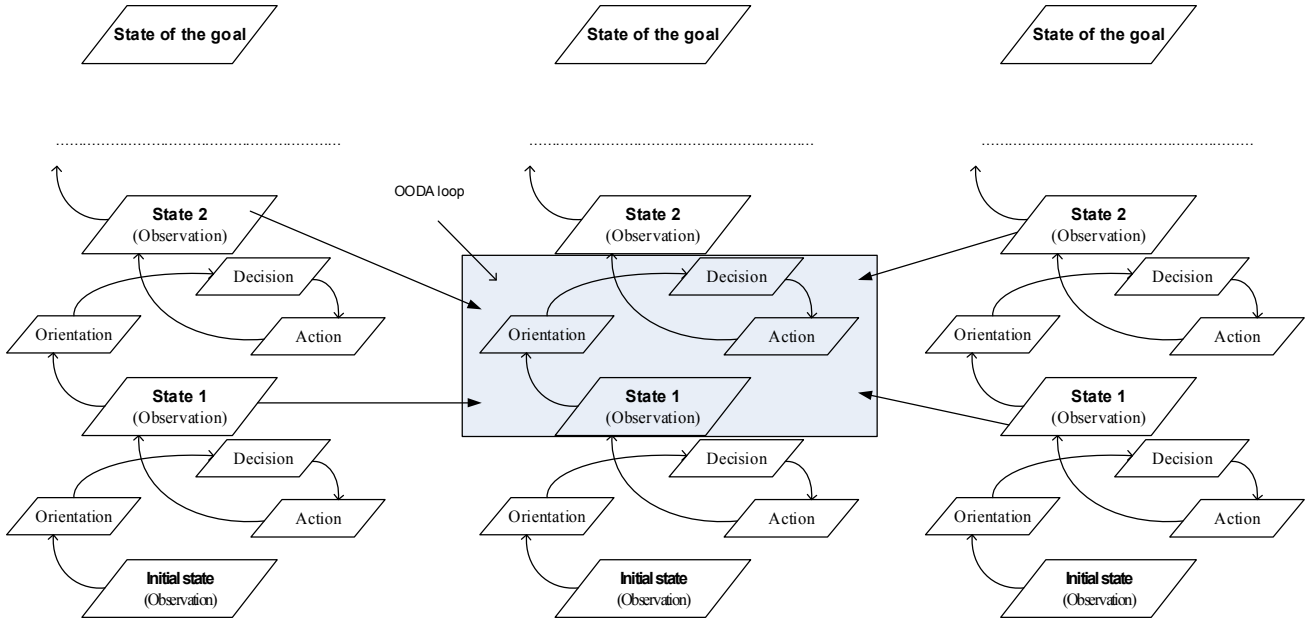


Fig. 2. Operation of facilities in a competitive environment

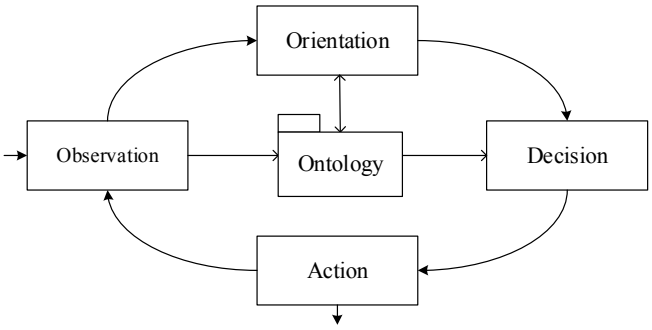


Fig. 3. Using ontologies in the OODA loop

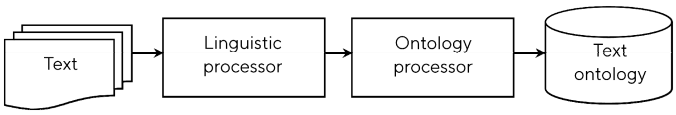


Fig. 4. Structural and functional diagram of the knowledge processor

The work of the ontology processor is supported by the corresponding database, the main components of which are: first, a set of rules, and second, a universal logical database MModWN [5] type WordNet [6]. The Knowledge Processor is used in a system of automated learning from text documents, which in turn is used to solve the task of semantic search in full-text databases. Among the systems developed in Ukraine, it should noted that the development of the staff of the Department of Mathematical Informatics at Taras Shevchenko National University of Kyiv is a system of text processing in natural language [7]. The system is designed to solve tasks such as analysis and synthesis of texts in natural language, automated generation of abstract text, automated indexing (definition of the subject) of the text. Our system is described in detail in [4, 8-12]. The main advantage of our approach is to build an Intelligent Agent (IA) that determines the value of the messages that are proposed to added to the ontology depending on the management plan chosen. The peculiarities of the functioning of a specialized IA are determined by its interest - a vector of estimates of the desirability of possible states of the agent. To describe the agent's interest, by which he distinguishes the states of the surrounding world and positions himself in it, a utility function is used, which is a numerical evaluation of his desirability for the agent [13-18]. Utilities are combined with action probabilities to determine the expected utility of each action. Suppose $U(S)$ the usefulness of the condition S from the point of view of the agent making the decision to perform some action A. An arbitrary undetermined action can cause a result state $Result_i(A)$ where the index i runs over all possible results. Before taking action A, the agent evaluates the likelihood of each of the possible outcomes $P(Result_i(A) | Do(A), E)$, where E is the set of parameters available to the agent of its state, and $Do(A)$ is the statement according to which in the current state the action is performed. Thus, it is possible to calculate the conditional utility of the action based on known status parameters:

$$EU(A \mid E) = \sum P(Result_i(A) \mid Do(A), E) \cdot U(Result_i(A)) \quad (2)$$

If the Maximum Expected Utility (MEU) guides a rational intellectual agent, it is forced to choose an action that

maximizes the expected utility for the agent. This is how a mechanism for motivating the behavior of a rational intellectual agent works, regardless of its scope. In the case of an information retrieval agent Maximum Expected Utility − *MEU*, its interest may be determined by evaluating the novelty of the messages received, which requires the use of methods of intellectual analysis of natural-language texts. We believe that the text is constructed as a message. The message structure is oriented to the perception of another agent, so it consists of two parts (Fig. 5): the ascertaining part, by which the addressee evaluates the relevance of the message (1) and defines its context (2), and the constructive part - potentially new to the reader in in this context (3).

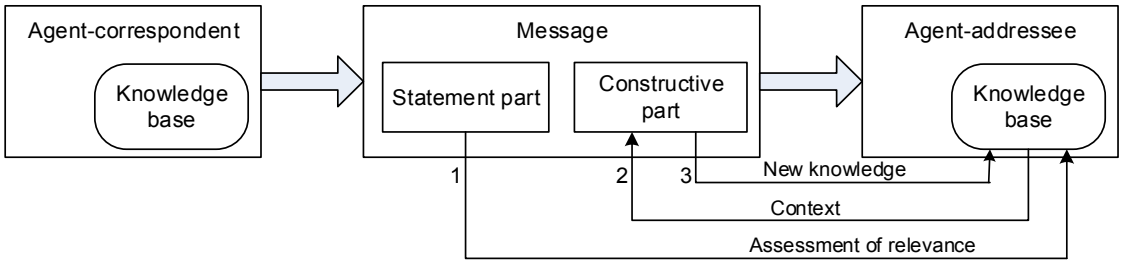


Fig. 5. Recognition of new knowledge in a message to fill their knowledge base

If the new knowledge is not a complete algorithm, but only a single fact or rule that clarifies the algorithms already known to the agent, changing their utility functions serves to evaluate the novelty of the fact (rule) and their importance to the agent. This approach is described in more detail in a monograph [9]. The representation of knowledge in the form of an ontology implies that any possible generalization, that is, a complex, complex concept, is always explicitly articulated, named, and appears as a separate concept in the knowledge base. At the orientation stage, a plan of action is drawn up. In order to reach the target state of IA, a plan to achieve this state with all possible alternatives must first be developed [10]. Planning is based on decomposition. The ZP planning task has three components: the set of states S , the set of actions A , the set of target states $Goal$ (goal states):

$$\text{ZP} = \langle St, A, Goal \rangle . \tag{3}$$

Further, we will assume that the goal state is the same. If there are several states of goal, then the goal can written as a disjunction of these states. Then achieving this state is a solution to some sub-task, so assuming the uniqueness of the goal state is normal. The action A , in turn, has four parts: action name, parameter list, prerequisite, and result. The plan itself is defined as a four-element tuple - <Multiple Actions, Multiple Organization Constraints, Multiple Causal Relationships, Multiple Prerequisites Multiple> [11]. The condition $S(i)$ is given as a set of facts with appropriate probabilistic estimates. The action is presented as a mapping from state $S(i)$ to state $S(j)$ with a corresponding probability p_{ij} , that is $S(i) \to S(j)$, a probability p_{ij} . The IA must be able to evaluate the condition in order to select the necessary actions. It is easier to do this with states in which he was already. It is more difficult to estimate future states. Heuristic functions or metacognition are used to evaluate them. So first, let's look at the assessment of the completed states, then the actions and finally their combination, leading to a new (future) state. Let $v(S(i))$ is an assessment of the condition $S(i)$. We will use software ontology to evaluate states that already had IA. The goal state $Goal$ is determined by the need for some set of attributes to reach certain values $z(x, Goal)\ \forall x \in X$. Each state $S(i)$ is defined by its many traits that acquire value $z(y, S(i))\ \forall y \in Y_i$. To evaluate a condition $S(i)$, it is necessary to map ψ the plurality of features and their status values $S(i)$ to the plurality of features and status values $Goal$. Obviously, the BR, namely the optional module of the Semantic Web Rule Language (SWRL) ontologies, should use such a mapping.

$$\psi : Y_i \xrightarrow{\hat{O}} X . \tag{4}$$

Then an assessment of the condition $v(S(i))$ is calculated

$$v(St(i)) = d(S(i), Goal) = \sum_{x \in X_W} \varphi(z(\psi(y), S(i)), z(x, Goal)) . \tag{5}$$

where X_W is many essential features. For example, to determine the weight of the elements of ontology [4, 12]. φ is a certain metric that depends on the specifics of the software [4, 12]. In our research for choosing IA actions, we will rely on the agent's rationality as an effort to minimize the cost of resources to reach the end state. Therefore, we assume that each action a_{ij} is uniquely determined by the cost of resources g_{ij}^k (cost of transition from state to state), where $k = 1, 2, ..., n_i$ - the number of alternatives α_k for making the transition a_{ij} . Therefore, in the future, we denote the action by three indexes: a_{ij}^k transition from state $S(i)$ to state $S(j)$, using the alternative α_k [13].

Because the lower the score, the better, the action estimate is directly proportional to the cost of resources, that is:

$$v(a_{ij}^k) = E \cdot g_{ij}^k , \tag{6}$$

where E is a scalar value that reduces the measurement of an action estimate to a single dimension with a state estimate. Choosing the best plan is at the decision stage. In general, the decision to choose an action based on an alternative is made according to some relationship between condition and action:

$$o_i\left(a_{ij}^k\right)=\delta\left(v\left(a_{ij}^k\right),v\left(S\left(j\right)\right)\right). \quad (7)$$

In particular, this relationship can be linear:

$$o\left(a_{ij}^k\right)=\omega v\left(a_{ij}^k\right)+\left(1-\omega\right)v\left(S\left(j\right)\right). \quad (8)$$

where $\omega\in\left[0,1\right]$ is part of the alternative action that the IA gives when making the decision, the other share belongs to the state in which it will pass.

After evaluating actions and states, the path selection task is reduced to the task of asynchronous dynamic programming [14]. We get the following model of transitions between states:

$$S\left(j\right)=a\left(S\left(i\right),o_i\right) \quad (9)$$

with the optimization criterion

$$S\left(j\right)=a\left(S\left(i\right),o_i\right) \quad (9)$$

$$\Theta\left(S\left(0\right),\vec{o}\right)\Rightarrow\min(\max). \quad (10)$$

Task (9) - (10) is a dynamic programming task. Using methods suitable for the solution of such tasks, we find the solution in the form of a path from transition to the initial state, that is, the path of the plan. In our further works, it is planned to complicate model (9) - (10) taking into account the states of competitors and to use the developed mathematical support for the modeling of military actions.

IV. CONCLUSIONS

This paper deals with the modeling of the behavior of an intelligent agent that functions in a competitive environment. An OODA loop is selected to model the behavior. To improve the efficiency of the OODA cycle, it is proposed to use the ontology of the domain within which the intelligent agent operates and the ontology of the tasks that arise in this field. The influence of the stages of the OODA loop on the ontology content and vice versa - the ontology content on the course of the stages is determined. The interaction between the ontology and the stages of the OODA loop is suggested. The task of planning the activities of an intellectual agent in a competitive environment is reduced to the task of dynamic programming.

[1] V. Lytvyn, V. Vysotska, D. Dosyn, O. Lozynska, and O. Oborska, "Methods of Building Intelligent Decision Support Systems Based on Adaptive Ontology," IEEE 2nd International Conference on Data Stream Mining and Processing, DSMP, 2018, pp. 145-150.

[2] T. Gruber, "A translation approach to portable ontologies," Knowledge Acquisition, 5 (2),1 993, pp. 199-220.

[3] S. Babichev, "An Evaluation of the Information Technology of Gene Expression Profiles Processing Stability for Different Levels of Noise Components," Data, vol. 3 (4), 2018, art. no. 48.

[4] S. Babichev, B. Durnyak, I. Pikh, and V. Senkivskyy, "An Evaluation of the Objective Clustering Inductive Technology Effectiveness Implemented Using Density-Based and Agglomerative Hierarchical Clustering Algorithms," Advances in Intelligent Systems and Computing, vol. 1020, 2020, pp. 532-553.

[5] S. Sachenko, S. Rippa, and Y. Krupka, "Pre-Conditions of Ontological Approaches Application for Knowledge Management in Accounting," International Workshop on Antelligent Data Acquisition and Advanced Computing Systems: Technology and Applications, 2009, pp. 605-608.

[6] G.A. Miller, "WORDNET: A lexical database for English," Communications of ACM, 11, 1995, pp. 39-41.

[7] M. Davydov, and O. Lozynska, "Mathematical method of translation into Ukrainian sign language based on ontologies," Advances in Intelligent Systems and Computing, vol. 871, 2018, pp. 89-100.

[8] V. Lytvyn, V. Vysotska, D. Dosyn, and Y. Burov, "Method for ontology content and structure optimization, provided by a weighted conceptual graph," Webology, vol. 15(2), 2018, pp. 66-85.

[9] V. Lytvyn, Y. Burov, P. Kravets, V. Vysotska, A. Demchuk, A. Berko, Y. Ryshkovets, S. Shcherbak, and O. Naum, "Methods and Models of Intellectual Processing of Texts for Building Ontologies of Software for Medical Terms Identification in Content Classification," CEUR Workshop Proceedings, Vol-2362, 2019, pp. 354-368.

[10] Y. Burov, V. Vysotska, and P. Kravets: "Ontological approach to plot analysis and modeling," CEUR Workshop Proceedings, Vol-2362, 2019, pp. 22-31.

[11] I. Khomytska, V. Teslyuk, A. Holovatyy, and O. Morushko, "Development of methods, models, and means for the author attribution of a text," Eastern-European Journal of Enterprise Technologies, vol. 3(2-93), 2018, pp. 41-46.

[12] N. Antonyuk, L. Chyrun, V Andrunyk., A. Vasevych, S. Chyrun, A. Gozhyj, I. Kalinina, and Y. Borzov, "Medical News Aggregation and Ranking of Taking into Account the User Needs," CEUR Workshop Proceedings, Vol-2362, 2019, pp. 369-382.

[13] L. Chyrun, L. Chyrun, Y. Kis, and L. Rybak, "Automated Information System for Connection to the Access Point with Encryption WPA2 Enterprise," Lecture Notes in Computational Intelligence and Decision Making, 1020, 2020, pp. 389-404.

[14] Y. Kis, L. Chyrun, T Tsymbaliak., and L. Chyrun, "Development of System for Managers Relationship Management with Customers," Lecture Notes in Computational Intelligence and Decision Making, 1020, 2020, pp. 405-421.

[15] L Chyrun., A. Kowalska-Styczen, Y. Burov, A. Berko, A. Vasevych, I. Pelekh, and Y. Ryshkovets, "Heterogeneous Data with Agreed Content Aggregation System Development," CEUR Workshop Proceedings, Vol-2386, 2019, pp. 35-54.

[16] L. Chyrun, Y. Burov, B. Rusyn, L. Pohreliuk, O. Oleshek, A. Gozhyj, I. Bobyk, "Web Resource Changes Monitoring System Development," CEUR Workshop Proceedings, Vol-2386, 2019, pp. 255-273.

[17] A. Gozhyj, L. Chyrun, A. Kowalska-Styczen, and O. Lozynska, "Uniform Method of Operative Content Management in Web Systems," CEUR Workshop Proceedings, Vol-2136, 2018, pp. 62-77.

[18] L. Chyrun, A. Gozhyj, I. Yevseyeva, D. Dosyn, V. Tyhonov, and M. Zakharchuk, "Web Content Monitoring System Development," CEUR Workshop Proceedings, Vol-2362, 2019, pp. 126-142.

IEEE Third International Conference on Data Stream Mining & Processing
August 21-25, 2020, Lviv, Ukraine

Information Resources Analysis System of Dynamic Integration Semi-Structured Data in a Web Environment

Andrii Berko, Irina Pelekh
Information Systems and Networks Department
Lviv Polytechnic National University
Lviv, Ukraine
Andrii.Y.Berko@lpnu.ua

Lyubomyr Chyrun, Ivan Dyyak
Programming Department
Ivan Franko National University of Lviv
Lviv, Ukraine
Lyubomyr.Chyrun@lnu.edu.ua

***Abstract*— The article describes the procedure for analyzing information resources of the system of dynamic integration of weakly structured data in the web-environment to determine the common features of information resources and identify links between them. Model-based shared resource definitions and language describing the rules of access to resources, examines the process of creating an object adapter defining the common features of the information resources and identify relationships between them using the rules of the method of "black box". The process of determining the structural-dynamic model of the domain Mashup-application is described.**

Keywords—information resource; dynamic integration; Mashups-application; object adapter; "black box" method; agent-oriented approach

I. INTRODUCTION

The growing number of resources and services available on the Internet and the opportunities that Web 2.0 offers are pushing end users to evolve from passive consumers of information into information producers who are able to access and manipulate existing information resources in order to generate new content. In terms of human-machine interaction, this requires new communication paradigms that should allow people to access content, move it to personal interactive workspaces where they can integrate it, and, if necessary, create new content using existing content. The solution of this problem requires the use of methods and means of dynamic data integration. The data integration system frees users from the need to know which data sources, other than integrated, they use, what properties of these sources and how to access them [1-4]. Today, due to the rapid development of the Internet, a huge number of information resources of various nature and content has accumulated and is constantly growing [5-9]. This raises the problem of their operational, dynamic integration, while highlighting the content of information and preserving the semantics of data [10-13]. The complexity and nature of the methods used to solve this problem depends significantly on the level of integration that needs to provided, the properties of individual data sources and the totality of sources as a whole and the necessary methods of integration [14-19]. Due to the rapid growth of Ukrainian and world information resources, methods of determining the common features of information resources and dynamic integration of information are in the focus of researchers of federated environments and are of scientific and practical interest to developers of modern distributed intelligent information systems, but numerous problems still remain unresolved. Therefore, at present especially urgent task to develop new approaches, technologies, architectural solutions and flexible tools for analyzing the information resources of the system of dynamic integration semi-structured data in a web environment for defining common information resources and identify relationships between them.

II. ANALYSIS OF RECENT RESEARCH AND PUBLICATIONS

Speaking about the dynamic integration of data in the web-environment, the research and development of data integration systems, working using the technology of Mash-Up [4, 20-23]. Mashups are an application development approach that allows users to combine data from multiple sources into a single integration tool. [4, 24-26] Some Meshes can created by a simple combination of JavaScript code with XML, which makes it possible to build a new innovative web service. Other, larger Meshes, which are the basis of relevant web sites, use the technology of services such as Google Maps and an address database, linking them together and showing project information on a map. Unlike web service compositions, where the focus is only on combining business services, the Mashup framework is expanding, which has more functionality and allows for the integration of heterogeneous resources, such as data services. Programs created with the help of Mash-Up technology are called Mashups or Mashup applications. Research shows that there is a high level of interest within Mashup. In addition, with the rapid development of IT-technologies, the needs for dynamic integration of heterogeneous data in the web-environment become especially urgent. Mashups technology opens up new and wide opportunities for data transmission to consumers. However, the fact remains unchanged that the user of Mashups should know, at least, how to write program code using programming languages (for example, Java Script, XML / HTML), how to use various web APIs. In order to solve this problem, there is a considerable amount of effort that is put into developing tools that are designed to support users with little programming knowledge to develop Mashups applications [27-32].

Since the content, nature and structure of a Mashup application cannot fully known in advance, because the Mashup is formed as a response to a user request. Therefore, it is advisable to choose such a method of constructing the

978-1-7281-3214-3/20/$31.00 ©2020 IEEE

functioning of the system, which would most correspond to the above criteria. Therefore, we propose the use of a combination of object-oriented approach to modeling and the method of "black box". According to the method of "black box" and agent-oriented approach [7] we perform the analysis of information resources of the system of dynamic integration of weakly structured data in the web-environment to determine the common features of information resources in the web-environment in order to dynamically integrate weakly structured data using mashups. A prominent feature of the agent-oriented approach is that the basic knowledge about the environment of the local agent should not contain all the information about all the information resources-a complete model. Tasks, algorithms, specifications of objects required to perform the task are loaded from the agent server in the form of a script and the desired representation - a slice of the domain model. Data representation is called a form of General knowledge representation of the subject area [7]. The representation of the same real information objects in different information systems may be different [33-36]. There are three main classes of data models: structured data model (deterministic data schema); unstructured data model (nondeterministic data schema); mixed data models (partially deterministic schema). As a type of model for designing a multi-agent system in order to exchange information between Autonomous distributed information resources, it is advisable to use a model with a partially deterministic data schema [37-45]. This model type is characterized by following properties:

- The set of types of properties and of relationship types of object with other objects dynamically only at the time of appearance of each real object in focus of the information system;
- Partially deterministic schema;
- Can be displayed not only the General properties and relations of domain objects, but also the individual;
- The scheme being completed rapidly with the arrival of a new fact that is some properties and relationships.

The subject area under study is the web-environment, where the elements are their own logical representations of the semantics and structure of the information object. A conceptual model of the web environment needs to provide [46-52]:

- Representations of data;
- Means and methods of representation of semantic rules that restrict the set of valid States (the data space);
- Means of identification data to select a subset of objects of the information system required for analysis of specific problems;
- Handling data.

A data structure is generally understood as a set consisting of data elements and a set of relationships between them. This definition covers all possible approaches to data structuring, but in each specific task, certain aspects of it are used [7, 53-57]. Considering a data structure without considering its representation in computer memory is called an abstract or logical structure. There are the following levels of globalism of logical structures in the computing environment: global networks; local networks; a separate personal computer; logical disks in a personal computer; logical folders; files; file elements; data elements. In turn, combinations of such logical structures form more of complex composite logical structures. There are simple (basic, primitive) and integrated (structured, composite, complex) data structures (types). From a logical point of view, simple data are indivisible units. Integrated data structures are those whose constituent parts are other data structures - simple or, in turn, integrated. These logical data structures are called domain entities (types). All entities are in some relationship with each other, their number and the number of relationships between them can be quite large depending on the point of view of the subject area. In addition, entities can have quite a few properties-characteristics depending on the domain model. Relationships between entities can also have as many properties as desired.

Some different (physical and logical) information objects may have common characteristic properties. Therefore, it is advisable to consider not only instances of entities (specific information objects), but a more abstract entity, which in a certain way generalizes these instances into a set with characteristic properties common to all elements - types of objects. For example, the text document type = {{*.Doc}, {*.Txt}, ...}- an artificially created type that has a common property for all instances-functionality = working with texts. It should noted that the same information object could simultaneously temporarily belong to several types of objects and participate in logical relationships with other entities in different ways, depending on the belonging to the type. This representation of the ownership structure or an instance of the structure to some type allows us to consider the domain as dynamic, and the transition from type to type - as a transition between representations of the domain model (SOFTWARE) depending on the focus of the task or user, that is, the transformation of the SOFTWARE model. "Black box" is a term used in engineering and Cybernetics to refer to an object or system whose principles of operation are unknown, except that a certain input signal corresponds to a certain output signal [2]. "Black box" models allow you to reflect the inputs and outputs of the system, which are necessary to study one of the sides of its functioning, so called models "input-output". When constructing such a model, the relationship between these inputs and outputs is established. The input-output model reflects the basic properties of the system, such as integrity and relative isolation through the presence of communication with the external environment [2]. When constructing an input-output model, the problem is to determine the inputs and outputs that need to be included in the model, since the model is constantly modified when studying the system. The real system interacts with the environment through an infinite number of ways, that is, through an infinite number of inputs and outputs. The criterion for selecting these inputs and outputs is the purpose of the system, the materiality of a system connection with the environment. In the model we are forced to display a finite number of interactions and thus there is a high probability of not including exactly those inputs and outputs that most significantly determine the properties of the system.

The problem of choosing the most important inputs and outputs is expressed in the fact that the connections with

the external environment that are not included in the system do not disappear, but in a certain way act independently of us and thus complicate the study of the behavior of the system. When constructing input-output models, unaccounted-for and unknown connections with the external environment are represented using simplified uncertainty models [2]. Modern approaches to the construction of input-output models are based on one of these forms of uncertainty stochastic, set-theoretic or a combination of them. In these cases, the system model is considered as a "black box" (Fig. 1) [2] with the following parameter groups: control (input) parameters X_i, which are called factors; initial parameters Y_i, which are called status parameters; W_i parameters-effects.

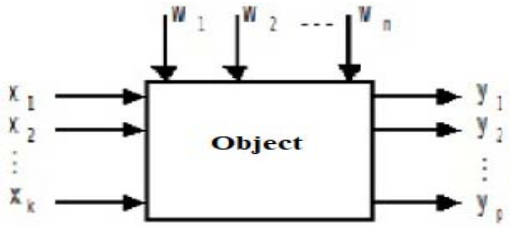


Fig. 1. System model in the form of a " black box " [2]

As noted in [2] the methods of identification theory are used to construct the model. In General, the identification problem is formulated as follows: based on the results of observation of the input and output variables of the system, it is necessary to build an optimal mathematical model in some sense. The main steps of identification are:

- The choice of class and the structure of the model and language of her descriptions; the choice of class and types of input X;
- Substantiation of criteria of similarity systems and models;
- The choice of method for the identification and development of appropriate algorithms for estimating the parameters of the system;
- Verification of adequacy of the model obtained because of identification.

III. Highlighting Problems and Statement of Purpose

Creating a model of the system of dynamic integration of weakly structured data in a web environment using mashup technology requires, first, work related to the analysis of information resources of the Mashup system to identify common features and identify links between them, using the correct methodology. The relevance of the task of determining the common features of information resources and identifying the links between them, distributed web-environment, is confirmed by the existence of a large class of tasks. These are the tasks of monitoring, integration, support and integrity of the regional information space, including the problems associated with the dynamic integration of heterogeneous information resources. Thus, the task of developing new approaches, technologies, architectural solutions and flexible tools for dynamic integration of weakly structured data in the web environment is urgent. Therefore, the introduction of new methodologies to identify common features of information resources and identify links between them in order to facilitate the storage, organization, access and analysis of information online is promising and constructive in terms of solving the problems. The aim of the work is to use existing approaches for the formation and analysis of the definition of common features of information resources and identify links between them in order to create a model of the system that allows end users to obtain unified information from various heterogeneous data sources.

IV. Analysis of the Obtained Scientific Results

Work with information resources, as with a set of data stored in computer systems, for the purpose of further effective use of identification, extraction, processing, analysis, exchange, within the tasks of owners of information resources, includes: storage;

- Search for the necessary information resource; automation of access to information resources;
- Extraction of information in a unified form; data processing (tasks); data analysis;
- Definition of common features of information resources; information exchange;
- Control of availability of all necessary resources;
- Control of all stages of work with the information resource (transaction monitoring);
- Output of the result in the specified form (data / task).

Exchange of results of completed tasks includes

- Determination of exchange participants (identification task);
- Determination of resource existence and access permissions to it (monitoring task);
- Verification of successful completion of the exchange transaction.

The task of identifying information resources or some information object (hereinafter - the object) includes several stages, among which are the main:

- Tasks the conditions for the allocation of some object among the totality of objects - definition of search image;
- Object recognition process-verification of compliance with the conditions of the search image of the characteristics of the scanning object.

The main tools used to ensure the integration of information resources include data converters, integrating data models, data model mapping mechanisms, object adapters (Wrappers), mediators(Mediators), ontological specifications, means of integration of schemes and integration of ontological specifications, as well as the architecture that ensures the interaction of tools used in a particular system of resource integration. Object adapters (Wrappers) are components associated with data sources to solve technical heterogeneity and metamodeling heterogeneity problems. Their main functions: accept source requests in some language, turn the request into the source language, execute the request, and send the results to the mediator. According to the "black box" approach proposed in [2], we analyze the information resources of a Mashup system in a web environment in order to determine the

common features of information resources and formulate links between them. In addition, complex objects can contain a set of simple objects, for example, a person can have one or more addresses. Based on the General resource definition model and the resource access rules description language, consider the process of creating an object adapter using the black box method rules. The automatic object adapter is based on three modules:

- Resource description interpreter-used to generate a resource data structure within the system without reading the resource description;
- Mapping device-used to generate internal mapping rules by reading the comparison configuration;
- Intercept device-used to intercept system operations according to resource and map information, access a new resource using transform rules, and finally to return mapping results to the system.

Let's take a closer look at the work of the object adapter to provide dynamic integration of information resources. As shown in Fig. 2 the object adapter consists of two subsystems: the object component and the resource conversion description configuration component.

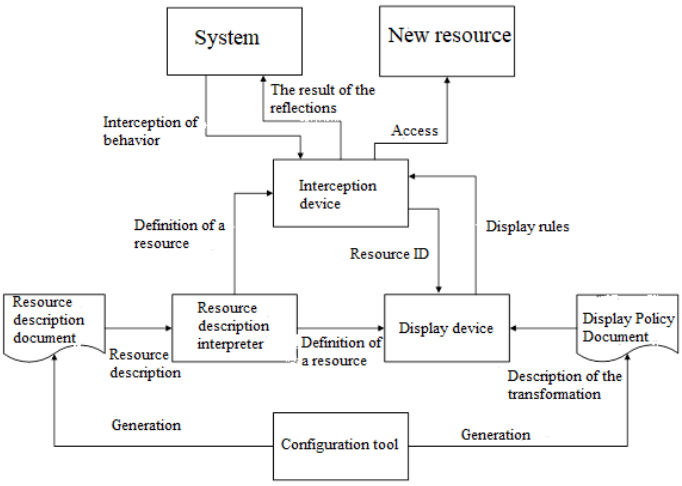


Fig. 2. Object adapter architecture identify common features of information resources and identify relationships between them

The object component performs multiple object mining or system migrations according to resource and mapping information and automatically generates reusable objects or a transformed system. The configuration component is used to generate the resource description and display rules that the object component needs. Fig. 2 shows the architecture of the object adapter to identify common features of information resources and identify links between them and describes the cooperation and the main data flow between modules. The external adapter configuration tool is used to generate a resource description and display rules. According to differences in functionality, business logic, data sources, etc., software systems have different resource access behaviors. Even with the same functionality and business logic, software systems still vary depending on differences in the experience, skills and personal practices of designers and developers. However, no matter how complex and diverse web-based systems are, operations at the system level are consistent, consistent and according to certain standards. The operating system must provide a complete set of operations to access all resources on the platform-a set of operation keys. The set of operation keys must take into account different operating systems, for example, a set of actions in the form of system calls in Unix/Linux or as an API in a system-level DLL file on Windows. Thus, the problem of intercepting the behavior of a software resource can solved by intercepting a set of operation keys. According to the method of "black box" and agent-oriented approach, we present a structural-dynamic model of the domain Mashup-application at a discrete time t. Let the information object is an information resource. Let us have an information resource with the following characteristics:

$$O_i = \langle D_i, T_i, L_i, F_i \rangle, \quad (1)$$

where D is the direction, valid values: {input, output}, T is the type of information resource with a set of allowable values: {pipe, file, network, DB}, L is the location information and the authentication information to access the resource, F is format of the information resource that is used to describe the syntax and semantics of the information resource.

The format of the information resource is defined as follows:

$$F_i = \langle F_{standart}, F_{description} \rangle, \quad (2)$$

where $F_{standart}$ is an indication of a standard resource type. Its value is NULL if the resource format is not of the standard type, $F_{description}$ is used to describe the format defined by the user of the resource. Its value is NULL for the resource in standard format.

Given the formula (1) the phase trajectory (state) of the I-th information resource at each discrete time t will have the form:

$$\varphi_i(t) = O_i(t)(D_i(t), T_i(t), L_i(t), F_i(t)). \quad (3)$$

Given that the phase space of the information resource can be represented as a set of States of the information resource, which at each time t describe a tuple of values of the characteristics of the information resource, the structural-dynamic model of the domain of the Mashup application at a discrete time t will have the following form:

$$\langle O(t), R(t), B(t) \rangle = \langle O(t), R(O(t)), B(O(t)) \rangle, \quad (4)$$

where O is a set of object types, R is a set of relations between object types, B is a set of valid operations on object types, and new object types, new relations, new rules of behavior at time t can arise.

The description of an information object can characterized by a set of properties:

- Complete (abstract) set of all possible properties-universe of properties;

- A subset of the universe of object properties essential for solving some problem.

Each task defines a set of fixed object properties, thereby determining the choice of object type. A parameter is a logical implementation of a property of an information object, so it is worth distinguishing between the parameter value and the parameter type. Different types (interpretations) can represent the same property. Four basic operations can performed on all logical structures:

- Creation is to allocate memory for the data structure;
- Destruction is the opposite in its actions before the creation operation;
- Selection is to access data within the structure itself;
- Update is allows you to change the data values in the data structure.

V. CONCLUSIONS

The order of the analysis of information resources of the system of dynamic integration of weakly structured data in the web-environment is described. The characteristic of the information object as an information resource is given and the description of its most important characteristics is given. The work of the object adapter is described and recommendations on its use for the process of determining the common features of information resources and identifying links between them are given. According to the method of "black box" and agent-oriented approach, the structural-dynamic model of the domain of Mashup-application at a discrete time *t* is depicted.

[1] R. Baumgartner, G. Gottlob, M. Herzog, and W. Slany, "Interactively adding web service interfaces to existing web applications," Int. Symposium on Applications and the Internet, 2004, pp. 74-80.

[2] G. Canfora, A.R. Fasolino, G. Frattolillo, and P. Tramontana, "Migrating interactive legacy systems to web services," IEEE CS Press, editor, European Conference on Software Mainteinance and Reengineering, 2006, pp. 23-32.

[3] H. Guo, C. Guo, F. Chen, and H. Yang, "Wrapping client-server application to web services for internet computing," International Conference on Parallel and Distributed Computing, Applications and Technologies, 2005, pp. 366–370.

[4] A.Y. Berko, and K.A. Aliekseyeva, "Quality evaluation of information resources in web-projects," Actual Problems of Economics, vol. 136(10), 2012, pp. 226-234.

[5] A. Laender, B. Ribeiro-Neto, and A. Silva, "A brief survey of web data extraction tools," SIGMOD record, 31(2), 2002.

[6] Y. Jiang and E. Stroulia, "Towards reengineering web sites to web-services providers," IEEE CS Press, editor, European Conference on Software Mainteinance and Reengineering, 2004, pp. 296–305.

[7] G. Wagner, "The Agent-Object-Relationship Metamodel: Towards a Unified View of State and Behavior," Information Systems, 28(5), 2003, pp. 475-504.

[8] V. Vysotska, and L. Chyrun, "Methods of information resources processing in electronic content commerce systems," Proceedings of 13th International Conference: The Experience of Designing and Application of CAD Systems in Microelectronics, CADSM, 2015.

[9] V. Vysotska, and L. Chyrun, "Analysis features of information resources processing," Proceedings of the International Conference on Computer Sciences and Information Technologies, CSIT, 2015, pp. 124-128.

[10] V. Andrunyk, L. Chyrun, and V. Vysotska, "Electronic content commerce system development," Proceedings of 13th International Conference: The Experience of Designing and Application of CAD Systems in Microelectronics, CADSM 2015-February, 2015.

[11] K. Alieksieieva, A. Berko, and V. Vysotska, "Technology of commercial web-resource processing," Proceedings of 13th International Conference: The Experience of Designing and Application of CAD Systems in Microelectronics, CADSM 2015-February, 2015.

[12] V. Vysotska, R. Hasko, and V. Kuchkovskiy, "Process analysis in electronic content commerce system," Proceedings of the International Conference on Computer Sciences and Information Technologies, CSIT, 2015, pp. 120-123.

[13] V. Lytvyn, V. Vysotska, O. Veres, I. Rishnyak, and H. Rishnyak, "The Risk Management Modelling in Multi Project Environment," Proceedings of the International Conference on Computer Sciences and Information Technologies, CSIT, 2017, pp. 32-35.

[14] V. Vysotska, V. Lytvyn, Y. Burov, A. Gozhyj, and S. Makara, "The consolidated information web-resource about pharmacy networks in city," CEUR Workshop Proceedings, 2018, pp. 239-255.

[15] V. Vysotska, V.B. Fernandes, and M. Emmerich, "Web content support method in electronic business systems," CEUR Workshop Proceedings, Vol-2136, 2018, pp. 20-41.

[16] A. Gozhyj, I. Kalinina, V. Vysotska, and V. Gozhyj, "The method of web-resources management under conditions of uncertainty based on fuzzy logic," Proceedings of the International Conference on Computer Sciences and Information Technologies, CSIT, 2018, pp. 343-346.

[17] A. Gozhyj, V. Vysotska, I. Yevseyeva, I. Kalinina, and V. Gozhyj, "Web Resources Management Method Based on Intelligent Technologies," Advances in Intelligent Systems and Computing, 871, 2019, pp. 206-221.

[18] 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, 689, Springer, 2018, pp. 204-216.

[19] M. Korobchinsky, V. Vysotska, L, Chyrun, and L. Chyrun, "Peculiarities of Content Forming and Analysis in Internet Newspaper Covering Music News," Proceedings of the International Conference on Computer Sciences and Information Technologies, 2017, pp. 52-57.

[20] T. Batiuk, V. Vysotska, and V. Lytvyn, "Intelligent System for Socialization by Personal Interests on the Basis of SEO-Technologies and Methods of Machine Learning,"CEUR workshop proceedings, Vol-2604, 2020, pp. 1237-1250.

[21] V. Lytvyn, A. Gozhyj, I. Kalinina, V. Vysotska, V. Shatskykh, L. Chyrun, Y. Borzov, "An intelligent system of the content relevance at the example of films according to user needs," CEUR Workshop Proceedings, Vol-2516, 2019, pp. 1-23.

[22] P. Kravets, V. Lytvyn, V. Vysotska, and Y. Burov, "Promoting training of multi-agent systems," CEUR Workshop Proceedings, Vol-2608, 2020, pp. 364-378.

[23] B. Rusyn, V. Vysotska, and L. Pohreliuk, "Model and architecture for virtual library information system," Proceedings of the International Conference on Computer Sciences and Information Technologies, CSIT, 2018, pp. 37-41.

[24] V. Lytvyn, and V. Vysotska, "Designing architecture of electronic content commerce system," Proceedings of the Internat. Conference on Computer Sciences and Information Technologies, 2015, pp. 115-119.

[25] O. Naum, L. Chyrun, O. Kanishcheva, and V. Vysotska, "Intellectual System Design for Content Formation," Proceedings of the International Conference on Computer Sciences and Information Technologies, CSIT, 2017, pp. 131-138.

[26] J. Su, A. Sachenko, V. Lytvyn, V. Vysotska, and D. Dosyn, "Model of Touristic Information Resources Integration According to User Needs," Proceedings of the International Conference on Computer Sciences and Information Technologies, CSIT, 2018, 113-116.

[27] V. Lytvyn, V. Vysotska, D. Dosyn, and Y. Burov, "Method for ontology content and structure optimization, provided by a weighted conceptual graph," Webology, vol. 15(2), 2018, pp. 66-85.

[28] V. Lytvyn, V. Vysotska, D. Dosyn, O. Lozynska, and O. Oborska, "Methods of Building Intelligent Decision Support Systems Based on Adaptive Ontology," IEEE 2nd International Conference on Data Stream Mining and Processing, DSMP, 2018, pp. 145-150.

[29] V. Lytvyn, Y. Burov, P. Kravets, V. Vysotska, A. Demchuk, A. Berko, Y. Ryshkovets, S. Shcherbak, and O. Naum, "Methods and Models of Intellectual Processing of Texts for Building Ontologies of Software for Medical Terms Identification in Content Classification," CEUR Workshop Proceedings, Vol-2362, 2019, pp. 354-368.

[30] N. Antonyuk, A. Vysotsky, V. Vysotska, V. Lytvyn, Y. Burov, A. Demchuk, I. Lyudkevych, L. Chyrun, S. Chyrun, and I. Bobyk, "Consolidated Information Web Resource for Online Tourism Based on Data Integration and Geolocation," Proceedings of the International Conference on Computer Sciences and Information Technologies, CSIT, 2019, pp. 15-20.

[31] A. Vysotsky, V. Lytvyn, V. Vysotska, D. Dosyn, I. Lyudkevych, N. Antonyuk, O. Naum, A. Vysotskyi, L. Chyrun, and O. Slyusarchuk, "Online Tourism System for Proposals Formation to User Based on Data Integration from Various Sources," Proceedings of the International Conference on Computer Sciences and Information Technologies, CSIT, 2019, pp. 92-97.

[32] A.Y. Berko, "Models of data integration in open information systems," Actual Problems of Economics, (10), 2010, pp. 147-152.

[33] A.Y. Berko, "Methods and models of data integration in E-business systems," Actual Problems of Economics (10), 2008, pp. 17-24.

[34] A. Berko, "Consolidated data models for electronic business systems," The Experience of Designing and Application of CAD Systems in Microelectronics, CADSM, 2007, pp. 341-342.

[35] M. Davydov, and O. Lozynska, "Mathematical method of translation into Ukrainian sign language based on ontologies," Advances in Intelligent Systems and Computing, vol. 871, 2018, pp. 89-100.

[36] S. Babichev, "An Evaluation of the Information Technology of Gene Expression Profiles Processing Stability for Different Levels of Noise Components," Data, vol. 3 (4), 2018, art. no. 48.

[37] S. Babichev, B. Durnyak, I. Pikh, and V. Senkivskyy, "An Evaluation of the Objective Clustering Inductive Technology Effectiveness Implemented Using Density-Based and Agglomerative Hierarchical Clustering Algorithms," Advances in Intelligent Systems and Computing, vol. 1020, 2020, pp. 532-553.

[38] O. Kuzmin, and M. Bublyk, "Economic evaluation and government regulation of technogenic (man-made) damage in the national economy," Computer sciences and information technologies, 2016, pp. 37-39.

[39] M.I. Bublyk, and O.M. Rybytska, "The model of fuzzy expert system for establishing the pollution impact on the mortality rate in Ukraine," Computer sciences and information technologies, 2017, pp. 253–256.

[40] M. Bublyk, O. Rybytska, A. Karpiak, and Y. Matseliukh, "Structuring the fuzzy knowledge base of the IT industry impact factors," Computer sciences and information technologies, 2018.

[41] N. Chukhray, N. Shakhovska, O. Mrykhina, M. Bublyk, and L. Lisovska, "Consumer aspects in assessing the suitability of technologies for the transfer," Computer sciences and information technologies, CSIT, 2019, pp. 142–147.

[42] N. Chukhray, N. Shakhovska, O. Mrykhina, M. Bublyk, and L. Lisovska, "Methodical Approach to Assessing the Readiness Level of Technologies for the Transfer," Advances in Intelligent Systems and Computing IV, 1080, Springer, 2020, pp. 259-282.

[43] N. Antonyuk, M. Medykovskyy, L. Chyrun, M. Dverii, O. Oborska, M. Krylyshyn, A. Vysotsky, N. Tsiura, and O. Naum, "Online Tourism System Development for Searching and Planning Trips with User's Requirements," Advances in Intelligent Systems and Computing IV, Springer Nature Switzerland AG, 1080, 2020, pp. 831-863.

[44] O. Lozynska, V. Savchuk, and V. Pasichnyk, "Individual Sign Translator Component of Tourist Information System," Advances in Intelligent Systems and Computing, Springer, 1080, 2020, pp. 593-601.

[45] A. Rzheuskyi, O. Kutyuk, O. Voloshyn, A. Kowalska-Styczen, V. Voloshyn, L. Chyrun, S. Chyrun, D. Peleshko, and T. Rak, "The Intellectual System Development of Distant Competencies Analyzing for IT Recruitment," Advances in Intelligent Systems and Computing IV, Springer, Cham, 1080, 2020, pp. 696-720.

[46] B. Rusyn, L. Pohreliuk, A. Rzheuskyi, R. Kubik, Y. Ryshkovets L. Chyrun, S. Chyrun, A. Vysotskyi, and V.B. Fernandes, "The Mobile Application Development Based on Online Music Library for Socializing in the World of Bard Songs and Scouts' Bonfires," Advances in Intelligent Systems and Computing IV, Springer, 1080, 2020, pp. 734-756.

[47] A. Bomba, M. Nazaruk, N. Kunanets, and V. Pasichnyk, "Modeling the Dynamics of Knowledge Potential of Agents in the Educational Social and Communication Environment," Advances in Intelligent Systems and Computing IV, Springer, Cham, 1080, 2020, pp. 17-24.

[48] R. Holoshchuk, V. Pasichnyk, N. Kunanets, and N. Veretennikova, "Information Modeling of Dual Education in the Field of IT," Advances in Intelligent Systems and Computing, Springer, 2020, pp. 637-646.

[49] V. Andrunyk, V. Pasichnyk, N. Antonyuk, ans T. Shestakevych, "A Complex System for Teaching Students with Autism: The Concept of Analysis. Formation of IT Teaching Complex," Advances in Intelligent Systems and Computing IV, Springer, 1080, 2020, pp. 721-733.

[50] T. Borovska, D. Grishin, I. Kolesnik, V. Severilov, I. Stanislavsky, and T. Shestakevych, "Research and Development of Models and Program for Optimal Product Line Control," Advances in Intelligent Systems and Computing IV, Springer, Cham, 1080, 2020, pp. 186-201.

[51] I. Oksanych, I. Shevchenko, I.,Shcherbak, S. Shcherbak, "Development of specialized services for predicting the business activity indicators based on micro-service architecture," Eastern-European Journal of Enterprise Technologies, 2(2-86), 2017, pp. 50-55.

[52] I. Galushka, and S. Shcherbak, "Devising a mathematical model for pattern-based enterprise data integration," Eastern-European Journal of Enterprise Technologies, 2(9), 2015, pp. 59-64.

[53] S.S. Shcherbak, "Interoperability web application models based on microformats," CriMiCo, 21st International Crimean Conference: Microwave and Telecommunication Technology, 2011, pp. 57-58.

[54] K. Morozov, I. Sidenko, G. Kondratenko, and Y. Kondratenko, "Increasing Web-Design Effectiveness Based on Backendless Architecture," CEUR workshop proceedings, Vol-2604, 2020, 894-905.

[55] K. Ivanova, G. Kondratenko, I. Sidenko, and Y. Kondratenko, "Artificial Intelligence in Automated System for Web-Interfaces Visual Testing," CEUR workshop proceedings, Vol-2604, 2020, pp. 1019-1031.

[56] V. Andrunyk, A. Vasevych, L. Chyrun, N. Chernovol, N. Antonyuk, A. Gozhyj, V. Gozhyj, I. Kalinina, M. Korobchynskyi, "Development of Information System for Aggregation and Ranking of News Taking into Account the User Needs," CEUR workshop proceedings, Vol-2604, 2020, pp. 1127-1171.

[57] V. Andrunyk, A. Berko, B. Rusyn, L. Pohreliuk, S. Chyrun, B. Dokhniak, I. Karpov, and M. Krylyshyn, "Information System of Photostock Web Galleries Based on Machine Learning Technology," CEUR workshop proceedings, Vol-2604, 2020, pp. 1032-1059.

IEEE Third International Conference on Data Stream Mining & Processing
August 21-25, 2020, Lviv, Ukraine

Design of a System for Dynamic Integration of Weakly Structured Data Based on Mash-Up Technology

Irina Pelekh, Andrii Berko, Vasyl Andrunyk
Information Systems and Networks Department
Lviv Polytechnic National University
Lviv, Ukraine
Andrii.Y.Berko@lpnu.ua

Lyubomyr Chyrun, Ivan Dyyak
Programming Department
Ivan Franko National University of Lviv
Lviv, Ukraine
Lyubomyr.Chyrun@lnu.edu.ua

***Abstract*— The article describes the design of a system for dynamic integration of weakly structured data using Mash-Up technology. Functional requirements to the system of dynamic data integration based on Mash-Up technology are characterized. An algorithm for constructing an ontological model of all integral systems and an algorithm for obtaining information resources from an integral system is proposed. The architecture and principles of the Shares and Discounts Mashup dynamic integration of weakly structured data are considered.**

Keywords*— *dynamic integration; Mash-Up technology; Mash-Up system design; mash-Up system architecture

I. INTRODUCTION

Due to the rapid growth of Ukrainian and world information resources, methods of dynamic integration of information based on Mash-Up technology and processing of information resources are in the focus of researchers of federated environments and are of scientific and practical interest to developers of modern distributed intelligent information systems, but numerous problems still remain unresolved [1-5]]. Unfortunately, there are no well-defined recommendations for the construction of dynamic data integration systems based on Mash-Up technology. Such projects are implemented mainly thanks to their own ideas and solutions.

II. ANALYSIS OF RECENT RESEARCH AND PUBLICATIONS

Content integration in Mash-Up systems is usually dynamic, that is, it occurs at runtime, based on user input. To achieve the minimum execution time, most Mash-Up systems support only simple types of data integration [6-9]. For example, they often use standardized object identifiers (such as latitude / longitude of geographic positions, or unique product numbers such as EAN or ISBN) to easily link different sources or services [10-12]. Query access to data sources or search engines is usually based on keywords, tags or category names, but without extensive post-processing, to view results from different sources. Hence, it can concluded that the current use of Mash-Up technology requires improved support for dynamic data integration in heterogeneous data objects. As a rule, the scheme of data integration Mash-Up is as follows: the task of user-defined data transformation to the service level and presentation of the finished collage of information resources (Fig.1) [13-21].

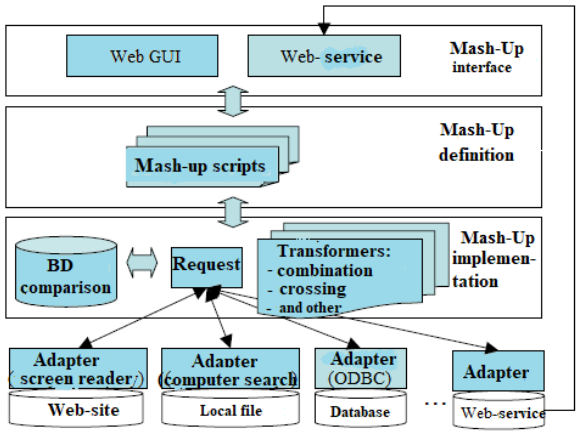


Fig. 1. Data integration Mash-Up scheme [2]

Let's consider some examples of the most famous modern Mash-Up systems. IFTTT [3] ("if this, then that" - "if this, then that") is a Mash-Up service that allows users to connect to various web applications (for example, Face book, Evernote, LinkedIn, Dropbox, etc.) using simple conditional statements known as "recipes" and create a simple automated sequence of operations that is performed when a certain action is performed. IFTTT allows users to create and share "recipes" that fit the judgment: "if this, then that", "this" is the part of the recipe that is called the trigger. It is quite easy to use and consists of only three tabs [22-27]:

- Tasks are a list of your active tasks.
- Recipes are a list of the most popular tasks that you can use as your own, that is, it is something like blanks tasks.
- Channels are a list of supported services, at the time of the study of 54 (For example, Twitter, Evernote, Google Calendar, LinkedIn, Google RSS Reader, Gmail, WordPress, etc.).

Google Alerts [4] is a service from the search giant. This system works based on the idea of monitoring the results of a search query according to the time changes. In fact, you can set Alerts to show new results on demand. The system is able to filter the results and select the most relevant data [28-36].

In the notification options list:

- The query itself (Google search query syntax is supported);
- Query type (all, news, blogs, videos, discussion, books);
- Message refresh rate (real-time, once a day, once a week);
- Filter the best results or all;
- Send results to e-mail or as an RSS feed.

Wappwolf [5] is a Mash - Up service for working with files. It is similar in idea with IFTTT, but with a bias towards file processing. The only event here is adding a file to the cloud storage folder (Dropbox, Google Drive, SkyDrive, Box are supported), but there can be quite a lot of actions:

- Synchronize with other cloud storage Box, SkyDrive, Google Drive, as well as with FTP server.
- A variety of simple operations for images: resize, grayscale, rotate, add watermark.
- Operations for audio files: convert to another format.
- Operations for text files: convert to PDF, eBook formats, kindle downloads, print via Google cloud printer.
- Operations for all file types: add to archive, rename, encrypt / decrypt.

Considering the principles of Mash-Up systems, they can divided into the following main groups:

- Formats and data access;
- Internal data model;
- Work with incoming and outgoing information flow (data mapping, providing functionality of data flow operators, data update, etc.);
- Functionality (extensibility, distribution, etc.).

III. Highlighting Problems and Statement of Purpose

The process of creating a dynamic data integration system based on Mash-Up technology requires a detailed analysis of all-important characteristics of the design and operation of such systems. It is as well preliminary development of recommendations for the design of this type of system and a detailed description of the processing of information resources in such a system, using the correct methodology. Thus, the task of developing new approaches, technologies, architectural solutions and flexible tools for dynamic integration of weakly structured data in the web environment is urgent. Therefore, the description of the design of the system of dynamic integration of weakly structured data using Mash - Up technology is promising and constructive in terms of solving the problems. The aim of the work is to develop recommendations for creating a system of dynamic data integration based on Mash-Up technology and semantic processing of information resources in such a system.

IV. Analysis of the Obtained Scientific Results

In [7] based on the analysis of activity of Mash-Up systems allocate the following States of activity:

1. Registration. At successful registration-transition to the second state, at unsuccessful registration-return again to the beginning of registration.
2. Authorization, if everything was successful, move on, if not go back to the beginning of authorization.
3. The formation of the problem. If the task is formed in accordance with the rules of the system, move on, if not – go back to the beginning of the third state.
4. The formation of the sources for the Mash-Up.
5. Search for the necessary information in the selected sources. If the search results are satisfactory-go ahead, if not go back to the search.
6. Extract information and move to the next state.
7. Storage of the received information in the form of a service.
8. Visual representation of the finished Mash-Up.

The most important States in the Mash-Up system, according to [7], is to search for the necessary information (the fifth state), extract the information found (the sixth state) and store it as a service (the seventh state). To improve and improve the result for the seventh state in [7] proposed a systematic procedure for determining the structure and content of input information resources, which can interpreted as a method for determining the structure and content of the input information. The use of this method makes it possible to improve the quality indicators of the result of the fifth state. In addition, since each next state depends on the previous one-it will increase the productivity of the next two important States of activity. Therefore, when designing a dynamic data integration system based on Mash-Up technology, the mentioned method is used. The relationship between functional and operational requirements for data integration systems and different areas of architecture such as application systems and technology architecture is shown in Fig. 2. Functional requirements for the application system describe the value that the system represents in terms of the implementation of the functions of the organization (business value). Operational (or operational) requirements for a software system specify aspects such as reliability, manageability, performance, security, compatibility. Moreover, this is not a complete list. Examples of operational requirements are the ability to access the system only authorized users, the level of availability of the application system 99, 99 % of the time, and the like. A technology architecture is a hardware and software infrastructure architecture that enables application systems to operate and fulfill the operational (non-functional) requirements that are presented to the application systems and information architecture. This leads to the following conclusion: "Functional requirements are provided by the application architecture, operational requirements are provided by the technological architecture."

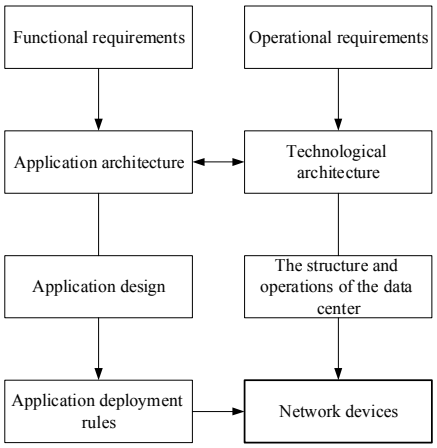


Fig. 2. Relationship of functional and operational requirements to application architecture and technology architecture

Although, of course, one remark should be made. A good technology architecture can provide security, availability, reliability, and a whole list of other operational requirements, but if the system is designed so that it does not take advantage of the technology architecture, it will still not function properly, and it will be difficult to implement and maintain. Similarly, a properly designed application system structure that accurately meets the requirements of business processes and is assembled from reusable components using state-of-the-art technology may not match the actual configuration of the hardware and system software used. For example, existing servers may not support system components, and network configuration and topology may not guarantee information flow needs. This shows that there is still a significant relationship between application architecture and technology architecture: a good technology architecture must built with support for application systems that play an important role in the organization's work. Therefore, the application architecture needs to make good use of the technology architecture to ensure that it meets all operational requirements. The functional requirements for the dynamic data integration system based on Mash-Up technology have developed, which are shown in Fig. 3. Using the method of determining the structure and content of the received input information [14] designed Shares and Discounts Mashup-search system discounts on the purchase of goods, services, and the like.

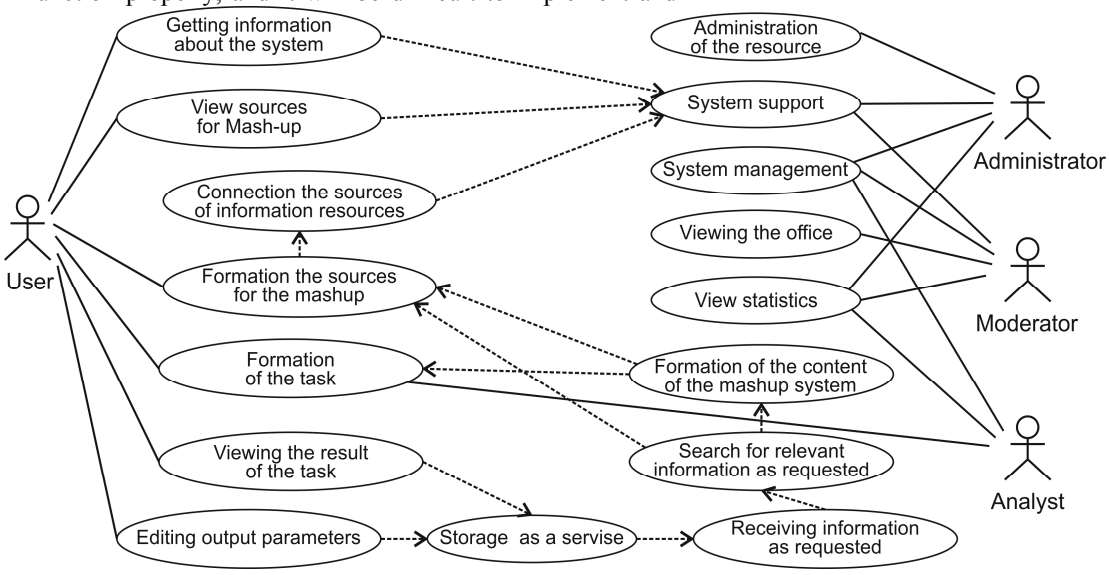


Fig. 3. Functional requirements for dynamic data integration system based on Mash-Up technology

The system automatically performs the display of information stored in heterogeneous information web-systems, which are integrated into the ontological model, which is later used to search for information resources, according to the user's request. The implementation of the system of dynamic integration of weakly structured data in web-systems is a set of completed modules that can used to build other systems. The functional decomposition of a program defines functions as abstract operations in terms of a task, not in the details of their implementation. According to the method of determining the structure and content of the received input information [7], an algorithm for constructing an ontological model of all systems that integrate was constructed (Fig. 4). When constructing any algorithm, the primary task is to determine the input and output data. The input data for the algorithm for obtaining the data structure of the system as an ontological model are block diagrams of databases of systems that are integrated; ontology of the subject area. The initial data is a General ontological model that describes the structure of integrated systems within their domain and the relationship between the elements of different systems. Such a model will modeled by means of the RDF language, its extension RDFs and using the OWL language. We describe the work of the algorithm for constructing an ontological model of all integrable systems. In addition, for this purpose we will introduce some notations of concepts. Let us have some database scheme of the system: $S = \{T_1,...,T_m\}$, where $T_1,...,T_m$ is system database schema tables S. $T_i = \{A_1,...,A_k\}$, $i = \overline{1,n}$, where $A_1,...,A_k$ is the attributes tables of the database schema. $R = \{R_1,...,R_z\}$ is the connection between the concepts of ontology. The algorithm for constructing an ontological model of all systems that are integrated contains 6 basic steps, namely: representation of database structure as RDF, i.e. sequential mapping of schema to RDF format.

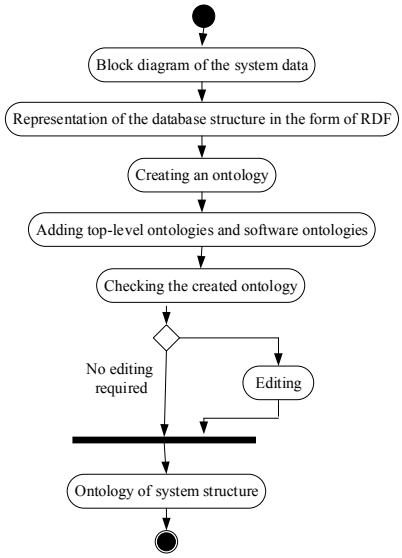


Fig. 4. Scheme of the algorithm for constructing an ontological model of all integrable systems

$$T_i \to T(RDF)_i,\ A_j \to A(RDF)_j,\quad i=\overline{1,n},\ j=\overline{1,k}, \qquad (1)$$

where $T(RDF)_i$ is the concepts of the ontology described by RDF; $A(RDF)_j$ is properties of concepts in ontology.

1. Adding semantic attributes and the creation of an ontology. This step is implemented by using a procedure to define common features of database elements and add links between them.
2. Adding top-level ontologies and domain ontologies. We implement this step using the OWL language, using the owl: import command. Thanks to the transitivity rule in RDF, additional ontologies expand subject areas and add new concepts and properties.
3. Checking the created ontology. This step is implemented by checking and analyzing the extracted ontology for "connectivity", that is, we check for missing semantic links anywhere. If so, then move on to step five, if not-move on to step six.
4. Edit the extracted ontology using the ontology editor (Protégé) and add links between concepts. Then go back to step 4.

Storing the resulting General ontology of the system structure in a file or metadata store in RDF format. The algorithm for obtaining information resources from the integrated system (Fig. 5) provides for the use of the previously obtained global ontological meta-model of integrable systems, including top-level ontologies and domain ontology. The input data of the algorithm are:

- Information about stored in databases resources;
- Global ontological goal-model.

The initial data, as a result of the algorithm, is a meta-model that combines the conceptual and substantive parts of the ontology.

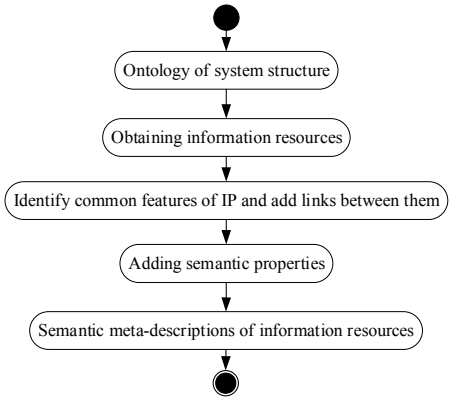


Fig. 5. The scheme of the algorithm for obtaining information resources from the system is integrated

Thus, such an ontology combines both information about the structure of integrated systems and metadata of objects stored in them, described in terms of the conceptual part. The obtained objective-model can used as a single interface for semantic integration of data of distributed systems and ensuring their interoperability. The work of the algorithm for obtaining information resources from the system, integrated consists of the following steps:

1. Obtaining information about stored information resources from a system that integrates using a global ontological meta-model. That is, we get each tuple from each table of a certain database schema.
2. Definition of the General features of the received tuples of tables of information resources and giving of semantic communications between them.
3. Adding semantic properties using semantic analyzer that works based on descriptor logic and imported ontology.

The result is a metadata ontology of information resources and the structure of the system stored in the metadata repository in RDF format. Information system architecture is a concept that defines the model, structure, functions and interrelation of information system components. Architecture of the designed system Shares and Discounts Mashup is shown in Fig. 6. The input data of the dynamic data integration system based on the technology Mash-Up Shares and Discounts Mashup are sources of information resources to which the system has access and the user request. The user sets the request and receives a response to it in the form of a collage of information resources available sources that correspond to the specified request. When designing the system interface, ergonomic characteristics, simplicity and compactness of information display on the screen, convenience of access to the main controls of the system operation modes, ensuring reliability of the system operation are taken into account. The interface is quite simple and understandable for any user.

Even without detailed instructions, everyone can understand what the Shares and Discounts Mashup system does. Everything is done in tones that are not debilitating to the human eye.

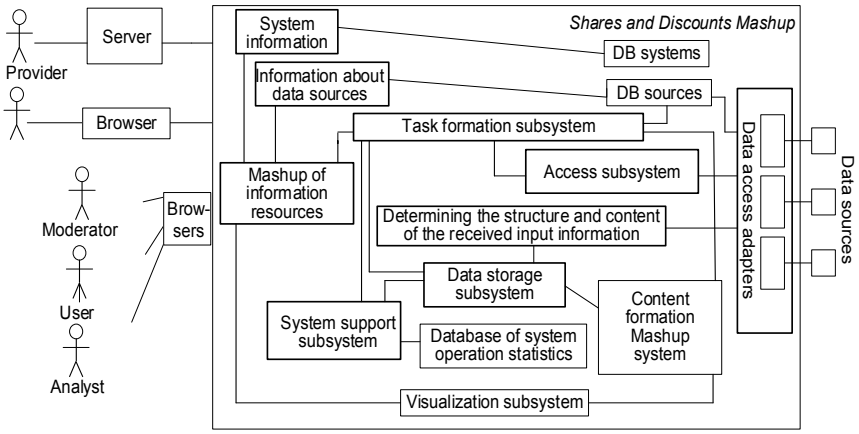


Fig. 6. Architecture of the system of dynamic integration of weakly structured data Shares and Discounts Mashup

The interface is written with elements of HTML syntax, because HTML is better suited for the design of web pages. The name of the system itself is unobtrusive, but still quite noticeable in the upper left corner of the page (Pic. 6). The inscription «Sale», which directly characterizes the search system for discount coupons, occupies the main place of the background. Navigation through the system is carried out by using menu items or simply flipping the page down. Since the system looks like a web site, it is necessary to provide at least some small description for the web site: its need, purpose and functions. To do this, a menu item with a name that does not require explanations "about us" is highlighted. Menu items are placed horizontally at the top of the site from left to right. Main menu items: "how does it work?", "Poshuk", "partners", "About us", "Contacts" (Fig. 7). By clicking on the menu item "How does it work?" You will go to the part of the site with a description of the rules of working with the system.

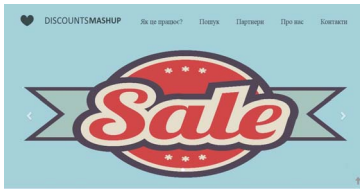


Fig. 7. Kind of system interface

The menu item "Search" corresponds to the search bar, where the user can enter a request to search for the desired discount coupon and the system will give a Mash-Up of results-information resources from available sources that most fully reflect the response to the user's request. The menu item "Partners" contains a list of possible sources for finding information resources with the ability to quickly go to the source site and read more about it. These sources are such well-known discount coupon sites: Pokupon, Biglion, Gara, Goodplace, Groupon, KupiSkidku, and the like. In General, the search is carried out on more than twenty different discount coupon sites. The "About us" menu item contains a description of the system. Menu item "Contacts" - feedback for user feedback, suggestions. Example of working with the system: the user enters "Spa treatments" into the search feed. Russian language is a condition for entering the request, since almost all Ukrainian discount coupon sites are described in Russian. The result of this query we can see in Fig. 8. Absolutely simple and clear scheme of operation of this system allows the user with any level of computer technology to search for a discount coupon to purchase the necessary goods or receive a certain service easily and quickly. Further, finding the desired discount, the user can directly go to the source of the discount coupon and read more about the promotion.

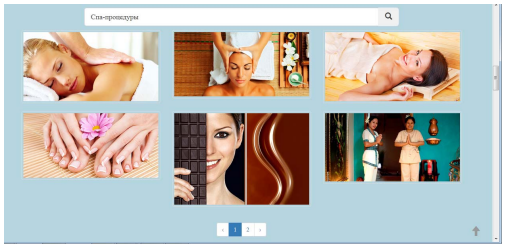


Fig. 8. The result of the system

V. CONCLUSIONS

To solve the problem of dynamic integration of weakly structured data, the design of the system of dynamic integration of weakly structured data Shares and Discounts Mashup is considered. The interrelations of functional and operational requirements with application architecture and technological architecture in dynamic data integration systems are shown. Functional requirements for dynamic data integration system based on Mash-Up technology are described. Described the process of processing information resources in the system of dynamic integration semi-structured data using the algorithm for constructing the ontological model of all systems that integrate and algorithm of information resources of system, integrations. The architecture of the system of dynamic integration of weakly structured data Shares and Discounts Mashup is described. The functioning of the designed system of dynamic integration of weakly structured data Shares and Discounts Mashup is considered. Further research will be devoted to the scientific search for expanding the scope of applications of the developed and proposed methods in the design of systems based on the technology Mash-Up dynamic data integration.

REFERENCES

[1] A.Y. Berko, and K.A. Aliekseyeva, "Quality evaluation of information resources in web-projects," Actual Problems of Economics, vol. 136(10), 2012, pp. 226-234.

[2] T. Fischer, F. Bakalov, and A. Nauerz, "An overview of current approaches to mashup generation," Proceedings of the International Workshop on Knowledge Services and Mashups, 2009, pp. 157-158.

[3] B. Rusyn, L. Pohreliuk, A. Rzheuskyi, R. Kubik, Y. Ryshkovets L. Chyrun, S. Chyrun, A. Vysotskyi, and V.B. Fernandes, "The Mobile Application Development Based on Online Music Library for Socializing in the World of Bard Songs and Scouts' Bonfires," Advances in Intelligent Systems and Computing IV, Springer, 1080, 2020, pp. 734-756.

[4] V. Andrunyk, L. Chyrun, and V. Vysotska, "Electronic content commerce system development," Proceedings of 13th International Conference: The Experience of Designing and Application of CAD Systems in Microelectronics, CADSM 2015-February, 2015.

[5] K. Alieksieieva, A. Berko, and V. Vysotska, "Technology of commercial web-resource processing," Proceedings of 13th International Conference: The Experience of Designing and Application of CAD Systems in Microelectronics, CADSM 2015-February, 2015.

[6] L. Giusy, and H. Hakim, "Mashups for data integration: An analysis," Technical Report UNSW-CSE-TR-0810, 2008, pp. 68-69.

[7] G. Canfora, A.R. Fasolino, G. Frattolillo, and P. Tramontana, "Migrating interactive legacy systems to web services," IEEE CS Press, editor, European Conference on Software Mainteinance and Reengineering, 2006, pp. 23-32.

[8] V. Vysotska, and L. Chyrun, "Methods of information resources processing in electronic content commerce systems," Proceedings of 13th International Conference: The Experience of Designing and Application of CAD Systems in Microelectronics, CADSM, 2015.

[9] V. Vysotska, and L. Chyrun, "Analysis features of information resources processing," Proceedings of the International Conference on Computer Sciences and Information Technologies, CSIT, 2015, pp. 124-128.

[10] V. Vysotska, R. Hasko, and V. Kuchkovskiy, "Process analysis in electronic content commerce system," Computer Sciences and Information Technologies, CSIT, 2015, pp. 120-123.

[11] V. Lytvyn, V. Vysotska, O. Veres, I. Rishnyak, and H. Rishnyak, "The Risk Management Modelling in Multi Project Environment," Proceedings of the International Conference on Computer Sciences and Information Technologies, CSIT, 2017, pp. 32-35.

[12] V. Vysotska, V. Lytvyn, Y. Burov, A. Gozhyj, and S. Makara, "The consolidated information web-resource about pharmacy networks in city," CEUR Workshop Proceedings, 2018, pp. 239-255.

[13] V. Vysotska, V.B. Fernandes, and M. Emmerich, "Web content support method in electronic business systems," CEUR Workshop Proceedings, Vol-2136, 2018, pp. 20-41.

[14] A. Gozhyj, I. Kalinina, V. Vysotska, and V. Gozhyj, "The method of web-resources management under conditions of uncertainty based on fuzzy logic," Proceedings of the International Conference on Computer Sciences and Information Technologies, CSIT, 2018, pp. 343-346.

[15] A. Gozhyj, V. Vysotska, I. Yevseyeva, I. Kalinina, and V. Gozhyj, "Web Resources Management Method Based on Intelligent Technologies," Advances in Intelligent Systems and Computing, 871, 2019, pp. 206-221.

[16] 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, 689, Springer, 2018, pp. 204-216.

[17] T. Batiuk, V. Vysotska, and V. Lytvyn, "Intelligent System for Socialization by Personal Interests on the Basis of SEO-Technologies and Methods of Machine Learning,"CEUR workshop proceedings, Vol-2604, 2020, pp. 1237-1250.

[18] V. Lytvyn, A. Gozhyj, I. Kalinina, V. Vysotska, V. Shatskykh, L. Chyrun, Y. Borzov, "An intelligent system of the content relevance at the example of films according to user needs," CEUR Workshop Proceedings, Vol-2516, 2019, pp. 1-23.

[19] P. Kravets, V. Lytvyn, V. Vysotska, and Y. Burov, "Promoting training of multi-agent systems," CEUR Workshop Proceedings, Vol-2608, 2020, pp. 364-378.

[20] B. Rusyn, V. Vysotska, and L. Pohreliuk, "Model and architecture for virtual library information system," Proceedings of the International Conference on Computer Sciences and Information Technologies, CSIT, 2018, pp. 37-41.

[21] V. Lytvyn, and V. Vysotska, "Designing architecture of electronic content commerce system," Proceedings of the Internat. Conference on Computer Sciences and Information Technologies, 2015, pp. 115-119.

[22] O. Naum, L. Chyrun, O. Kanishcheva, and V. Vysotska, "Intellectual System Design for Content Formation," Proceedings of the International Conference on Computer Sciences and Information Technologies, CSIT, 2017, pp. 131-138.

[23] J. Su, A. Sachenko, V. Lytvyn, V. Vysotska, and D. Dosyn, "Model of Touristic Information Resources Integration According to User Needs," Proceedings of the International Conference on Computer Sciences and Information Technologies, CSIT, 2018, 113-116.

[24] V. Lytvyn, V. Vysotska, D. Dosyn, and Y. Burov, "Method for ontology content and structure optimization, provided by a weighted conceptual graph," Webology, vol. 15(2), 2018, pp. 66-85.

[25] V. Lytvyn, V. Vysotska, D. Dosyn, O. Lozynska, and O. Oborska, "Methods of Building Intelligent Decision Support Systems Based on Adaptive Ontology," IEEE 2nd International Conference on Data Stream Mining and Processing, DSMP, 2018, pp. 145-150.

[26] V. Lytvyn, Y. Burov, P. Kravets, V. Vysotska, A. Demchuk, A. Berko, Y. Ryshkovets, S. Shcherbak, and O. Naum, "Methods and Models of Intellectual Processing of Texts for Building Ontologies of Software for Medical Terms Identification in Content Classification," CEUR Workshop Proceedings, Vol-2362, 2019, pp. 354-368.

[27] S. Babichev, "An Evaluation of the Information Technology of Gene Expression Profiles Processing Stability for Different Levels of Noise Components," Data, vol. 3 (4), 2018, art. no. 48.

[28] S. Babichev, B. Durnyak, I. Pikh, and V. Senkivskyy, "An Evaluation of the Objective Clustering Inductive Technology Effectiveness Implemented Using Density-Based and Agglomerative Hierarchical Clustering Algorithms," Advances in Intelligent Systems and Computing, vol. 1020, 2020, pp. 532-553.

[29] N. Antonyuk, A. Vysotsky, V. Vysotska, V. Lytvyn, Y. Burov, A. Demchuk, I. Lyudkevych, L. Chyrun, S. Chyrun, and I. Bobyk, "Consolidated Information Web Resource for Online Tourism Based on Data Integration and Geolocation," Conference on Computer Sciences and Information Technologies, CSIT, 2019, pp. 15-20.

[30] A. Vysotsky, V. Lytvyn, V. Vysotska, D. Dosyn, I. Lyudkevych, N. Antonyuk, O. Naum, A. Vysotskyi, L. Chyrun, and O. Slyusarchuk, "Online Tourism System for Proposals Formation to User Based on Data Integration from Various Sources," Computer Sciences and Information Technologies, CSIT, 2019, pp. 92-97.

[31] O. Lozynska, V. Savchuk, and V. Pasichnyk, "Individual Sign Translator Component of Tourist Information System," Advances in Intelligent Systems and Computing, Springer, 1080, 2020, pp. 593-601.

[32] T. Borovska, D. Grishin, I. Kolesnik, V. Severilov, I. Stanislavsky, and T. Shestakevych, "Research and Development of Models and Program for Optimal Product Line Control," Advances in Intelligent Systems and Computing IV, Springer, Cham, 1080, 2020, pp. 186-201.

[33] I. Oksanych, I. Shevchenko, I.,Shcherbak, S. Shcherbak, "Development of specialized services for predicting the business activity indicators based on micro-service architecture," Eastern-European Journal of Enterprise Technologies, 2(2-86), 2017, pp. 50-55.

[34] I. Galushka, and S. Shcherbak, "Devising a mathematical model for pattern-based enterprise data integration," Eastern-European Journal of Enterprise Technologies, 2(9), 2015, pp. 59-64.

[35] K. Morozov, I. Sidenko, G. Kondratenko, and Y. Kondratenko, "Increasing Web-Design Effectiveness Based on Backendless Architecture," CEUR workshop proceedings, Vol-2604, 2020, 894-905.

[36] K. Ivanova, G. Kondratenko, I. Sidenko, and Y. Kondratenko, "Artificial Intelligence in Automated System for Web-Interfaces Visual Testing," CEUR workshop proceedings, Vol-2604, 2020, pp. 1019-1031.

IEEE Third International Conference on Data Stream Mining & Processing
August 21-25, 2020, Lviv, Ukraine

Adaptive Algorithm for Radar-System Parameters Tuning by means of Motion Zone Estimation

Andrii Cheredachuk
National University of Kyiv-Mohyla Academy
Kyiv, Ukraine
andry.cheredarchuk@gmail.com

Galyna Kriukova
National University of Kyiv-Mohyla Academy
Kyiv, Ukraine
kriukovagv@ukma.edu.ua

Andrii Malenko
Glushkov Institute of Cybernetics NAS of Ukraine
Kyiv, Ukraine
malenko.andrii@gmail.com

Maksym Sarana
National University of Kyiv-Mohyla Academy
Kyiv, Ukraine
m.sarana@ukma.edu.ua

Oleksandr Sudakov
National Taras Shevchenko University of Kyiv
Kyiv, Ukraine
saa@univ.kiev.ua

Sergii Vodopyan
National University of Kyiv-Mohyla Academy
Kyiv, Ukraine
ORCID: 0000-0002-8406-7434

Yevhenii Volynets
National University of Kyiv-Mohyla Academy
Kyiv, Ukraine
volynets@ukma.edu.ua

***Abstract*—This paper focuses on development of algorithm for parameters tuning for Doppler radar-based sensing systems for motion monitoring. The algorithm adapts parameters for better system performance according to estimated motion zone, i.e. mean and maximum distance to objects and their velocities. A new approach for determination of radar height position and tilt angle is proposed and experimentally tested.**

The method uses radar measurement points cloud only and may be applied for optimization of distance and velocity ranges and reduction of artifacts. Error of tilt angle estimation is 0.7-1.6 degrees in range of 0-45 degrees. Height estimation error depends on moving object height and is less than 1 m. Main goal of the algorithm is estimation of motion zone and corresponding adaptive change of radar-system configuration parameters for optimising system's field of view and performance. The proposed approach may be used for indoor and outdoor motion detection and activity classification in applications such as health diagnostics, health monitoring, surveillance, occupancy sensing, and automotive.

***Index Terms*—adaptive algorithm, parameters tuning, radio frequency sensor, noise reduction**

I. Introduction

Motion detectors have found wide use in commercial applications, e.g. surveillance video cameras with infrared night vision, passive infrared sensors (PIR), thermal cameras, radar sensors, and many more data stream mining systems [1]–[4]. Recently, the radar-based sensors have become widely employed in robotics and active safety applications, as low power and cost effective devices for motion detection, able to track and classify a number of objects simultaneously and estimate corresponding distances and shapes. Various fusion techniques for sensor data, such as radar-detections and video camera data allows to develop new algorithms for object classification and recognition [5], [6]. Such an approach may be used for the development of hybrid intelligent systems [7].

Galyna Kriukova is grateful to the Charity Foundation "Believe in Yourself" for financing her sabbatical.

The Frequency-Modulated Continuous-Wave (FMCW) technique providing high resolution measurements is widely used in robotics and automotive industry, both for instrumentation and measurement. The main idea of radars with FMCW technique is to get the range and velocity information from the beat frequency, which is composed of Doppler frequency and propagation delay.

Most radio frequency sensors use the 24, 60, and 77GHz radio bands [8]. Spectrum regulations and standards developed by the European Telecommunications Standards Institute (ETSI) and U.S. Federal Communications Commission (FCC) restrict new products from using the 24GHz ultra-wide band and 77GHz band for city infrastructure, industrial factory and building applications, including those requiring human-machine interaction (see Electronic Communications Committee Decision (04)10).

Sensors using 60GHz band are capable of gathering rich high accurate point-cloud data, making this band preferable for radar-sensing applications in worldwide industrial setting [8]. Mentioned previously point-cloud radar data (in the 3-dimensional space x, y, z-axes, and radial velocity data) enables object and motion detection, localisation and classification in both indoor and outdoor applications. Nevertheless, in the setting of multi-target detection problem, common FMCW radars will suffer from the range-velocity ambiguity problem, which typically causes missed or ghost targets [3].

In order to automatically improve radar-based system, the algorithm for system performance estimation should be developed. Such an idea is widely used for video-systems, for example, a number of objective blind image quality assessment methods for better object detection and classification were proposed [9]. To the best of our knowledge, for radar-based systems such an approach is not very well studied.

Gathering high accuracy meaningful data requires corresponding maximum and resolution for both range and velocity from the sensor. Increasing of radar distance range requires

decreasing of velocity range and vice-versa. In order to optimise contrary parameters by tuning chirp configuration, it is beneficial to estimate and match actual field of view and motion zone of sensor, therefore problem of position restoring arises: find height and tilt angle of radar mounting from given point-cloud radar data.

Main idea of our approach is the following: given default chirp configuration, the online algorithm estimates the motion zone and field of view of the radar-based system, afterwards, with estimations of maximum range and velocity observed, the algorithm adapts the chirp configuration parameters to optimise system's field of view and performance.

In this paper, we consider a new algorithm for system parameters tuning based on 3D reconstruction method of motion zone estimation for radar-based system with corresponding tuning and optimisation of system parameters for better performance. We integrate domain knowledge to data science problem of parameters optimisation, at the same time incorporating theoretical models to analyse data, i.e. applying theory-guided data science approach discussed in [10].

Machine learning methods are widely used is signal processing, e.g. for noise reduction or anomaly detection, in particular for UWB radar signal data [3], [11]–[13].

The organisation of the paper is following: background information and recent developments, followed by the description of hardware and software system architecture, overview how chirp-configuration depends on system performance according to theory of Doppler radar sensing, then we propose method for noise reduction, motion detection and classification, provide experimental results and discuss potential applications and related challenges.

II. System and Method Description

The radar-based system can be divided into two hardware and software subsystems. The hardware subsystem includes the RF subsystem (transmitter, receiver and antenna array) and Digital Signal Processing (DSP) subsystem. The software subsystem consists of control and algorithm parts.

A. Hardware subsystem

The hardware configuration of the proposed radar system includes the RF/analog subsystem and DSP subsystem. We based our tests on Texas Instruments IWR6843, commercial FMCW radar device operating in the 60–64GHz band [14]. Functional diagram of the sensor is presented in Fig. 1. The antenna array is controlled by DSP subsystem with 3 transmitting and 4 receiving antenna elements.

B. Software subsystem

In this section, we explain details of our adaptive method. The software subsystem has hardware control parts and algorithmic part. It is based on firmware and millimeter wave software development kit (mmWave SDK) provided by Texas Instruments [15]. It allows user to specify the chirping profile, displays the detected objects and other information in real-time.

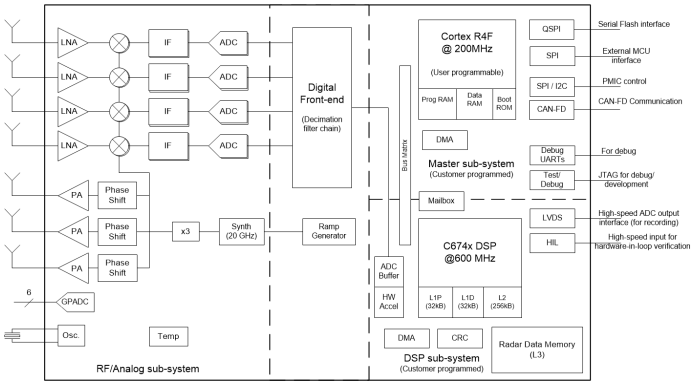


Fig. 1. IWR6843 Functional Block Diagram [14].

C. Chirp configuration

In FMCW radars with linear slope, the transmitting signal is a sweep in frequency, which is usually called as a "chirp", i.e. a single tone with its frequency changing linearly with time. A set of such chirps mold a "frame", which is used as the observation window for the radar processing. Various parameters of the chirp ramp (i.e. sweep bandwidth, frequency slope, time of chirping, sampling frequency, etc) impact the system performance [16].

For example, the range of distance over which a radar-based sensor detects objects is a crucial parameter for the following applications of the system. Detection of objects at large distance is limited by both the signal-to-noise ratio of the received signal, and the intermediate frequency (IF) bandwidth supported by the radar device. The relationship between maximum detection range and the intermediate frequency (IF) bandwidth is shown in (1):

$$R_{max} = \frac{IF_{max} \cdot c}{2S}, \tag{1}$$

where IF_{max} is maximum intermediate frequency (IF) bandwidth supported, S is slope of the transmitted chirp, and c is speed of light.

At the same time, in majority of applications resolving two closely spaced objects as two separate objects, rather than detect them as one, is a crucial ability for the system. Therefore, another important metric is the smallest distance between two objects that allows them to be detected as separate, which is usually referred as the range resolution. This parameter depends on the bandwidth of chirp sweep that the radar sensor can provide. The relationship is the following (2).

$$R_{res} = \frac{c}{2B}, \tag{2}$$

where c is speed of light, and B is sweep bandwidth of FMCW chirp. Correspondingly, for better range resolution the larger sweep bandwidth should be used. For example, if radar device support a 4GHz sweep bandwidth, that allows a range resolution of as low as approximately 4cm.

As sweep bandwidth B of FMCW chirp is proportional to its slope S. Combining (1) and (2) we have that range resolution and maximum of detection range are proportional

$$R_{res} \propto R_{max}.$$

Therefore, to optimise range resolution, it's better to minimise R_{max} to actual motion zone.

The same relationship between system parameters is related to velocity of the object, which along with distance detection is another critical parameter. The chirp cycle time (that is, the time difference between the starts of two consecutive chirps) influences the maximum measurable velocity by FMCW radar-based system. This correspondingly depends on the minimum inter-chirp time allowed and how fast the frequency sweep can be performed, therefore, for maximum of velocity detection we have:

$$V_{max} = \frac{\lambda}{4T_c}, \tag{3}$$

where T_c is total chirp time, including chirp time and idle time, and λ is wavelength of the signal used.

Along with range resolution, separating out objects with small velocity differences is an important ability of radar-based system, therefore, good velocity resolution is required. Corresponding velocity resolution parameter depends on the duration of transmit frame, i.e. increasing the number of chirps in a frame improves the velocity resolution.

$$V_{res} = \frac{\lambda}{2NT_c}, \tag{4}$$

where T_c is total chirp time, N is number of chirps in a frame, λ is wavelength of the signal used.

Once again, combining (3) and (4) we get

$$V_{res} \propto V_{max}.$$

Doppler radars are generally ambiguous in either range or doppler, or both. In the case when actual object velocity is more than unambiguous V_{max} various distortion and ghost targets may be observed. In Fig. 2, 3 and 4 we compare point-clouds of detected object with sensor with different chirp configs.

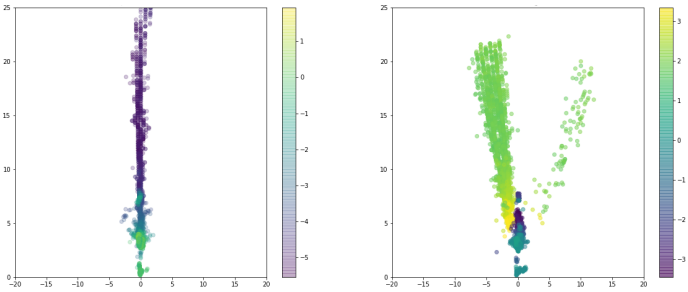


Fig. 2. Comparison of detected points of approaching object with speed approx. 4.2 m/s, for chirp with $V_{max} = 15.3$ m/s (left) and $V_{max} = 3.4$ m/s (right). Points colour corresponds to detected radial velocity in m/s.

It's easy to see that sensor performs much better when V_{max} exceeds object speed. Moreover, otherwise inference in

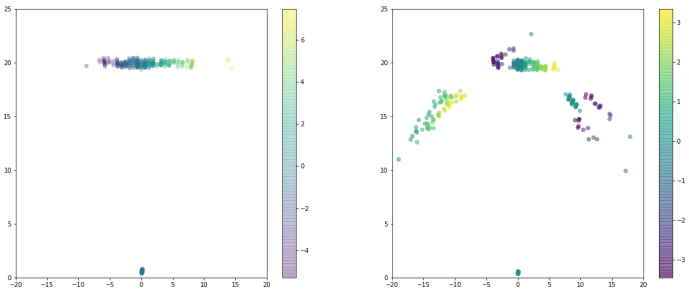


Fig. 3. Comparison of detected points of object moving tangentially with speed approx. 13.9 m/s, for chirp with $V_{max} = 15.3$ m/s (left) and $V_{max} = 3.4$ m/s (right). Points colour corresponds to detected radial velocity in m/s.

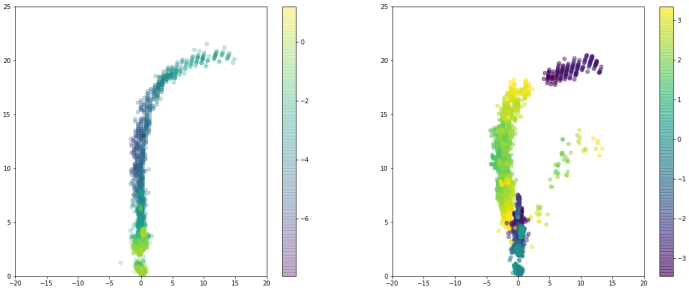


Fig. 4. Comparison of detected points of object with altering direction and speed, for chirp with $V_{max} = 15.3$ m/s (left) and $V_{max} = 3.4$ m/s (right). Points colour corresponds to detected radial velocity in m/s.

target localisation, ghost targets and other distortions occur. Thereby, as other configuration parameters are interconnected and influence system parameters, adaptive chirp configuration tuning is of interest for better performance in the real-world applications.

Main idea of the algorithm is following: given initial parameters we estimate motion zone and velocities distribution. Then we adapt system parameters corresponding to range maximum and velocity maximum to meet these requirements. Consistently, we adapt range resolution and velocity resolution parameters. Then the process is repeated iteratively.

D. Noise reduction

Due to the range-velocity ambiguity problem and various reflections or other noise sources, from multi-path propagation of a transmitted radar wave or due to interference from other radar sensors, in real-world applications radar-based sensor often produces ghost (imaginary) targets and other noise, which have nothing in common with real objects, like human or vehicles.

To eliminate noisy points from records, we apply the following idea: extended Kalman filter (EKF) based multi-hypothesis tracker is used to group points into trajectories. We've explored usage of Hidden Markov Model (HMM, see [17]) for tracking as well, but EKF with motion model embedded performs better.

As the tracker doesn't use information about estimated radial object velocity, we use this information to detect noisy

erroneous trajectories and remove corresponding points from data for further analysis.

Radial velocity of the object may be seen as a derivative of the distance to object with respect to time:

$$v = \frac{dR}{dt}.$$

Therefore, integrating it we should get

$$R = \int v\, dt,$$

or

$$\Delta R = \int_{t_1}^{t_2} v\, dt. \tag{5}$$

Therefore, having trajectory

$$\{(t_i, x_i, y_i, z_i, v_i) |\, i = 0, \ldots, n\}$$

we may calculate approximation of the integral (5)

$$\hat{r}_k = \int_{t_0}^{t_k} v\, dt \approx \sum_{i=1}^{k} (t_i - t_{i-1}) v_i,$$

and compare the value $\hat{r}_k$ with actual distance to object

$$r_k = \sqrt{x_k^2 + y_k^2 + z_k^2}.$$

For concerted trajectories values $\hat{r}_k$ and r_k have to agree with each other up to some constant, corresponding to object distance at initial time t_0. To verify this, it's enough just to consider linear regression for inputs $\{\hat{r}_k | k = 1, \ldots, n\}$ and outputs $\{r_k | k = 1, \ldots, n\}$

$$r = a\hat{r} + b,$$

where ideally a should be equal to 1 and b corresponds to distance to object at time t_0. Nevertheless, we may estimate the coefficient of determination R^2 of the prediction. The best possible score is 1, in our numerical experiment we've used different values of threshold to label trajectory as "good" or "noisy". At the same time, estimated value b, which ideally corresponds to object distance at initial time, has to fit to feasible distance, therefore, an inequality

$$0 \leq b \leq R_{max} + \varepsilon$$

must hold for some acceptable level of noise ε, for example

$$0 \leq b \leq R_{max} + R_{res}.$$

This approach allows us to select trajectories corresponding to salient motion, remove noisy and recurrent movement (e.g. fan vibration or plant oscillation). Thereby the method enables tuning chirp parameters according to estimated motion zone.

III. Experiment Results

Our experiments are conducted with a commercial FMCW 60–64GHz band radar chip IWR6843 made by Texas Instruments [14]. The experiment is performed in a natural environment for various system mounting. Initial system parameters are the following according to selected chirp configuration: range resolution 0.28m, velocity resolution 0.25m/s, maximum velocity 15m/s, maximum range 25m, angle resolution is up to 5° on the sides of sensor's field of view.

In order to verify motion zone reconstruction algorithm performance, radar sensor is mounted at the height of 1.5m, 2m and 2.5m and tilt angle 0°, 15°, 30° and 45°. We record identical sandglass-shaped trajectories of person moving in front of radar sensor, and then we use our method to reconstruct shape or corresponding point clouds.

To measure the performance of the algorithm, we estimate the plane underlying the point cloud, compare its position (height and tilt angle) with corresponding mounting height and tilt angle. Getting the corresponding plane may be reduced to the total least squares problem [18], [19] and solved by means of singular value decomposition (SVD) method, i.e. for set of points we calculate centroid (or centre of mass), subtract it from each point, compute SVD decomposition. Eigenvector corresponding to the least eigenvalue defines plane as a normal to the plane vector.

In Fig. 5, 6, 7, 8 the noise removal and motion zone reconstruction process is illustrated for tilt angles 0°, 15°, 30° and 45°, and for height 1.5m, 2.0m and 2.5m accordingly.

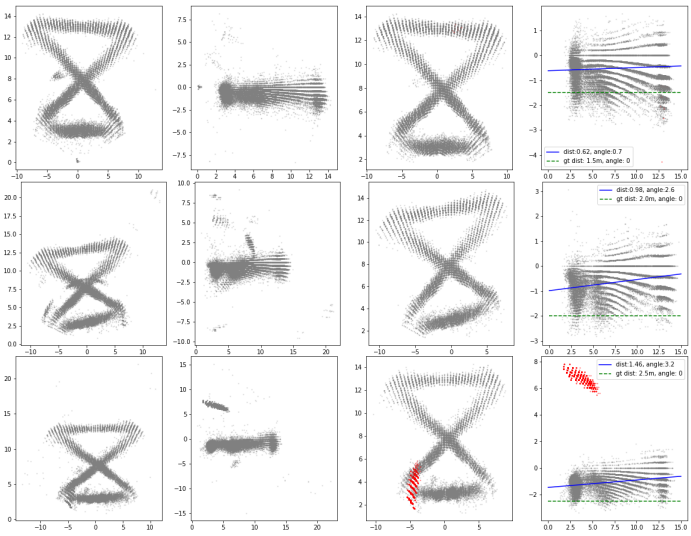


Fig. 5. Motion zone reconstruction for tilt angle 0°.

Two left-most plots depict point-cloud (view from above, i.e. (x, y)-plane, and side view, i.e. (y, z)-plane, correspondingly).

Two right-most plots correspond to the de-noised point-cloud by means of tracker in the same projections. Points corresponding to static trajectories (recurrent motion) are pictured with red colour. In (y, z)-plane green dashed line corresponds

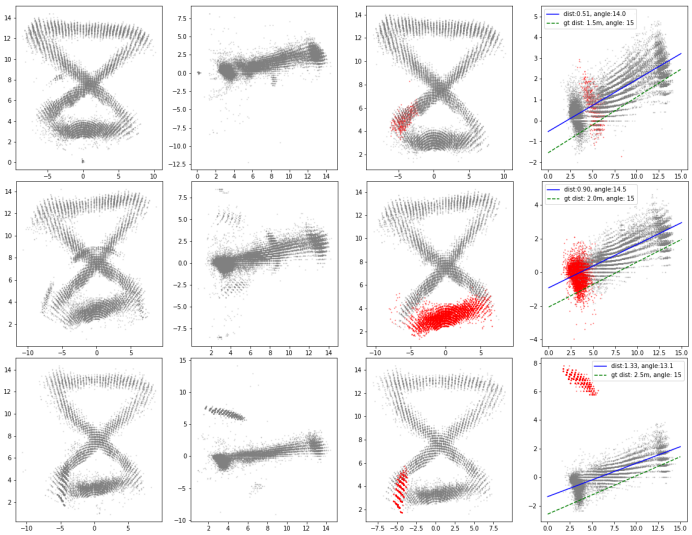


Fig. 6. Motion zone reconstruction for tilt angle 15°.

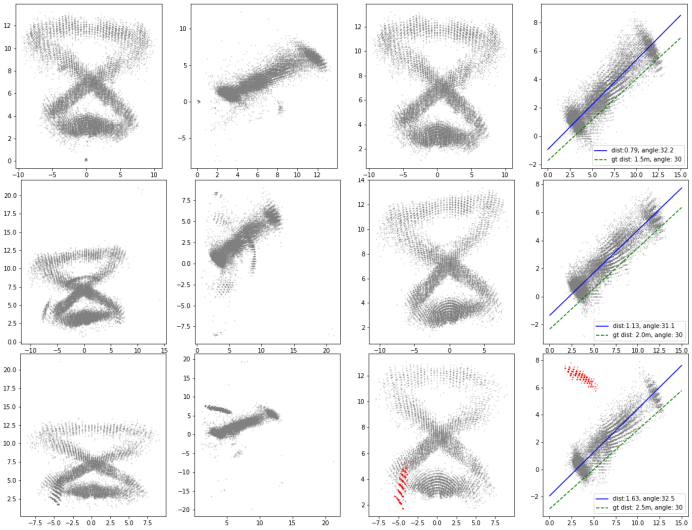


Fig. 7. Motion zone reconstruction for tilt angle 30°.

to ground truth of footing level, blue one corresponds to estimated plane of motion zone (approximately at the height 1.0m above the ground).

Results of motions zone reconstruction are presented in Table I. As we can see, the error of angle recovery is less than 10°, approximately 2°, which corresponds to measurement error and/or IWR6843 angle resolution.

Difference between estimated middle plane of point-cloud and ground level definitely should be positive. As we can see, it approximately equals 1m, which corresponds to estimated mass centre of the human body. Evidently, for each mounting height and tilt angle we may observe different points of the human body, and corresponding middle plane of point-cloud may shift.

Therefore, we assume, that the angle error is more appropriate measure for estimating performance of motion zone

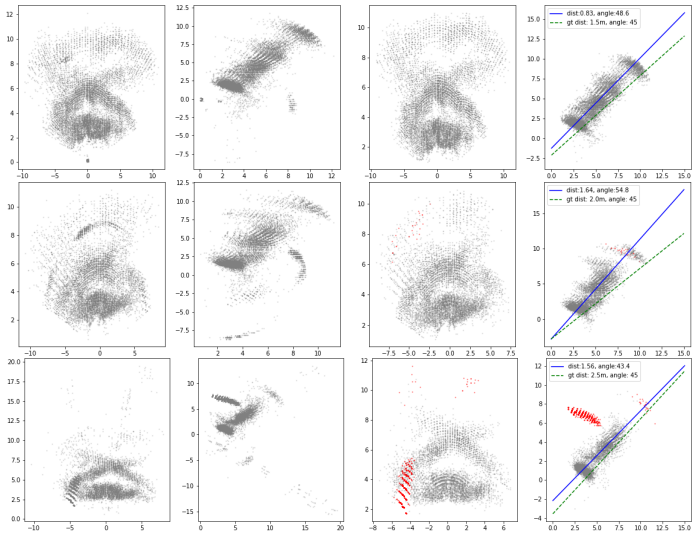


Fig. 8. Motion zone reconstruction for tilt angle 45°.

reconstruction.

TABLE I
MOTION ZONE RECONSTRUCTION

	Ground Truth		Estimated		Difference	
Figure	***angle***	***height***	***angle***	***height***	***angle***	***height***
Fig. 5	0°	1.5m	0.7°	0.62	0.7°	0.88
Fig. 5	0°	2.0m	2.6°	0.98	2.6°	1.02
Fig. 5	0°	2.5m	3.2°	1.46	3.2°	1.04
Fig. 6	15°	1.5m	14.0°	0.51	1.0°	0.99
Fig. 6	15°	2.0m	14.5°	0.90	0.5°	1.1
Fig. 6	15°	2.5m	13.1°	1.33	1.9°	1.17
Fig. 7	30°	1.5m	32.2°	0.79	2.2°	0.71
Fig. 7	30°	2.0m	31.1°	1.13	1.1°	0.87
Fig. 7	30°	2.5m	32.5°	1.63	2.5°	0.87
Fig. 8	45°	1.5m	48.6°	0.83	3.6°	0.67
Fig. 8	45°	2.0m	54.8°	1.64	9.8°	0.36
Fig. 8	45°	2.5m	43.4°	1.56	1.6°	0.94

According to estimated motion zone, the maximum range doesn't need to exceed 15m, therefore value of 25m is excessive, accordingly, chirp configuration may be modified to decrease maximum detection range and improve range resolution up to 0.17cm. Moreover, having long history of records, we may get estimation of maximum velocity observed, and modify corresponding parameters accordingly.

IV. CONCLUSIONS

In the paper we consider a new method for de-noising and static clutter removal for motion zone reconstruction. This approach may be used for adaptive change of chirp configuration parameters for optimising radar system's field of view and performance.

The algorithm works in the following loop: having some initial radar system configuration (such as chirp slope, time of chirping, number of chirps per frame, etc), the motion detection data are collected, after noise reduction the motion zone boundaries are estimated (such as velocity and range maximum for detected objects), finally, the radar-system parameters

are updated to meet the requirements of the corresponding motion zone, optimise system's field of view and performance.

Such an approach allows to tune system's parameters without any additional manual work and maintenance, whereas preserving data security and privacy. The presented method is a step towards a better understanding of noise reduction and system optimization according to the environment and usage history.

We hope that such an approach may be of interest and can make a significant contribution to development of novel indoor and outdoor safety systems. Moreover, we assume there is a good potential that combining the system configuration tuning and sensor fusion techniques will enable a powerful tool for various object detection and tracking frameworks.

In order to broaden the approach, the sensor fusion (i.e. video-camera data and radar-system motion detection and localization) may be used to optimize the motion zone estimation and boost the sensor algorithms for better system's accuracy and performance. We plan to consider hybrid machine learning algorithms for sensor fusion as a part of our future research.

ACKNOWLEDGMENT

Authors would like to express deep and sincere gratitude to Mr. Sergii Bodnarchuk for providing invaluable guidance and support throughout this research.

REFERENCES

[1] S. N. A. Wahab, S. Ramli, and N. M. Zainudin, "Temperature determining method from motion detection using thermal images," in *2015 IEEE 11th International Colloquium on Signal Processing Its Applications (CSPA)*, March 2015, pp. 26–29.

[2] O. Sudakov and A. Malenko, "Realistic mathematical model of passive infrared sensor's signal," in *2019 10th IEEE International Conference on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications (IDAACS)*, vol. 2, Sep. 2019, pp. 757–760.

[3] I. Ryu, I. Won, and J. Kwon, "Detecting ghost targets using multilayer perceptron in multiple-target tracking," *Symmetry*, vol. 10, no. 1, p. 16, 2018. [Online]. Available: https://www.mdpi.com/2073-8994/10/1/16

[4] Y. Bodyanskiy, O. Vynokurova, Z. Szymański, I. Kobylin, and O. Kobylin, "Adaptive robust models for identification of nonstationary systems in data stream mining tasks," in *2016 IEEE First International Conference on Data Stream Mining Processing (DSMP)*, 2016, pp. 263–268.

[5] N. Shvai, A. Hasnat, A. Meicler, and A. Nakib, "Accurate classification for automatic vehicle-type recognition based on ensemble classifiers," *IEEE Transactions on Intelligent Transportation Systems*, vol. 21, no. 3, pp. 1288–1297, 2020.

[6] N. Shvai, A. Meicler, A. Hasnat, E. Machover, P. Maarek, S. Loquet, and A. Nakib, "Optimal ensemble classifiers based classification for automatic vehicle type recognition," in *2018 IEEE Congress on Evolutionary Computation (CEC)*, 2018, pp. 1–8.

[7] 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," in *2016 IEEE First International Conference on Data Stream Mining Processing (DSMP)*, 2016, pp. 23–28.

[8] R. Jacobi and A. Aginskiy, "Choosing 60-GHz mmwave sensors over 24-GHz to enable smarter industrial applications." [Online]. Available: http://www.ti.com/lit/wp/spry328/spry328.pdf

[9] A. Hasnat, N. Shvai, A. Sanogo, M. Khata, A. Llanza, A. Meicler, and A. Nakib, "Application guided image quality estimation based on classification," in *2019 IEEE International Conference on Image Processing (ICIP)*, 2019, pp. 549–553.

[10] A. Karpatne, G. Atluri, J. Faghmous, M. Steinbach, A. Banerjee, A. Ganguly, S. Shekhar, N. Samatova, and V. Kumar, "Theory-guided data science: A new paradigm for scientific discovery from data," *IEEE Transactions on Knowledge and Data Engineering*, vol. 29, no. 10, pp. 2318–2331, 10 2017.

[11] W. Wang, X. Zhou, B. Zhang, and J. Mu, "Anomaly detection in big data from uwb radars," *Security and Communication Networks*, vol. 8, no. 14, pp. 2469–2475, 2015. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/sec.745

[12] W. Wang, M. Zhang, D. Wang, Y. Jiang, Y. Li, and H. Wu, "Anomaly detection based on kernel principal component and principal component analysis," in *Communications, Signal Processing, and Systems*, Q. Liang, J. Mu, M. Jia, W. Wang, X. Feng, and B. Zhang, Eds. Singapore: Springer Singapore, 2019, pp. 2222–2228.

[13] I. Karlashevych and V. Pravda, "Use of cluster analysis method to increase the efficiency and accuracy of radar data processing," *Computational Problems Of Electrical Engineering*, vol. 7, no. 1, pp. 33–36, 2017. [Online]. Available: http://ena.lp.edu.ua:8080/handle/ntb/41500

[14] "IWR6843 single-chip 60-GHz to 64-GHz intelligent mmWave sensor integrating processing capability." [Online]. Available: http://www.ti.com/product/IWR6843

[15] "mmWave software development kit (SDK)." [Online]. Available: http://www.ti.com/tool/MMWAVE-SDK

[16] V. Dham, "Programming chirp parameters in TI radar devices," May 2017, Application Report. [Online]. Available: http://www.ti.com/lit/an/swra553/swra553.pdf

[17] G. Kriukova and M. Glybovets, "High-performance data stream mining by means of embedding hidden Markov model into reproducing kernel Hibert spaces," in *2018 IEEE Second International Conference on Data Stream Mining Processing (DSMP)*, 2018, pp. 207–211.

[18] S. Shklyar, "Conditions for the consistency of the total least squares estimator in an errors-in-variables linear regression model," *Theory of Probability and Mathematical Statistics*, vol. 83, pp. 175–190, 2011. [Online]. Available: https://doi.org/10.1090/S0094-9000-2012-00850-8

[19] S. V. Masiuk, A. G. Kukush, S. V. Shklyar, M. I. Chepurny, and I. A. Likhtarov, *Radiation Risk Estimation: Based on Measurement Error Models*, ser. De Gruyter Series in Mathematics and Life Sciences. De Gruyter, 2017. [Online]. Available: https://books.google.com.ua/books?id=Sq3bDgAAQBAJ

IEEE Third International Conference on Data Stream Mining & Processing
August 21-25, 2020, Lviv, Ukraine

Application of Ontologies And Meta-Models for Dynamic Integration of Weakly Structured Data

Andrii Berko, Irina Pelekh, Liliya Chyrun,
Myroslava Bublyk, Ihor Bobyk, Yurii Matseliukh
Lviv Polytechnic National University
Lviv, Ukraine
Andrii.Y.Berko@lpnu.ua

Lyubomyr Chyrun
Programming Department
Ivan Franko National University of Lviv
Lviv, Ukraine
Lyubomyr.Chyrun@lnu.edu.ua

***Abstract*— The possibilities of using ontologies and meta-models to create a system of dynamic integration of weakly structured data are considered. The process of transformation of weakly structured data into structured information is described. Dynamic integration of weakly structured data at different levels of complexity: physical, logical, and global is considered. The operation principles of the Universal browser for dynamic integration of weakly structured data are proposed to be applied when creating various web-systems.**

Keywords*— *dynamic integration; ontology; meta-model; web-system, weakly structured data

I. INTRODUCTION

Constant innovation, the rapid development of social progress, the discovery of new technological innovations cause the active use of web-systems in various spheres of human activity [1-3]. The use of this type of information resources makes it possible to quickly find the necessary information without leaving home or office [4-9]. Today, due to the constant development of information technologies in web-systems, quite significant volumes of weakly structured data of various nature have been accumulated and continue to grow rapidly. Hence, there is a need for their rapid, dynamic integration for the convenience of presenting these data and subsequent use [10-14]. The data integration task is to connect data from different sources and provide the user with a single (unified) representation of this data, including the ability to highlight information of interest to the user on request. The data integration system allows you to free the user from the need to independently select sources that contain the information that the user needs, access each source separately, and manually compare and combine data from different sources. The data integration role increases as the volume and need for data sharing increases. The problem of dynamic data integration is extremely multidimensional and remains unresolved. The complexity and nature of the methods used to solve it depend significantly on the integration level that needs to provide, the properties of individual data sources and the totality of sources as a whole, and the necessary integration methods.

II. ANALYSIS OF RECENT RESEARCH AND PUBLICATIONS

Scientific and engineering communities of the world have been dealing with the problem of the dynamic integration of information resources for many years. Only in recent years, modern infrastructure (WEB and GRID environments) and open service-oriented architectures in the field of information technology (SOA, OGSA) have appeared and are actively developing. It is as well as significant advances in the development of relevant international basic standards of information exchange (XML, RDF, TM, OWL) allow to create fundamentally new IP models. Such models make it possible to build globally-distributed applications that implement technological chains, which can use not only own IP but also those that can be offered by other organizational structures [3, 15-21]. This takes into account the fact that while working with such an application, it is possible to replace one information service with another, delete those that have lost value or relevance, and add new ones. The integration process is hampered by the heterogeneity of data sources, according to the integration level [22-27].

- When integrating at the physical level, different file formats can be used in the data sources.
- At the logical level of integration, there may be the heterogeneity of the data models used for different sources and differences in data schemes, although the same data model is used. Some sources can be websites and others can be object databases, and so on.
- When integrating globally, different ontologies may correspond to different data sources. For example, each of the sources can represent information resources that model a certain fragment of the subject area with its own conceptual system, and these fragments intersect.

There are the following problems of data integration [28]:

- Heterogeneity: data sources use different models (and even metamodels).
- Autonomicity: the sources are operated independently of each other, independently designed to solve specific, different problems, using different methods.
- Distributivity: sources are physically or logically accessible only through network protocols of remote

978-1-7281-3214-3/20/$31.00 ©2020 IEEE

access, in particular, information sources can be distributed on the Internet.

There are these types of data schema mismatch [29-32]:

- Heterogeneity conflicts (different data models are used for different sources).
- Naming conflicts (different schemes use different terminology, which leads to homonymy and synonymy in naming).
- Semantic conflicts (different levels of abstraction are chosen for modelling similar entities of the real world).
- Structural conflicts (the same entities are represented in different sources by different data structures).

Among the main tools used to ensure the integration of information resources are data converters that integrate data models, data model mapping mechanisms, object adapters (Wrappers), mediators (Mediators), ontological specifications, means of integration of schemes and integration of ontological specifications [3-7, 33-37], as well as architecture that provides interaction of tools used in a particular system of resource integration. Integration using ontological specifications allows you to highlight the data stored in each data source, and link mutually with logical and global ontologies by displaying (Fig. 1) [8-12, 38-43].

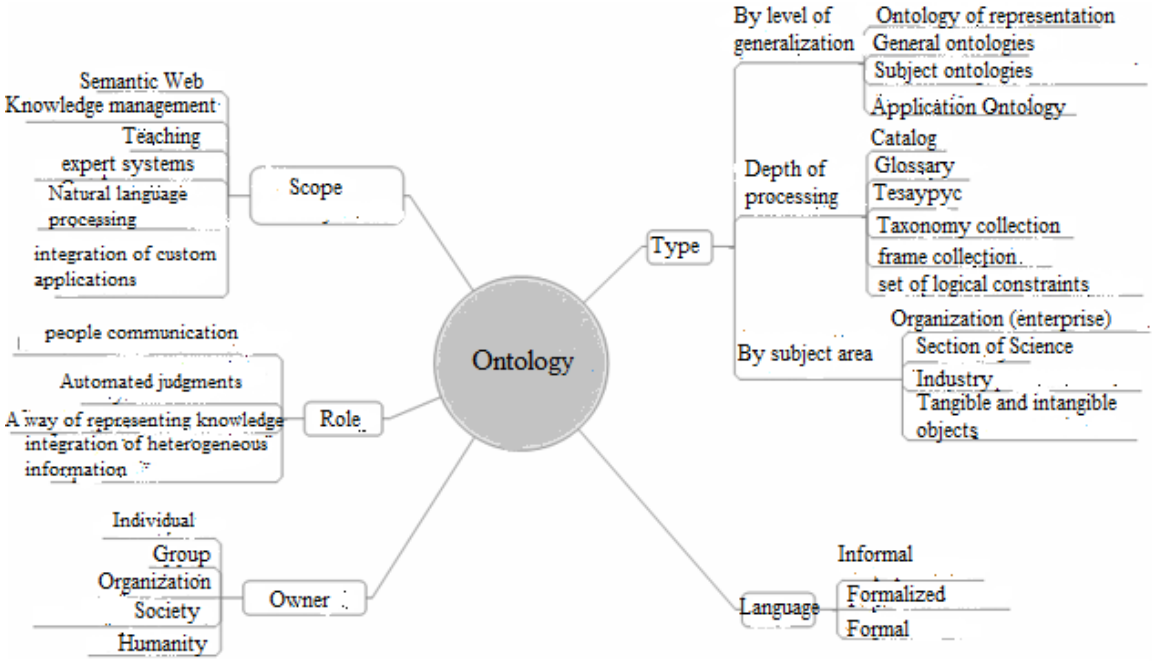


Fig. 1. Systematization of knowledge in the field of ontologies

Ontology is a formal specification of a separating conceptual model. Ontology consists of domain entity classes, properties of these classes, relationships between these classes, and statements built based on these classes, their properties, and relationships between them. Fig. 1 shows applications, roles, types, presentation languages, and ontology owners [44-49]. Domain ontology modelling is an expensive and time-consuming task that requires expert input. However, the core community should only do this work once, which is able to develop and maintain dictionaries using the collaborative tools developed by the Semantic Web community. Nevertheless, also, non-experts can add and display (map) new data sources using widely recognized auxiliary ontologies, such as Dublin Core, FOAF, SOAP, DOAP, etc. For a specific subject area, the following approach is considered reasonable: first search for existing ontologies, and then fill the gaps between them on the basis of a bottom-up approach from own concepts to more General ones (Fig. 2) [3].

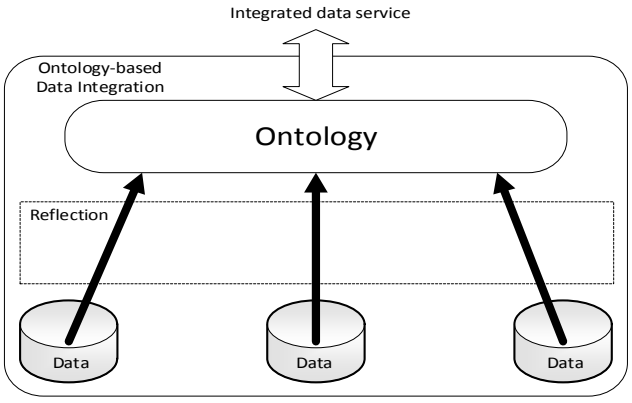


Fig. 2. Ontology-based data integration

III. Analysis of the Obtained Scientific Results

Data integration is considered at different levels of complexity (Fig. 3) [4, 20-21].

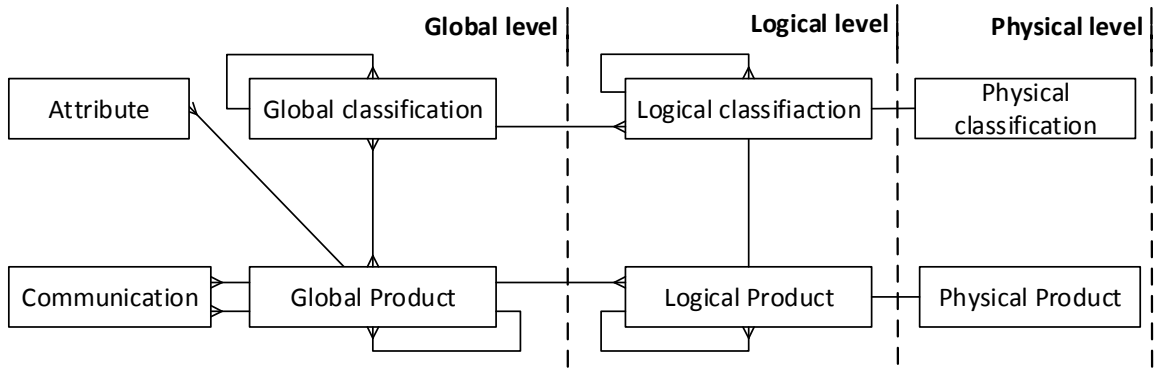


Fig. 3. Data integration at different levels of complexity [4]

The use of different methods and methods of integration at each level, depending on the complexity of the level, greatly simplifies the process of integration of weakly structured data, especially in web-based systems. At the physical level, where all data is from a single data source, it is not converted, but only placed as different combinations or images. The database browser performs these functions. The second level is logical. Data is still obtained from a single source, but it is obtained using adapters. Adapters (wrappers) are components associated with data sources to solve technical heterogeneity and metamodeling heterogeneity problems. Their main functions: accept source requests in some language, turn the request into the source language, execute the request, and send the results to the mediator. Means of intermediaries supported a unified meta descriptions of the integrated data sources. As a rule, semantic intermediaries are developed for a specific narrow subject area. Mediation mechanisms rely on ontological specifications of sources. An integrated ontology of the sources used is created for the mediator. Such systems also require an integrating data model with advanced data semantics modelling capabilities. At the global level (system integration level), it is proposed to use a Universal structured data browser [4], the principles of which are shown in Fig. 4. [4]

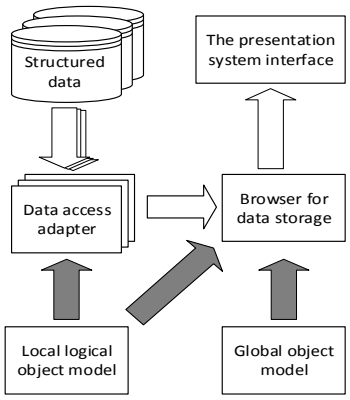


Fig. 4. System of operation of the Universal structured data browser [4]

The use of a Universal structured data browser is extremely important in web-systems, because even structured information that is taken from the Internet is not always correctly filed and depicted. In order to properly submit to the user the information obtained from several web sites, it first needs to be correctly syntactically and semantically issued. As already mentioned, the integration of weakly structured data needs to carry out at three levels. At the top, global level, we integrate data using a universal structured data browser. Therefore, before using the universal structured data browser, the data needs to be structured. The transformation of weakly structured data into structured data is a rather time-consuming process that requires high accuracy, so the so-called Universal browser of weakly structured data is used to work with weakly structured data, which is described in detail in [4] and shown in Fig. 5. [4]

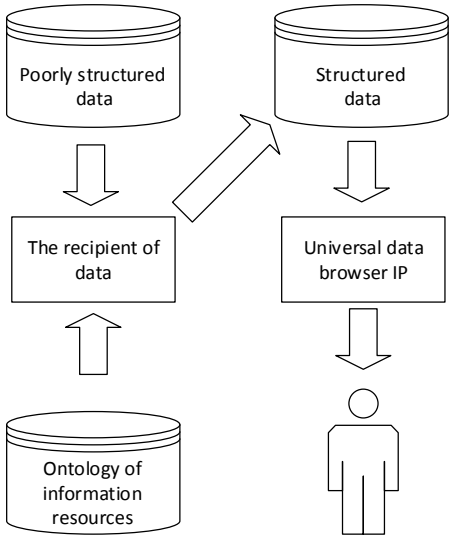


Fig. 5. Conceptual model of universal browser of weakly structured data [4]

The integration of data in an information system naturally involves the integration in some form of metadata that determines their sources. One of the traditional tasks of metadata integration in structured data integration systems is schema integration. Difficulties in addressing it in specific situations may relate to the presence of conflicts, for example:

- Heterogeneity conflicts (different data models are used for different sources);

- Name conflicts (different schemes use different terminology, resulting in homonymy and synonymy in naming);
- Semantic conflicts (different levels of abstraction are chosen for modelling similar entities of the real world);
- Structural conflicts (the same entities are represented in different sources by different data structures).

Let's consider the work of the Universal browser of weakly structured data and how for our case it is possible to turn weakly structured data into structured data. Therefore, for us, the input information is a certain class of weakly structured data. Within this class, we find data that consists of two parts - hierarchical classification and object-specific data. This data may be partially structured. As an example, take the product catalogue of the manufacturer or seller. The mechanism for converting weakly structured data into structured data is shown in [4] (Fig. 6).

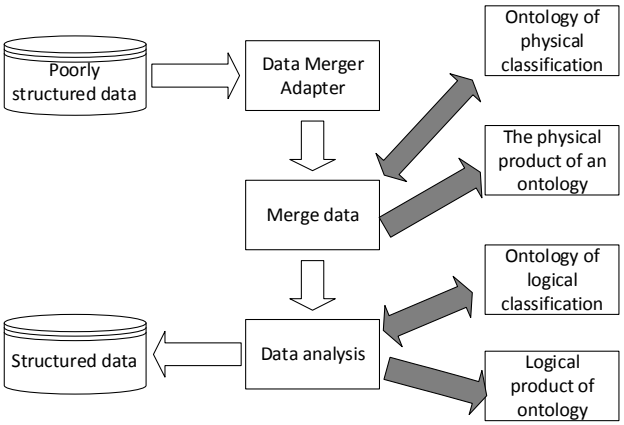


Fig. 6. The principle of transformation of weakly structured data [4]

We will show how this mechanism can applied to the operation of our Universal browser of weakly structured data. Typically, each of the weakly structured data sources has a different format. Sites are analysed using adapters that can analyse web pages and get the necessary information from them. If this is a product classification page, the adapter receives this classification. If the page contains General product information, the specific product data will retrieved. A Web site usually has a home page from which a variety of links can be used to, for example, view the entire catalogue of information on the classification of a particular product, as well as information describing that product. Before analysing data sources and getting data from them, you need to analyse web pages and install ontologies for local products. In addition, an adapter must be installed that can analyse a specific data source. Data is populated and updated by using data source model information that is stored in the ontology product. The process of recording accurate data in the product database is happening at the next stage. The Universal structured data browser works with this already structured data. Therefore, the main task is to convert weakly structured data into structured data. Weakly structured data after entering the system is transformed by using a data Association adapter that combines data using the ontology of physical classification. The data, after conversion, goes to the analysis stage. When analysing data, the system of the universal browser of weakly structured data by using the ontology of logical classification converts them into already structured data, which the universal browser of structured data works with. Working with already structured data is as follows: using the adapter, data is obtained from the database, and then the data is transferred to the browser for data storage, which then presents the data to the user. The storage browser receives and sends a user request to the system. To work with weakly structured data and convert it into structured data, the system of the Universal browser of weakly structured data is used, which is shown in Fig. 5. The system of the Universal structured data browser, which is shown in Fig. 3, works with already structured data. If the generic structured data browser system uses only one metadata database, the Generic structured data browser system will use information from multiple classification ontologies and ontology products to search for and view the data. The structured database serves as a data source for the Universal structured data browser system. The data access adapter allows the system to access information that is stored in a structured database. The system does not have any specific access functions that will allow access to the structured database, so the adapter interprets data from the ontology product to generate dynamic queries to select information from the structured database. Therefore, the main idea of dynamic integration of weakly structured data in web-systems is that ontologies and meta-models are used and integration occurs at three levels of complexity. The work of the web-system, which works on the principle of dynamic integration of weakly structured data, using ontologies and meta-models, is as follows: the user sends a request through the system interface; the system analyses the request and dynamically receives data from various web-sources. The Universal browser of weakly structured data works with this weakly structured data, further. It is converts them into structured information. The Universal browser of structured data works with already structured data semantically and syntactically forms them and gives out already issued data through the user interface.

IV. CONCLUSIONS

The main provisions of dynamic integration of weakly structured data in web-systems using ontologies and meta-models are formulated. The functioning of the Universal browser of dynamic integration of weakly structured data in web-systems is described. The transformation principles of weakly structured data into structured information are considered. The data integration at different levels of complexity is presented: physical, logical, and global. The conceptual model of the universal data storage structure is described. The described principles of operation of the Universal browser can apply when creating web-systems for dynamic integration of weakly structured data of different nature.

REFERENCES

[1] V. Lytvyn, V. Vysotska, D. Dosyn, and Y. Burov, “Method for ontology content and structure optimization, provided by a weighted conceptual graph,” Webology, vol. 15(2), 2018, pp. 66-85.

[2] V. Lytvyn, Y. Burov, P. Kravets, V. Vysotska, A. Demchuk, A. Berko, Y. Ryshkovets, S. Shcherbak, and O. Naum, “Methods and Models of Intellectual Processing of Texts for Building Ontologies of Software for Medical Terms Identification in Content Classification,” CEUR Workshop Proceedings, Vol-2362, 2019, pp. 354-368.

[3] D.O. Briukhov, and S.S. Shumilov, "Ontology Specification and Integration Facilities in a Semantic Interoperation Framework," Proc. of the Second Intern. Workshop ADBIS'95. Moscow, 1995, 195-200.

[4] G. Arnicans, and G. Karnitis, "Intelligent Integration of Information From Semi-Structured Web Data Sources on the Basis of Ontology and Meta-Models," 7th International Baltic Conference on Databases and Information Systems Vilnius, Lithuania, 2006, pp. 177-186.

[5] A.Y. Levy, "Logic-Based Techniques in Data Integration. Logic-based Techniques in Data Integration," Logic Based Artificial Intelligence. Edited by J. Minker. Kluwer Publishers, 2000.

[6] I. Manolescu, D. Florescu, and D. Kossman, "Answering XML Queries over Heterogeneous Data Sources," Proc. Of the 27th VLDB Conference, Roma, Italy, 2001.

[7] R. Baumgartner, G. Gottlob, M. Herzog, and W. Slany, "Interactively adding web service interfaces to existing web applications," Int. Symposium on Applications and the Internet, 2004, pp. 74-80.

[8] G. Canfora, A.R. Fasolino, G. Frattolillo, and P. Tramontana, "Migrating interactive legacy systems to web services," IEEE CS Press, editor, European Conference on Software Mainteinance and Reengineering, 2006, pp. 23-32.

[9] H. Guo, C. Guo, F. Chen, and H. Yang, "Wrapping client-server application to web services for internet computing," International Conference on Parallel and Distributed Computing, Applications and Technologies, 2005, pp. 366–370.

[10] A.Y. Berko, and K.A. Aliekseyeva, "Quality evaluation of information resources in web-projects," Actual Problems of Economics, vol. 136(10), 2012, pp. 226-234.

[11] K. Alieksieieva, A. Berko, V. Vysotska, "Technology of commercial web-resource processing," Proceedings of 13th International Conference: The Experience of Designing and Application of CAD Systems in Microelectronics, CADSM 2015-February, 2015.

[12] V. Lytvyn, V. Vysotska, D. Dosyn, O. Lozynska, and O. Oborska, "Methods of Building Intelligent Decision Support Systems Based on Adaptive Ontology," IEEE 2nd International Conference on Data Stream Mining and Processing, DSMP, 2018, pp. 145-150.

[13] Y. Burov, V. Vysotska, and P. Kravets "Ontological approach to plot analysis and modeling," CEUR Workshop Proceedings, Vol-2362, 2019, pp. 22-31.

[14] S. Sachenko, S. Rippa, and Y. Krupka, "Pre-Conditions of Ontological Approaches Application for Knowledge Management in Accounting," International Workshop on Antelligent Data Acquisition and Advanced Computing Systems: Technology and Applications, 2009, pp. 605-608.

[15] M. Davydov, and O. Lozynska, "Mathematical method of translation into Ukrainian sign language based on ontologies," Advances in Intelligent Systems and Computing, vol. 871, 2018, pp. 89-100.

[16] S. Babichev, "An Evaluation of the Information Technology of Gene Expression Profiles Processing Stability for Different Levels of Noise Components," Data, vol. 3 (4), 2018, art. no. 48.

[17] S. Babichev, B. Durnyak, I. Pikh, and V. Senkivskyy, "An Evaluation of the Objective Clustering Inductive Technology Effectiveness Implemented Using Density-Based and Agglomerative Hierarchical Clustering Algorithms," Advances in Intelligent Systems and Computing, vol. 1020, 2020, pp. 532-553.

[18] B. Rusyn, L. Pohreliuk, A. Rzheuskyi, R. Kubik, Y. Ryshkovets, L. Chyrun, S. Chyrun, A. Vysotskyi, and V.B. Fernandes, "The Mobile Application Development Based on Online Music Library for Socializing in the World of Bard Songs and Scouts' Bonfires," Advances in Intelligent Systems and Computing IV, Springer, 1080, 2020, pp. 734-756.

[19] V. Andrunyk, L. Chyrun, V. Vysotska, "Electronic content commerce system development," Proceedings of 13th International Conference: The Experience of Designing and Application of CAD Systems in Microelectronics, CADSM 2015-February, 2015.

[20] V. Vysotska, V. Lytvyn, Y. Burov, A. Gozhyj, S: Makara, "The consolidated information web-resource about pharmacy networks in city," CEUR Workshop Proceedings, 2018, pp. 239-255.

[21] Vysotska, V., Fernandes, V.B., Emmerich, M.: Web content support method in electronic business systems. In: CEUR Workshop Proceedings, Vol-2136, 20-41. (2018)

[22] A. Gozhyj, I. Kalinina, V. Vysotska, and V. Gozhyj, "The method of web-resources management under conditions of uncertainty based on fuzzy logic," Proceedings of the International Conference on Computer Sciences and Information Technologies, CSIT, 2018, pp. 343-346.

[23] A. Gozhyj, V. Vysotska, I. Yevseyeva, I. Kalinina, and V. Gozhyj, "Web Resources Management Method Based on Intelligent Technologies," Advances in Intelligent Systems and Computing, 871, 2019, pp. 206-221.

[24] B. Rusyn, V. Vysotska, and L. Pohreliuk, "Model and architecture for virtual library information system," Proceedings of the International Conference on Computer Sciences and Information Technologies, CSIT, 2018, pp. 37-41.

[25] V. Lytvyn, and V. Vysotska, "Designing architecture of electronic content commerce system," Proceedings of the Internat. Conference on Computer Sciences and Information Technologies, 2015, pp. 115-119.

[26] T. Batiuk, V. Vysotska, and V. Lytvyn, "Intelligent System for Socialization by Personal Interests on the Basis of SEO-Technologies and Methods of Machine Learning,"CEUR workshop proceedings, Vol-2604, 2020, pp. 1237-1250.

[27] J. Su, A. Sachenko, V. Lytvyn, V. Vysotska, and D. Dosyn, "Model of Touristic Information Resources Integration According to User Needs," Proceedings of the International Conference on Computer Sciences and Information Technologies, CSIT, 2018, 113-116.

[28] A. Alhegami, and H. Alsaeedi, "A framework for incremental parallel mining of interesting association patterns for big data," International Journal of Computing, 19(1), 2020, pp. 106-117.

[29] M. Komar, V. Dorosh, P. Yakobchuk, V. Golovko, and A. Sachenko, "Deep Neural Network for Image Recognition Based on the Caffe Framework," IEEE Second International Conference on Data Stream Mining & Processing, 2018, pp. 102-106.

[30] I. Paliy, A. Sachenko, V. Koval, and Y. Kurylyak, "Approach to Face Recognition Using Neural Networks," Proceedings of the International Workshop on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications, 2005, pp. 112-115.

[31] A. Sachenko, V. Kochan, and V. Turchenko, "Instrumentation for Gathering Data," IEEE Instrumentation & Measurement Magazine, Vol. 6(3), 2003, pp.34-40.

[32] O. Kuzmin, and M. Bublyk, "Economic evaluation and government regulation of technogenic (man-made) damage in the national economy," Computer sciences and information technologies, CSIT, 2016, pp. 37-39.

[33] M.I. Bublyk, and O.M. Rybytska, "The model of fuzzy expert system for establishing the pollution impact on the mortality rate in Ukraine," Computer sciences and information technologies, CSIT, 2017, pp. 253–256.

[34] M. Bublyk, O. Rybytska, A. Karpiak, and Y. Matseliukh, "Structuring the fuzzy knowledge base of the IT industry impact factors," Computer sciences and information technologies, 2018.

[35] N. Chukhray, N. Shakhovska, O. Mrykhina, M. Bublyk, and L. Lisovska, "Consumer aspects in assessing the suitability of technologies for the transfer," Computer sciences and information technologies, CSIT, 2019, pp. 142–147.

[36] N. Chukhray, N. Shakhovska, O. Mrykhina, M. Bublyk, and L. Lisovska, "Methodical Approach to Assessing the Readiness Level of Technologies for the Transfer," Advances in Intelligent Systems and Computing IV, 1080, Springer, 2020, pp. 259-282.

[37] L. Giusy, and H. Hakim, "Mashups for data integration: An analysis," Technical Report UNSW-CSE-TR-0810, 2008, pp. 68-69.

[38] G. Canfora, A.R. Fasolino, G. Frattolillo, and P. Tramontana, "Migrating interactive legacy systems to web services," IEEE CS Press, editor, European Conference on Software Mainteinance and Reengineering, 2006, pp. 23-32.

[39] V. Vysotska, R. Hasko, and V. Kuchkovskiy, "Process analysis in electronic content commerce system," Computer Sciences and Information Technologies, CSIT, 2015, pp. 120-123.

[40] V. Lytvyn, V. Vysotska, O. Veres, I. Rishnyak, and H. Rishnyak, "The Risk Management Modelling in Multi Project Environment," Proceedings of the International Conference on Computer Sciences and Information Technologies, CSIT, 2017, pp. 32-35.

[41] 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, 689, Springer, 2018, pp. 204-216.

[42] V. Lytvyn, A. Gozhyj, I. Kalinina, V. Vysotska, V. Shatskykh, L. Chyrun, Y. Borzov, "An intelligent system of the content relevance at the example of films according to user needs," CEUR Workshop Proceedings, Vol-2516, 2019, pp. 1-23.

[43] P. Kravets, V. Lytvyn, V. Vysotska, and Y. Burov, "Promoting training of multi-agent systems," CEUR Workshop Proceedings, Vol-2608, 2020, pp. 364-378.

[44] V. Lytvyn, and V. Vysotska, "Designing architecture of electronic content commerce system," Proceedings of the Internat. Conference on Computer Sciences and Information Technologies, 2015, pp. 115-119.

[45] O. Naum, L. Chyrun, O. Kanishcheva, and V. Vysotska, "Intellectual System Design for Content Formation," Proceedings of the International Conference on Computer Sciences and Information Technologies, CSIT, 2017, pp. 131-138.

[46] V. Lytvyn, V. Vysotska, D. Dosyn, and Y. Burov, "Method for ontology content and structure optimization, provided by a weighted conceptual graph," Webology, vol. 15(2), 2018, pp. 66-85.

[47] N. Antonyuk, A. Vysotsky, V. Vysotska, V. Lytvyn, Y. Burov, A. Demchuk, I. Lyudkevych, L. Chyrun, S. Chyrun, and I. Bobyk, "Consolidated Information Web Resource for Online Tourism Based on Data Integration and Geolocation," Conference on Computer Sciences and Information Technologies, CSIT, 2019, pp. 15-20.

[48] A. Vysotsky, V. Lytvyn, V. Vysotska, D. Dosyn, I. Lyudkevych, N. Antonyuk, O. Naum, A. Vysotskyi, L. Chyrun, and O. Slyusarchuk, "Online Tourism System for Proposals Formation to User Based on Data Integration from Various Sources," Computer Sciences and Information Technologies, CSIT, 2019, pp. 92-97.

[49] O. Lozynska, V. Savchuk, and V. Pasichnyk, "Individual Sign Translator Component of Tourist Information System," Advances in Intelligent Systems and Computing, Springer, 1080, 2020, pp. 593-601.

IEEE Third International Conference on Data Stream Mining & Processing
August 21-25, 2020, Lviv, Ukraine

Big Data-Based Approach to Automated Linguistic Analysis Effectiveness

Andriy Lutskiv
Computer Systems and Networks Department
Ternopil Ivan Puluj National Technical University
Ternopil', Ukraine
Big Data Engineering Department
Dataengi
Vilnius, Lithunia
0000-0002-9250-4075

Nataliya Popovych
Department of Multicultural Education and Translation
State University "Uzhhorod National University"
Uzhhorod, Ukraine
Translation Services and Italian Customer Support Workconsult GmbH
Leipzig, Germany
0000-0001-6949-0771

***Abstract* — The authors of the finding develop customizable corpus tool to build corpus of historical and religious texts. Big Data approach to Natural Language Processing and Natural Language Understanding was used to achieve the goal of such corpus data platform developing. Calculation of qualitative and quantitative characteristics, building search queries belong to the most important features of the adaptable text corpus effectiveness. Number of computer-based calculations and amount of processing data have been reduced and parallelized to achieve higher performance on the levels of computational methods and implemented system. The higher level of efficiency as a trade-off between effectiveness and computational time has been achieved by choosing proper parameters of computational methods. Latent-Semantic Analysis is used as one of the core methods for making queries. The methods applied are mostly based on Singular Value Decomposition. Parameters of the decomposition are analyzed and justified. Suggested approach has been verified on the test data available in different languages.**

Keywords — Big Data, Corpus Linguistic, Latent-Semantic Analysis, Singular Value Decomposition, Natural Language Processing, adaptable text corpus, corpus-based translation studies

I. INTRODUCTION

While doing lexico-stylistic research of the text belonging to different genres and styles linguists and translators face the problems of:

- identifying the equivalence level in source and target languages;
- creating different search queries based on semantic closeness and text matching;
- finding lexico-semantic or stylistic equivalence inadequacies;
- automating marked up language search for resolving specific linguistic tasks created to process large amount of text data etc.

Adaptable text corpora are aimed to resolve such and similar linguistic problems. Usual text corpus contains marked-up source texts with the developed tag sets, relations between words and groups of words, qualitative and quantitative characteristics [1]. Such calculations cannot be done in real-time mode because they are time and memory consuming. Rather small amounts of source data turn into a large amount of different additional information needed for deeper, more sophisticated, multilayered and complex text analysis to meet specific linguistic needs. Amount of data which are stored by corpus can be qualified as Big Data. Thus, approaches which are used for Big Data processing are suitable for linguistic corpus development. Such corpus tool development involves scientific and Big Data engineering problem solving such as choosing proper mathematical model and methods, algorithms and data structures, programming technologies, software frameworks and libraries. The development verification is actualized by the analysis of the test text datasets processing. Corpus tool implementation is verified by professional linguist who has gained enough experience in using automated marked up language search and has got sufficient expertise in this domain. Hence, similar problems can be resolved effectively only through the joint efforts of Big Data engineers and linguists to maximize the desirable results.

Great amount of different corpus tools and corpora are available today. The thorough analysis and classification of corpora and corpus tools is presented in the former papers [2] of the authors. The corpus tools and corpora under mentioned classification have some disadvantages such as lack of customization, high ambiguity indicator, low effectiveness in large amount of text data processing. Lack of customization means that linguist cannot use custom or his own models of text components and custom methods for text processing. The use of input texts in ancient languages is also questionable, because it requires specific additional linguistic methods or approaches to the queries and additional preprocessing of them. This and the former papers are the results of research done for developing adaptable text corpus tool aimed at processing specific types of texts on the example of religious text data [2].

Using different Natural Language Processing methods, adaptable corpus tool [3] resolves the following tasks:

1) Extracts terms (words), sentences, groups of 2, 3, 4 and 5 terms (n-grams). N-grams with higher probabilities can be treated as collocations or colligations.

978-1-7281-3214-3/20/$31.00 ©2020 IEEE

2) Part of speech tags (POS-tags) and from syntactical parses extracted terms and sentences.

3) Calculates quantitative characteristics, such as frequencies and term-frequencies-invert-document-frequencies (TF-IDFs) [4] of tagged terms and n-grams..

One of the most important function of the text corpus is making search queries to find similarities between words and groups of words. Focusing on this function implementation it was decided [5] to use Latent-Semantic Analysis (LSA) approach. LSA [6-8] is successfully used in the web search services for different tasks which are of great importance for corpus tools, i.e., information retrieval, information classification (filtering) and cross-language retrieval. The latter is of extreme importance for multilingual corpus queries. Singular Value Decomposition (SVD) [4,9,10] is a key component of this technique and allows reducing computational and memory complexities. Value as an initial parameter should be set for the effective use of the optimal SVD rank. Thus, objectives of this finding are focused on:

- finding optimal SVD rank with saving maximum energy [4];
- measuring computation time of max SVD rank and optimal SVD rank to evaluate effectiveness;
- making numerical and expert analysis of chosen rank values for the verification of the suggested approaches.

It is important to note that among different computational challenges of LSA estimating optimal SVD rank was mentioned by Susan T. Dumais [6]. This finding resolves the tasks for specific text query such as historical and religious texts in English, Ukrainian and Russian.

The approaches suggested in this finding are implemented as software. By carrying out computational experiment these approaches have been verified. Different editions of the Bible in English (HCSB[11], NIV[12], KJVA[13]), Ukrainian (GYZ[14], OH62[15]) and Russian(NRT[16]) were chosen as source data for computational experiment.

The main advantage of the source data chosen as a test dataset consists in the well-knownness and availability of texts in the different languages.

II. Mathematical Model and Computational Methods

A. Mathematical model of the linguistic corpus

A number of sets (***L***, ***T***, ***A***, ***B***, ***C***, ***S***) have been suggested [5] to use as a mathematical model of the linguistic corpus. ***T*** - set of natural language terms or words *t*. Set ***L*** is a superset over the ***T***-sets that comprises terms from all languages:

$$\boldsymbol{L}=\{\boldsymbol{T}_1,\boldsymbol{T}_2,\ldots,\boldsymbol{T}_n\},\ \boldsymbol{T}=\{t_1,t_2,\ldots,t_n\},\ n\in\boldsymbol{N}. \quad (1)$$

There is a subset $\boldsymbol{T'}$ of set ***T*** which contains terms that belong to book ingested into corpus:

$$\boldsymbol{T'}=\{\boldsymbol{t'}_1,\boldsymbol{t'}_2,\ldots,\boldsymbol{t}_n\},\ n\in\boldsymbol{N}. \quad (2)$$

Set ***A*** — the whole corpus of one language which comprises books (set ***B***). Books are divided into chapters or documents (set ***C***). Chapter contains sentences (set ***S***) which comprise words $(t'_{ijkr};r)$ — pairs of term and its position in the sentence.

$$\boldsymbol{A}=\{\boldsymbol{B}_1,\boldsymbol{B}_2,\ldots,\boldsymbol{B}_n\},\quad \boldsymbol{B}_i=\{\boldsymbol{C}_{i1},\boldsymbol{C}_{i2},\ldots,\boldsymbol{C}_{im}\},\quad i\in\overline{1,n},$$

$$\boldsymbol{C}_{ij}=\{\boldsymbol{S}_{ij1},\boldsymbol{S}_{ij2},\ldots,\boldsymbol{S}_{ijp}\},\quad j\in\overline{1,m}, \quad (3)$$

$$\boldsymbol{S}_{ijk}=\{(t'_{ijk1};\boldsymbol{1}),(t'_{ijk2};\boldsymbol{2}),\ldots,(t'_{ijkr};r)\},\quad k\in\overline{1,p},$$

$$\boldsymbol{T'}=\{t_1,t_2,\ldots,t_h\},\quad h\in\overline{1,q},$$

$$n,\quad m,\quad q,\quad p,\quad r\in\boldsymbol{N},\ \boldsymbol{S}\subset\boldsymbol{C}\subset\boldsymbol{B}\subset\boldsymbol{A},\ \boldsymbol{T'}\subset\boldsymbol{A}.$$

For navigating through the terms in the corpus each set has its own index which points to the position. In these sets: *n* — number of the books in the corpus, *m* — number of the chapters in any book of the corpus, *p* — number of the sentences in any chapter, *r* — number of the ordered terms in any sentence, *q* — number of unique terms in the corpus.

At the first stage of book ingestion into the corpus terms and sentence are tokenized, POS-tagged and syntactically parsed. Thus, $\widetilde{\boldsymbol{S}}_{\boldsymbol{ijk}}$ — set of syntactically parsed sentences and $(\widetilde{t}'_{ijkr};r)$ — tagged ordered terms in the sentences:

$$F_{tagging}: S_{ijk}\to\tilde{S}_{ijk},\ i\in\overline{1,n},\ j\in\overline{1,m},\ k\in\overline{1,p},$$

$$\tilde{S}_{ijk}=\{(\widetilde{t}'_{ijk\,1};1),(\widetilde{t}'_{ijk\,2};2),\ldots,(\widetilde{t}'_{ijk\,r};r)\}, \quad (4)$$

$$n,m,p,v,r\in\boldsymbol{N}.$$

One of the most time and memory consuming processes is n-grams extraction. According to the recommendations given by Y. Lin and coauthors [17] this task has been accomplished. N-gram extraction can be performed only after having tokenized sentences and terms. N-grams is a contiguous sequence of n-terms. Size of n-gram depends on the analyzed language lexical specificity in grammar, but the most common cases are 2-, 3-, 4- and 5-grams. 2-grams (bigrams) and 3-grams (trigrams) which have been POS-tagged can be written as follows:

$$\{(\widetilde{t}'_{ijk\,v};v),(\widetilde{t}'_{ijk\,v+1};v+1)\}, \quad (5)$$

$$i\in\overline{1,n},\ j\in\overline{1,m},\ k\in\overline{1,p},\ v\geq 1,\ v+1\leq r,$$

$$\{(\widetilde{t}'_{ijk\,v};v),(\widetilde{t}'_{ijk\,v+1};v+1),(\widetilde{t}'_{ijk\,v+2};v+2)\}, \quad (6)$$

$$i\in\overline{1,n},\ j\in\overline{1,m},\ k\in\overline{1,p},\ v\geq 1,\ v+2\leq r,$$

$$n,m,p,v,r\in\boldsymbol{N}.$$

To treat n-grams as collocations and colligations probabilities of n-gram appearance in the text should be calculated:

$$P\left((\widetilde{t}'_{ijk\,v+1};v+1)|(\widetilde{t}'_{ijk\,v};v)\right)=$$

$$=\frac{count((\widetilde{t}'_{ijk\,v};v),(\widetilde{t}'_{ijk\,v+1};v+1))}{count((\widetilde{t}'_{ijk\,v};v))}, \quad (7)$$

$$i\in\overline{1,n},\ j\in\overline{1,m},\ k\in\overline{1,p},\ v\geq 1,\ v+1\leq r$$

$$n,m,p,v,r\in\boldsymbol{N}.$$

Hence, the most important characteristics for dictionary building are the frequencies of each POS-tagged term in the chapter, frequencies of each POS-tagged term in the book and TF-IDFs. Consider $z^{C_{ij}}$ — number of terms in *j*-chapter of *i*-book, and z^{B_i} — number of terms in *i*-book, $z_{t'_h}$ - number of *h*-term $\widetilde{t'}_h$. Frequencies for each POS-tagged term in the chapter $v_{\widetilde{t'}}^{(C_{ij})}$ and in the book $v_{\widetilde{t'}}^{(B_{ij})}$ can be obtained as given below:

$$\left\{\left(v_{\widetilde{t'}_1}^{(C_{ij})}\right),\left(v_{\widetilde{t'}_2}^{(C_{ij})}\right),\dots,\left(v_{\widetilde{t'}_h}^{(C_{ij})}\right)\right\},\ v_{\widetilde{t'}_i}^{(C_{ij})}=\frac{z_{\widetilde{t'}_h}}{z^{C_{ij}}}, \quad (8)$$

$$\left\{\left(v_{\widetilde{t'}_1}^{(B_{ij})}\right),\left(v_{\widetilde{t'}_2}^{(B_{ij})}\right),\dots,\left(v_{\widetilde{t'}_h}^{(B_{ij})}\right)\right\},\ v_{\widetilde{t'}_i}^{(B_{ij})}=\frac{z_{\widetilde{t'}_h}}{z^{B_{ij}}}, \quad (9)$$

$$i\in\overline{1,n},\ j\in\overline{1,m},\ h\in\overline{1,q},\ n,m,q\in\boldsymbol{N}.$$

Task of finding the most important terms (set $\boldsymbol{T'_{KW}}$) in the document can be achieved by different methods, such as suggested by V. Lytvyn, V. Vysotska, D. Uhryn, M. Hrendus, O. Naum, U.Shandruk, P.Pukach, O. Brodyak, especially for Slavic languages [18, 19, 20]. Authors of this finding used approach based on stop-words filtering, finding lemmas, calculating TF-IDFs of each lemmas and ranging lemmas by TF-IDF criteria. Stop-words removing based on the dictionary (set $\boldsymbol{T'_{SW}}$), which is predefined by the linguist:

$$\boldsymbol{T'_{KW}}=\boldsymbol{T'}\setminus\boldsymbol{T'_{SW}},\ \boldsymbol{T'_{KW}}\subset\boldsymbol{T'},\ \boldsymbol{T'_{SW}}\subset\boldsymbol{T'}. \quad (10)$$

From the set $\boldsymbol{T'_{KW}}$ according to the language rules POS-tagged lemmas are obtained:

$$\boldsymbol{F_{lemmatization}}:\boldsymbol{T'_{KW}}\to\boldsymbol{T''_{KW}},\ \boldsymbol{T''_{KW}}=\left\{\widetilde{t''}_1,\widetilde{t''}_1,\dots,\widetilde{t''}_h\right\}. \quad (11)$$

It is considered that w^{B_i} is a number of chapters of *i*-book, and $w^{B_i\widetilde{t''}_h}$ is a number of some *h*-lemmas $\widetilde{t''}_h$ in *i*-book. Analogous to term frequencies lemmas' frequencies $\frac{z_{\widetilde{t''}_h}}{z^{C_{ij}}}$ should be calculated. TF-IDFs (value $\xi_{\widetilde{t''}_h}^{(C_{ij})}$) calculation can be written as follows:

$$\left\{\left(\xi_{\widetilde{t''}_1}^{(C_{ij})}\right),\left(\xi_{\widetilde{t''}_2}^{(C_{ij})}\right),\dots,\left(\xi_{\widetilde{t''}_h}^{(C_{ij})}\right)\right\}, \quad (12)$$

$$\xi_{\widetilde{t''}_h}^{(C_{ij})}=\frac{z_{\widetilde{t''}_h}}{z^{C_{ij}}}\cdot\log\left(\frac{|w^{B_i}|}{\left|w^{B_i\widetilde{t''}_h}\right|}\right), \quad (13)$$

$$i\in\overline{1,n},\ j\in\overline{1,m},\ h\in\overline{1,q},\ n,m,q\in\boldsymbol{N}.$$

Relations between lemmas, n-grams and documents can be represented as a lemma-document and n-gram-document matrices. Lemmas and documents are represented as vectors of TF-IDFs, thus lemmas are rows and documents (chapters of the book) are columns of the matrix $\boldsymbol{M}_{LD}$:

$$\boldsymbol{M}_{LD}=\begin{array}{c}\begin{matrix}C_{i1} & C_{i2} & \dots & C_{im}\end{matrix}\\ \begin{pmatrix}\xi_{\widetilde{t''}_1}^{(C_{i1})} & \xi_{\widetilde{t''}_1}^{(C_{i2})} & \dots & \xi_{\widetilde{t''}_1}^{(C_{im})}\\ \xi_{\widetilde{t''}_2}^{(C_{i1})} & \xi_{\widetilde{t''}_2}^{(C_{i2})} & \dots & \xi_{\widetilde{t''}_2}^{(C_{im})}\\ \vdots & \vdots & \ddots & \vdots\\ \xi_{\widetilde{t''}_q}^{(C_{i1})} & \xi_{\widetilde{t''}_q}^{(C_{i2})} & \dots & \xi_{\widetilde{t''}_q}^{(C_{im})}\end{pmatrix}\end{array}\begin{matrix}\widetilde{t''}_1\\ \widetilde{t''}_2\\ \vdots\\ \widetilde{t''}_q\end{matrix} \quad (14)$$

Analogous to lemma-document matrix n-gram-document matrices for 2-, 3-, 4- and 5-grams can be written as: $\boldsymbol{M}_{2GD}$, $\boldsymbol{M}_{3GD}$, $\boldsymbol{M}_{4GD}$ and $\boldsymbol{M}_{5GD}$. These matrices contain TF-IDFs of each n-gram which show importance of this lemma or n-gram.

Thus, all of the sets and matrices described above represent elements of the text corpus and relations between these elements. The same data elements were used in the developed corpus tool as a source for searching queries.

B. Processing Data for Making Search Queries

In making search queries to find similarities between text corpus elements (terms, groups of terms or n-grams) by cosine similarity between these elements vector representations (scalar product) [4,10] can be applied. Finding similarities between term $\widetilde{t''}_{h-1}$ and term $\widetilde{t''}_h$ in the book $\boldsymbol{B_i}$ can be represented by the following:

$$\cos\left(\varphi_{(\widetilde{t''}_{h-1},\widetilde{t''}_h)}\right)=\frac{\sum_{j=1}^{m}\xi_{\widetilde{t''}_{h-1}}^{(C_{ij})}\cdot\xi_{\widetilde{t''}_h}^{(C_{ij})}}{\sqrt{\sum_{j=1}^{m}\left(\xi_{\widetilde{t''}_{h-1}}^{(C_{ij})}\right)^2}\cdot\sqrt{\sum_{j=1}^{m}\left(\xi_{\widetilde{t''}_h}^{(C_{ij})}\right)^2}}, \quad (15)$$

$$i\in\overline{1,n},\ j\in\overline{1,m},\ h\in\overline{1,q},\ n,m,q\in\boldsymbol{N}.$$

Finding similarities between chapter C_{im-1} and chapter C_{im} in the book $\boldsymbol{B_i}$:

$$\cos\left(\varphi_{(C_{im-1},C_{im})}\right)=\frac{\sum_{h=1}^{q}\xi_{\widetilde{t''}_h}^{(C_{im-1})}\cdot\xi_{\widetilde{t''}_h}^{(C_{im})}}{\sqrt{\sum_{h=1}^{q}\left(\xi_{\widetilde{t''}_h}^{(C_{im-1})}\right)^2}\cdot\sqrt{\sum_{h=1}^{q}\left(\xi_{\widetilde{t''}_h}^{(C_{im})}\right)^2}} \quad (16)$$

$$i=\overline{1,n},\ j=\overline{1,m},\ h=\overline{1,q},\ n,m,q\in\boldsymbol{N}.$$

In (15) and (16) *m* is the number of the chapters and *q* – is the number of lemmas (terms). If corpus has large amount of terms and chapters (documents) the task of finding closeness between them becomes more computational and memory resources consuming. To reduce these complexities it is suggested to use LSA approach. This method allows increasing effectiveness of search by reducing computational complexity and filtering less important terms or concepts. LSA consists of four main steps:

1) preparing of the lemma-document matrix;

2) transforming of lemma-document matrix elements from frequencies to TF-IDFs and their normalization;

3) reducing dimensionality by using SVD which key parameter is a rank;

4) retrieving information in reduced by SVD space by finding similarities based on cosine distance between vectors.

C. Singular Value Decomposition application

SVD can be applied to all matrices, not only to square matrices, and it always exists. The essence of the SVD [9] method is decomposition of higher dimensional (*h*x*m*) rectangular matrix $\boldsymbol{M}$ of rank $r\in[0,\min(h,m)]$ into 3 matrices of lower dimensional matrices:

$$\boldsymbol{M}=\boldsymbol{U}\cdot\boldsymbol{S}\cdot\boldsymbol{V}^T=\sum_{i=1}^{r}\sigma_i\boldsymbol{u}_i\boldsymbol{v}_i^T, \quad (17)$$

where:

- an orthogonal matrix $\boldsymbol{U}$ ($\boldsymbol{U}\in\mathbf{R}$(h x h)) with column vectors $\boldsymbol{u}_i$, $i=1,\dots,h$;

- an orthogonal matrix $\boldsymbol{V}$ ($\boldsymbol{V} \in \mathbf{R}$(m x m)) with column vectors $\boldsymbol{v}_j, j = 1, \dots, m$;
- matrix $\boldsymbol{S}$ (m x h) with $\boldsymbol{S}_{ii} = \sigma_i \geq 0$ and $\boldsymbol{S}_{ij} = 0, i \neq j$.

By convention non-zero diagonal elements (singular values) of matrix $\boldsymbol{S}$ are ordered: $\sigma_1 \geq \sigma_2 \geq \sigma_r$. Instead of full SVD factorization low-rank matrices were used. This allows approximating matrix $\boldsymbol{M}$ and reduce amount of computations.

By using SVD matrix $\boldsymbol{M}$ as a sum of low-rank matrices M'_i can be represented:

$$\boldsymbol{M} = \sum_{i=1}^{r} \sigma_i \boldsymbol{M'}_i, \quad (18)$$

where M'_i are weighted by the singular value σ_i: vectors $\boldsymbol{u}_i \cdot \boldsymbol{v}_i^T$ are scaled by corresponding singular value σ_i.

If the lower value k of the rank than r is chosen, than matrix $\widehat{\boldsymbol{M}}$ will be k-ranked approximated and may be represented as follows:

$$\widehat{\boldsymbol{M}}^k = \sum_{i=1}^{k} \sigma_i \boldsymbol{u}_i \boldsymbol{v}_i^T = \sum_{i=1}^{k} \sigma_i \boldsymbol{M}_i. \quad (19)$$

Frobenius norm of matrix $\boldsymbol{M}$ was used to measure errors between $\boldsymbol{M}$ and $\widehat{\boldsymbol{M}}^k$. It is the difference between square root of the sum of squares of the elements of $\boldsymbol{M}$ and $\widehat{\boldsymbol{M}}^k$. Based on (18) calculation of the matrix $\boldsymbol{M}$ approximation error can be written as follows:

$$\boldsymbol{\varepsilon} = \sum_{i=1}^{r} (\sigma_i)^2 - \sum_{j=1}^{k} \left(\sigma_j\right)^2, \; i \in \overline{1,r}, j \in \overline{1,k}, r \geq k. \quad (20)$$

According to Eckart-Young theorem [21] SVD can be used to reduce a matrix $\boldsymbol{M}$ to k-rank $\widehat{\boldsymbol{M}}^k$ in a principled and optimal in the spectral norm sense manner.

One of the objectives of this finding is to obtain optimal value of k, using computational resources and assessing semantic quality of the processing texts. This value was used as an input parameter in this computational experiment. To assess the quality of data processing results was the task of the linguist. Some specific subject area terms were used to verify correctness and effectiveness of the procedure.

III. Text Corpus Data Platform Implementation Details

Text corpus data platform comprises data processing subsystem and module with the user interface for data ingestion, querying, numerical and visual reports building. Text corpus data platform comprises developed software corpus tool and Hadoop software stack based on HDP 3.1.4 Data platform [22]. Corpus tool was developed on Java 8 with Apache Spark 2.3.2.

A. Data Processing Workflow

Text corpus storage contains source and calculated data for building search queries. Text data processing workflow was used [3, 6] to obtain all these data. The effectiveness of these workflows was achieved by means of using efficient computational methods and distributed implementations of these methods on the software level.

SQL database texts were imported from different sources into PostgreSQL. Now these data sources are: books of Bible in SQLite format, UTF-8 encoded text files and PDF-files. Importing was done by using of pgloader tool, specifically developed text parsers and Apache Tika framework. PostgreSQL used as a data source for Apache Spark. Workflow is implemented in Apache Spark which stores results of data processing into RDBMS PostgreSQL 9.6 and into distributed storage HDFS or AWS S3. RDBMS was used for storing structured data which are used for fast queries from user interface of the corpora. Distributed storage (distributed matrices) was applied for storing data which will be then recalculated. Calculated data comprise:

1) syntactically parsed sentences: stored in RDBMS;

2) dictionary of POS-tagged terms with calculated frequencies: stored in RDBMS;

3) dictionary of POS-tagged 2-, 3-, 4-, 5-grams with frequencies: stored in RDBMS;

4) term-document matrix with TF-IDF of each term in each document: stored partially in RDBMS and in distributed storage;

5) N-gram-document matrices with TF-IDF of each N-gram (2-, 3-, 4-, 5-grams) in each document: stored partially in RDBMS and in distributed storage;

*6) results of term-document and N-gram-document matrices SVD decomposition (**U**, **S** and **V** matrices) for making search queries:* stored partially in RDBMS and in distributed storage.

B. Components of text corpus Dataplatform

Big data processing framework Apache Spark 2.3.2 (Apache Spark SQL, Apache Spark MLLib) which contains distributed data structures and distributed computational algorithms implementation was used for corpus data platform building. Distributed data structures were applied as Datasets for data ingestion and as basic NLP processing operations. Distributed matrices were put from MLLib for SVD-decomposition in distributed environment. Apache Spark can transparently work with PostgreSQL data source, distributed file system HDFS and AWS S3 as storages for storing preprocessing data [5].

NLP was implemented by using Stanford Core NLP, LanguageTool and Apache Spark MLLib. Native mathematical libraries [23] were used to achieve higher efficiency in linear algebra calculation (e.g. vectors and matrix processing, SVD).

C. Applied Data Structures

To increase effectiveness of computations Spark distributed data structures: Datasets, DataFrames and RDDs were applied. Main stages of text data processing of Datasets Spark SQL were implemented. Broadcast-function was applied to reduce effect of "data skew" [24] by distributing small portions of data between executors KryoSerializer for Data transmission between executors was applied.

Mathematical calculations and features extraction were accomplished with Spark ML library CountVectorizer/CountVectorizerModel and IDF/IDFModel for frequencies and TF-IDF calculations, NGram for n-gram extraction, SingularValueDecomposition with DistributedMatrix for SVD decomposition in distributed mode were applied. For sentence extraction and tokenization data structures from Stanford Core NLP and LanguageTool – CoreSentence and AnalyzedSentence were used.

IV. COMPUTATIONAL EXPERIMENT

Experiment results were obtained after having developed software execution on the Spark in local and cluster modes. Computational cluster based on HDP 3.1.4 was deployed into private cloud [22]. Source data are shown in the table 1.

TABLE I. INPUT DATA CHARACTERISTICS

Book	Year of publishing	Target translated language	Number of books	Number of Chapters	Number of stories	Number of verses
HCSB	2009	English	66	1189	1139	31102
NIV	2011	English	66	1189	1139	31102
KJVA	2011	English	77	1336	1252	35488
GYZ	2019	Ukrainian	66	1189	1175	31160
OH62	1962	Ukrainian	66	1189	-	31170
NRT	2014	Russian	66	1189	1011	31163

Results of the first stage, which comprises data ingestion, sentences and terms tokenization, syntactical sentence parsing and terms POS-tagging, n-gram extraction are shown in the table 2. Due to a huge amount of time needed for data processing, the task was carried out on the Spark computational cluster: one book processing took about few hours in local mode. For example, on the PC with 8 core CPU i7-8565U, 16 GB of RAM, M2 SSD processing of GYZ took 6 hours 8 minutes and HCSB – 2 hours 3 minutes. Differences between the processing times of English HCSB and Ukrainian GYZ are caused by different implementations of POS-taggers and syntactical analyzers. Stanford Core NLP with neural network pretrained models was applied for the English texts processing, dictionary-based LanguageTool was used for Ukrainian with text processing.

TABLE II. TOKENIZATION AND N-GRAM EXTRACTION RESULTS

Book	Terms with POS	Sent.[b]	N-gram			
			2	*3*	*4*	*5*
HCSB	18151	40714	163066	363079	458298	469681
NIV	17589	39916	160766	361805	457308	469159
KJVA	18695	33787	184470	460385	629910	685215
GYZ	20516[a]	32586	202254	393040	440980	430940
OH62	22122[a]	34724	205196	401351	445917	434053
NRT	19053	40101	188176	358443	394041	376893

a. Large amount of unique terms caused that these terms were not well POS-tagged and lemmatized, due to lack of the POS-tagged dictionaries in the investigated linguistic domain.

b. syntactically parsed sentences.

According to recommendation in [4], i.e., "*A useful rule of thumb is to retain enough singular values to make up 90% of the energy in* ***S***", singular values for k in the range from 1 up to maximum value possible rank for document-lemma matrix were calculated. For example, for HCSB size of matrix $\boldsymbol{M}$ is 1189x17036, so the maximum value which was chosen was 1189. According to (18) all singular values σ, σ^2 (table 3) were calculated and computation time was also measured (fig.1). The estimation of k, σ and energy for HCSB, NIV, KJVA, GYZ, OH62 and NRT are shown in the table 4.

Estimated percent of energy for few of the ranks (or singular values number) for HCSB are also shown in the table 3. As it is shown on the example for HCSB text, $k=640$ allows saving 90% of energy in accordance with the rule mentioned above. But to verify weather this value of k is enough or there should be chosen lower value indicator depends on linguistic specific tasks. Expert analysis is based on the comparison of the search queries results, i.e., if search queries results are the same for different k the smallest value of k are to be chosen. Very close results were obtained for different books in the same language. Queries for different k-values were completed for the most important key words. Expert analysis has shown that to provide sufficient accuracy of query results it is enough to take 70% of energy – query results will be the same. But to obtain the same percentage of energy in the Ukrainian and Russian languages it is needed higher k-values than that of taken for the English language.

TABLE III. HCSBS INGULAR VALUE ESTIMANTION RESULTS

k	σ	σ^2	energy	% of energy
1	559.305954	312823.15	312823.15	3.68309697
5	342.023847	116980.312	946890.838	11.1484421
15	249.005987	62003.9818	1749214.03	20.594783
32	202.231197	40897.4569	2585566.83	30.4417795
58	158.832546	25227.7776	3397625.57	40.0027441
101	127.027755	16136.0505	4256360.01	50.1132561
166	103.515251	10715.4072	5105188.75	60.107141
263	84.7473086	7182.10632	5950186.95	70.0559262
409	68.7590158	4727.80226	6796251.11	80.0172616
640	53.3103309	2841.99138	7645710.78	90.0185749
820	43.9410372	1930.81475	8070542.26	95.0204283
1189	6.52722843	42.604711	8493481.25	100

Dependencies between sizes of SVD rank and calculation times are depicted in figure 1.

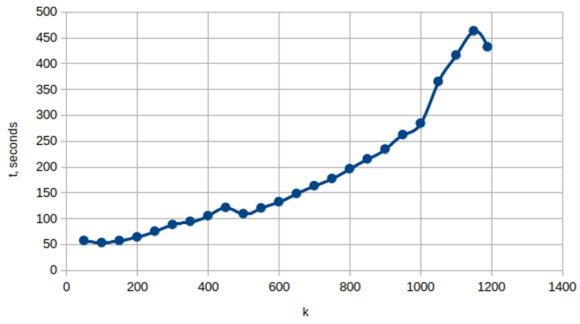


Fig. 1. Dependency between calculation of k and time.

In the same way the results for 2-, 3-, 4- and 5-grams were estimated.

V. CONCLUSIONS

A number of sets and matrices was suggested to be used as a mathematical model of text corpus. Peculiarities of using LSA for text data processing based on the example of historical and religious text corpora have been analyzed. Approaches to calculating time optimization on the level of Apache Spark data structures and computational methods have been offered. Adaptable corpus tool as a software tool for specific text corpus data processing has been developed.

TABLE IV. SINGULAR VALUES ESTIMATION RESULTS FOR DIFFERENT LANGUAGES

% of power	HCSB		NIV		KJVA		GYZ		OH62		NRT	
	k	*σ*	*k*	*σ*	*k*	*σ*	*k*	*σ*	*k*	*σ*	*k*	*σ*
1-5	1	559.306	1	567.252	1	621.033	1	335.590	1	329.412	1	319.345
10	5	342.024	4	351.128	4	387.220	14	104.982	15	103.100	14	110.345
20	15	249.006	15	234.481	14	265.186	53	82.586	54	81.997	49	85.285
30	32	202.231	34	186.762	31	201.597	108	72.130	109	71.772	101	74.081
40	58	158.833	63	151.746	60	156.554	176	65.340	177	64.861	166	66.851
50	101	127.028	107	122.297	105	123.830	258	59.644	260	59.250	247	60.726
60	166	103.515	178	99.063	178	99.516	357	54.517	360	54.120	343	55.169
70	263	84.747	272	82.468	288	81.472	476	49.369	479	49.131	461	49.836
80	409	68.759	418	67.186	453	65.901	624	43.877	627	43.608	608	44.097
90	640	53.310	647	52.288	714	51.128	819	37.160	822	36.921	805	37.054
95	820	43.941	825	43.096	918	42.158	950	32.424	955	32.153	940	32.069
100	1189	6.527	1189	5.787	1336	5.727	1189	5.345	1189	6.914	1189	3.869

Apache Spark computational cluster has been used as a core component of corpus data platform. Computational experiment has been successfully carried out. Text data preprocessing methods of finding optimal rank for SVD-decomposition on the example of specific texts in different languages (English, Ukrainian and Russian languages) were assessed and verified. Optimal rank for SVD-decomposition search queries of sufficient level of accuracy can be built. Optimal SVD-decomposition ranks have been calculated for term-document, lemma-document and n-gram-document matrices.

REFERENCES

[1] T. McEnery and A. Hardieб Corpus linguistics: Method, theory and practice, Cambridge: Cambridge University Press, 2012.

[2] A. Lutskiv, N. Popovych, “Adaptable Text Corpus Development for Specific Linguistic Research”, in: Proceedings of IEEE International Scientific and Practical Conference Problems of Infocommunications. Science and Technology, Kyiv, pp. 217-223, 2019.

[3] D. Jurafsk and J.H. Martin, Speech and Language Processing, 2nd ed. Prentice-Hall, Inc., Upper Saddle River, NJ, USA, 2009.

[4] A. Rajaraman, J. Leskovec and J.D. Ullman, Mining of Massive Datasets [Online]. Available: http://infolab.stanford.edu/~ullman/mmds/book.pdf

[5] A. Lutskiv, N. Popovych, “Big Data Approach to Developing Adaptable Corpus Tools”, Proceedings of the 4th International Conference on Computational Linguistics and Intelligent Systems (COLINS 2020). Volume I: Main Conference, pp.374-395, April 23-24, 2020.

[6] S.T. Dumais, “Latent Semantic Analysis”, Annual Review of Information Science and Technology, Volume38, Issue1, 2004, pp. 188-230, https://doi.org/10.1002/aris.1440380105

[7] T.K. Landauer, D.S. McNamara, S. Dennis and W. Kintsch, Handbook of Latent Semantic Analysis, Routledge Taylor & Francis Group, Lawrence Erlbaum Associates, Inc., 2011.

[8] S. Ryza, U. Laserson, S. Owen, and J. Wills, “Advanced Analytics with Spark. Patterns for Learning from Data at Scale,” 2nd ed., Sebastopol: O’Reilly Media, Inc., pp. 115-136.

[9] D. Kalman, “A Singularly Valuable Decomposition: The SVD of a Matrix”, The College Mathematics Journal, vol. 27, Issue 1, 1996, pp. 2-23. doi: https://doi.org/10.1080/07468342.1996.11973744

[10] M. Deisenroth, A. Faisal and O. Cheng, Mathematics for Machine Learning, Cambridge University Press, 2020. URL: https://mml-book.github.io/

[11] Christian Standard Bible, Holman Bible Publishers, Nashville, 1920, 2011.

[12] Holy Bible. New International Version, Zondervan Publishing House, 1140, 2011.

[13] Holy Bible: King James Version, Study Edition, Containing The Old Testament, Apocrypha, and New Testament. American Bible Society.1462, (2011).

[14] Novitnij pereklad Bibliyi Oleksandra Hyzhi, Druk KT Zabelina-Fil'kovs'ka, Kyiv, 2013.

[15] I. Ohiyenko, “Bibliya Abo Knȳhȳ Svyatoho Pȳs’ma Staroho ĭ Novoho Zapovitu”, Izmovȳ davn’oyevreĭs’koyi ĭ hrets’koyi na ukrayins’ku doslivno nanovo perekladena, UBT. Kyiv, 1962.

[16] Byblyja Novj russkyj sovremennj perevod Slovo Zhyzny. Mezhdunarodnoe Byblejskoe Obshhestvo Biblica, 2014.

[17] Y. Lin, J.-B. Michel, E. Lieberman Aiden, J. Orwant, W. Brockman and S. Petrov, “Syntactic Annotations for the Google Books Ngram Corpus”, proceedings of the 50th Annual Meeting of the Association for Computational Linguistics, Jeju, Republic of Korea, pp. 169–174, , 8-14 July 2012.

[18] V. Lytvyn, V. Vysotska, D. Uhryn, M. Hrendus and O. Naum, “Analysis of statistical methods for stable combinations determination of keywords identification”, Eastern-European Journal of Enterprise Technologies, 2/2(92) 2018, pp. 23-37.

[19] U. Shandruk, “Quantitative Characteristics of Key Words in Texts of Scientific Genre”, Ukrainian Scientific Journal, vol. 2362, 2019, pp. 163-172.

[20] 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”, EasternEuropean Journal of Enterprise Technologies, 2(2-86), 2017, pp. 4-12.

[21] M.M. Lin, “Discrete Eckart-Young Theorem for Integer Matrices”, SIAM Journal on Matrix Analysis and Applications, 32(4), 2011, pp.1367-1382. DOI: 10.1137/10081099X

[22] A. Lutskiv, “Provisioning Hortonworks Data Platform onto OpenStack with Terraform” https://dataengi.com/2018/09/21/terraform-hdp/. Accessed 28 Feb 2020.

[23] Z. Kang, “Using Native Math Libraries to Accelerate Spark Machine Learning Applications”, Cloudera Data Science, 2019. [https://docs.cloudera.com/documentation/guru-howto/data_science/topics/ght_native_math_libs_to_accelerate_spark_ml.html. Accessed 28 Feb 2020].

[24] D. Mykhalyk, “Spark DataSkew Problem”, 2019 [https://dataengi.com/2019/02/06/spark-data-skew-problem

IEEE Third International Conference on Data Stream Mining & Processing
August 21-25, 2020, Lviv, Ukraine

Streaming Process Discovery Method for Semi-Structured Business Processes

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
voytyshyn@gmail.com

***Abstract*—A streaming process discovery method for semi-structured business processes is proposed. The method is further development of the Fuzzy Miner which originally was intended for stationary event data sets. The streaming Fuzzy Miner supports concept drift of a business process model so that recent events are considered more important as the earlier ones. The method was used to represent business process models in the real-time business process monitoring information technology.**

Keywords—process mining, streaming process discovery, concept drift, Fuzzy Miner, RTBPM

I. INTRODUCTION

Today, an average company exploits IT ecosystem with the main purpose to automate the business processes. Usually, execution of such business processes leaves so-called "digital footprint" in databases, log files etc. This "digital footprint" hides the information about how the business processes are performed in reality. And, one of the challenges is to mine valuable insights from such kind of data taking into account its process nature. Occupying the place between classical data mining and business process management (BPM), process mining [1] is the applied discipline that addresses such a challenge.

In practice, complexity of companies' business processes varies in wide range [2,3,4]: from well-structured and predefined to unstructured with a high level of uncertainty. Often, business processes from the 1st category are implemented by means BPM systems, whilst the ones from other category at best allows only partial automation. However, especially interesting are co-called "semi-structured" business processes that lays in-between those two categories. Usually, such processes highly depend on people's decisions and are predictable to some extent; however, they are not structured enough to be prescribed with such languages as BPMN [5].

Current paper is devoted to a streaming process discovery method with the purpose to deal with semi-structure business processes. Fuzzy Miner [6,7,8] was chosen as a basic algorithm. The proposed method was implemented as a part of the real-time business process monitoring (shortly RTBPM) information technology [9,10].

The rest of the paper is organized as follows: the context of the task is outlined in following section; then, the task is stated in a formal manner; the proposed method is represented in section IV; implementation of the method and a calculations example are in sections V and VI respectively; ideas for future steps and conclusions close the paper.

II. CONTEXT

The purpose of the RTBPM information technology is to provide near real-time information of how semi-structured business processes are actually performed. Whilst classical BPM systems controls business processes execution by pre-scriptive models (e.g. BPMN), RTBPM monitors execution of business processes representing the real-life behavior in a descriptive manner. The approach taken by RTBPM works better (comparing to classical BPM) for semi-structured business processes since such processes usually have continuously evolving nature and usually are too complex to be modeled by such strict languages as BPMN.

One of the key features of RTBPM is representation of business process control flow model using event data (i.e. "digital footprint" left by the process in IT systems). As the monitoring is done in the near real-time mode, the process model is generated and updated from a continuous event data stream. In the process mining area, such task is named streaming process discovery [2,11,12,13,14,15].

From the software architecture perspective, RTBPM is based on the Lambda architecture pattern [16] which allows both streaming and batch processing of event data. The event data pipeline (i.e. scalable event data queue) ensures handling the data stream in near real-time mode. As the result, implementation of the proposed streaming process discovery method was integrated in the RTBPM's data pipeline.

III. TASK STATEMENT

The task is to develop a method with the purpose to visualize a business process model from an event data stream. The core of the method is a streaming process discovery algorithm. Similarly to the general process discovery problem [2], the formal task statement is the following.

Let S be a continuous event data stream. A streaming process discovery algorithm D_S is a function that maps S onto a process model G so that the model represents the behavior seen in the already received event data.

The method should fulfill the following criteria: (a) the process model is updated in near real-time mode once event data are received; (b) visual representation of the process model is understandable for the end users who are not experts in business process management or process mining; (c) the

978-1-7281-3214-3/20/$31.00 ©2020 IEEE

visual model supports process simplification so that less frequent behavior can be filtered; (d) the algorithm D_S supports concept drift handling of the process model [17] so that "older" behavior is "smoothly" replaced by "newer" activities.

IV. Streaming Fuzzy Miner

Fuzzy Miner is a process discovery algorithm that meets from the requirements from the prior section [18] (initially, it was designed for semi-structured or even unstructured business processes). However, the original Fuzzy Miner [6,7,8] aims stationary data sets (not data streams). So, this initial version (with some improvements) was integrated as a batch data processing component into RTBPM [19]. The streaming version is proposed in the rest of the current section.

A. Overview of the Algorithm

From the high-level perspective, the Fuzzy Miner algorithm consists of the 3 steps: (a) measuring log-based metrics; (b) measuring derivative metrics; (c) building fuzzy model. In RTBPM, the steps (a) and (b) are performed on the server side whilst step (c) is done on the side of the client web application. From the volume of processing data standpoint, the 1st step is most time and resource consuming since it requires scanning event log (which may be quite big). In turn, steps (b) and (c) do not require scanning of the entire event log and, consequently, are not so resource demanding. Thus, only step (a) requires adaptation to data streams. The goal of such adaptation is arranging calculations in the way that the changes in a business process model can be shown to the end user in near real-time mode (using as minimal computational resources as possible).

B. Handling Case Completion

For the batch Fuzzy Miner, it is assumed that all the cases from the processed event log are completed. However, the streaming version has to deal mainly with incomplete cases. So, it is necessary to define the case completion criterion $R: c \to \{0,1\}$, where $c \in \boldsymbol{C}$, $\boldsymbol{C}$ is a set of all possible cases (completed and incomplete). The characteristic of completion is an attribute of a case: $\#_{completed}(c) \in \{0,1\}$ (the symbolic notation is taken from [2]).

Due to diversity of business processes from different application areas, the criterion R is defined (and configured in RTBPM) for each business process individually. It is worth noting that the case completion check procedure is called from both streaming and batch data processing components of RTBPM.

C. Start and End Events

Originally, fuzzy model did not support the process start and end events (it was assumed that a business process could begin and finish at any of the process steps) [6]. However, from the end user standpoint, it is important to understand what are the entry and exit points. Due to this necessity, the batch [19] and streaming versions of Fuzzy Miner support the start and end events. The start event precedes to each of cases' first event, and the end event follows to each of completed cases' last event (incomplete cases do not have end events). The start and end events are measured by the Fuzzy Miner metrics like any regular event.

D. Dealing with Concept Drift

Over time, the process model may change significantly (such task is usually called concept drift [17]). The streaming Fuzzy Miner has to take this into account process model generation so that the most recent behavior has higher priority than older.

In order to handle process concept drift, cases are grouped by the following criteria:

- Each group include not more than M cases.
- Start time $\#_{start}(c_i)$ of any case from group i is not greater than start time $\#_{start}(c_{i+1})$ of any case from group $i+1$: $\max(\#_{start}(c_i)) \le \min(\#_{start}(c_{i+1}))$, where $c_i \in C_i$, $c_{i+1} \in C_{i+1}$, C_i and C_{i+1} are sets of cases from groups i and $i+1$ respectively.

The defined case grouping approach makes possible log-based metrics pre-aggregation for each group of cases.

E. Measuring Log-based Metrics

First of all, it should be emphasized that log-based metrics calculation approach described below is based on the assumptions that metric values are non-negative.

A log-based metric for the entire event log is calculated as the following:

$$L = \sum_{i=1}^{n} \lambda_i L_i \tag{1}$$

where L is log-based metric value of the entire event log; L_i is a log-based metric value for case group i; λ_i is a weight coefficient that defines importance of cases from group i; n – number of case groups.

The weight coefficients from (1) has to fulfill the conditions: $0 < \lambda_i < \lambda_{i+1} < 1$ for $i = 1..N-2$ and $\lambda_n = 1$. For example, the coefficients can be defined as follows:

$$\lambda_i = \left(\frac{i}{N}\right)^p \tag{2}$$

where $p \ge 0$ (as bigger p as less significant the earlier cases are; when $p = 0$, no concept drift is taken into account).

In turn, a log-based metric L_i from (1) is calculated according to the formula:

$$L_i = L_i^* + \sum_k l_i^k \tag{3}$$

where L_i^* is the metric value for all the completed cases of group i, whereas l_i^k is the metric value for an incomplete case k from group i. Formula (3) emphasizes that the metric values for completed cases are pre-aggregated, whilst the metrics for incomplete cases are recalculated once new events related to corresponding cases are received.

The proposed approach to pre-aggregate log-based metrics is visualized on Fig. 1.

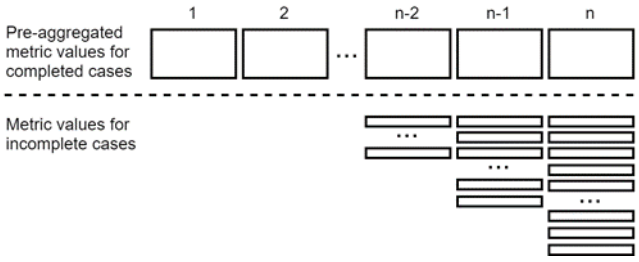


Fig. 1. Log-based metrics pre-aggregation approach

V. Implementation of Streaming Fuzzy Miner

The implementation of the proposed streaming Fuzzy Miner is integrated into the RTBPM event data pipeline that consists of the following stages (Fig. 2): (1) receiving event data and storing them into the database; (2) classifying the events; (3) checking if cases are completed; (4) measuring the log-based metrics. The streaming Fuzzy Miner is executed on stage 4. It takes newly received events, calculates log-based metrics, and saves the calculated values into the database.

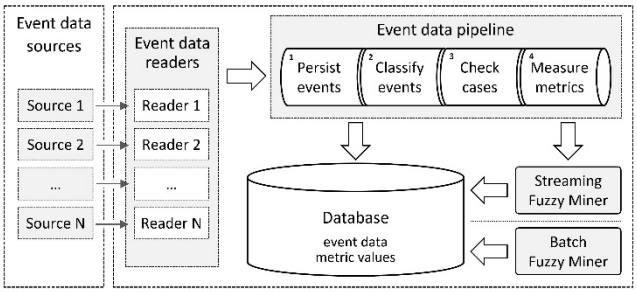


Fig. 2. Fuzzy Miner integration into the RTBPM data pipeline

To deal with concept drift, the following new metrics were added to the existing set of Fuzzy Miner metrics: drift unary significance metric, drift binary significance metric, and drift binary correlation metric. These new metrics were put in-between the log-based and derivate metrics so that the existing metrics' measurement logic remained almost unchanged.

The concept drift handling approach is also implemented for the batch version so that both Fuzzy Miner versions use the unified calculation procedure.

VI. Calculations Example

To demonstrate results of the proposed streaming process discovery method, the "Road Traffic Fine Management Process" data set was used [20].

Process model generated by the previous version of batch Fuzzy Miner [19] without handling of concept drift is depicted on Fig. 3.

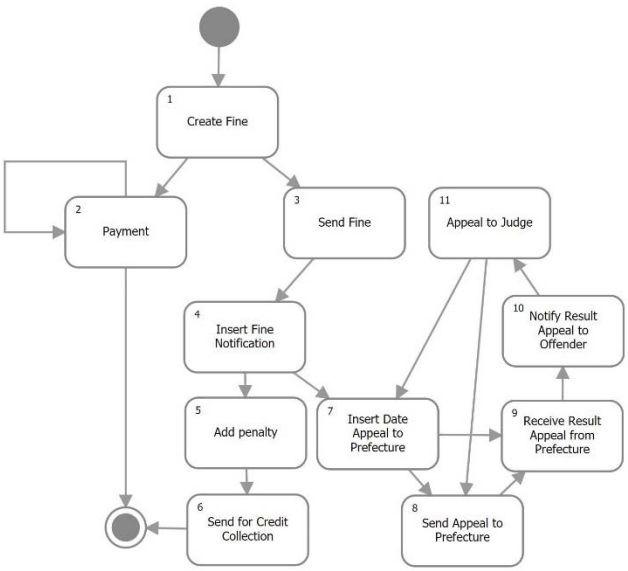


Fig. 3. Process model without considereation of concept drift (batch Fuzzy Miner)

As can be seen from the model, once event 7 is reached, the processes gets into an infinite loop so that there is not possibility to go to the end event. Obviously, the model on Fig. 3 represents incorrect behavior which is not present in the event log. Such situation is caused by the fact that significance of "old" cases is the same as significance of the recent ones (it is worth mentioning that the data set [20] contains events for approximately 13 years). The outcomes of the streaming Fuzzy Miner (with concept drift handling) is shown on Fig. 4. This process model does not contain infinite loops, so the issue from Fig. 3 is resolved.

Both executions of Fuzzy Miner used the same configuration (Table 1).

TABLE I. Fuzzy Miner Configuration

#	Category	Parameter	Value
1	Metrics (applicable for all the metrics)	Active	Yes
2		Weight	1.0
3		Reverce	No
4		Attenuation factor	N-root, n = 2.7
5	Concept drift (applicable for streaming Fuzzy Miner only)	Maximum number of cases in a group	1000
6		Concept drift factor	p = 3
7	Node filter	Significance cut-off	0.0358
8	Edge filter	Best edges / Fuzzy edges	Fuzzy edges
9		Significance / Correlation ratio	0.75
10		Edges cut-off	0.20
11		Ignore self-loops	Yes
12		Interpret absolute	No
13	Concurrency filter	Filter concurrency	Yes
14		Preserve	0.1296
15		Balance	0.70

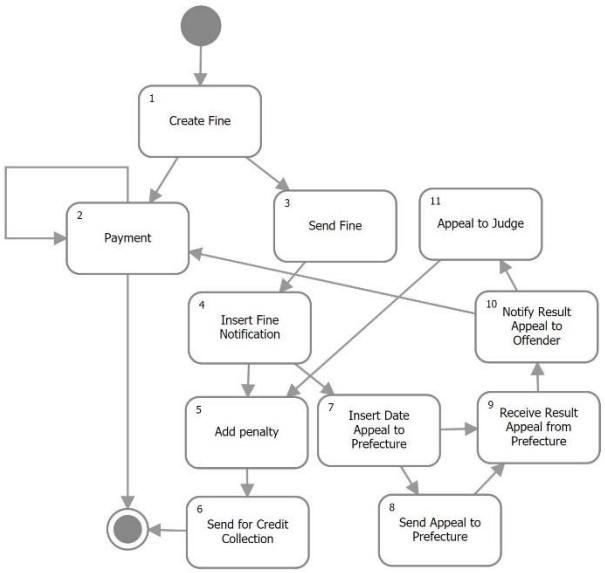


Fig. 4. Process model with consideration of concept drift (streaming Fuzzy Miner)

VII. NEXT STEPS

It is envisioned the following future steps of development of the proposed method:

- Quality measurement of a generated process model so that it can be detected automatically that the model represent behavior which is not present is the event log (like in the example from the previous section).
- Automated selection of the configuration parameters in order to maximize quality of the generated model.
- Automatic detection of concept drift points.
- Decision points detection. Such functionality allows to provide automatic suggestions to the end users based on maximization of valuable process outputs.
- Conversion of fuzzy model into BPMN. This makes possible conformance checking based on comparison with so-called "ideal" BPMN process model.
- Generating a model of collaboration between participants of a business process (so-called social mining task). For example, the approach from [21] can be applied for this purpose.

VIII. CONCLUSIONS

The proposed streaming process discovery method takes advantages from the original Fuzzy Miner (such as ability to deal with semi-structured business processes) and ensures handling of an event data stream so that the process model is updated in near real-time mode with minimal computational resources. Since business processes tend to evolve over time (which is especially relevant for semi-structured processes), the represented method is able to deal with process model concept drift so that older cases are considered as less significant than recent ones.

Implementation of proposed method was done within the scope of the RBPMN information technology [10]. The streaming version was integrated in the event data pipeline. Also, the batch version [19] was updated appropriately so that it supports concept drift handling, as well. It worth mentioning that the represented method (with some minor modifications) was used in the information technology from [22].

From the envisioned future tasks, the automatic quality check of process models is the one of high priority. The purpose of such quality measurement is to automatically detect discrepancies in process models (like the one on Fig. 3). Then, Fuzzy Miner can be automatically configured according to the process model maximization criterion.

REFERENCES

[1] W.M.P. van der Aalst, et al., "Process mining manifesto," in Business Process Management Workshops. BPM 2011 International Workshops. Lecture Notes in Business Information Processing, vol. 99, F. Daniel, K. Barkaoui, S. Dustdar, Eds. Springer, Berlin, Heidelberg, 2012, pp. 169-194. doi: 10.1007/978-3-642-28108-2_19

[2] W.M.P. van der Aalst, Process mining: data science in action, 2nd ed. Berlin: Springer, 2016. doi: 10.1007/978-3-662-49851-4

[3] P. Harmon, Business Process Change: A Guide for Business Mangeres and BPM and Six Sigma Professionals, 2nd ed., Morgan Kaufmann, 2007.

[4] S. Kemsley, "The Changing Nature of Work: From Structured to Unstructured, from Controlled to Social," In: Rinderle-Ma S., Toumani F., Wolf K. (eds) Business Process Management. BPM 2011. Lecture Notes in Computer Science, vol 6896. Springer, Berlin, Heidelberg, 2011. doi: 10.1007/978-3-642-23059-2_2

[5] Business Process Model and Notation (BPMN), OMG BPMN 2.0, 2011.

[6] Ch.W. Günther, "Process Mining in Flexible Environments," Ph.D dissertation, Eindhoven University of Technology, Eindhoven, The Netherlands, 2009. Accessed on: Jan. 29, 2019. [Online]. Available: http://www.processmining.org/blogs/pub2009/process_mining_in_flexible_environments

[7] Ch.W. Günther and W.M.P. van der Aalst, "Fuzzy Mining – Adaptive Process Simplification Based on Multi-perspective Metrics," in Proceedings of the 5th International Conference on Business Process Management. BPM 2007. Lecture Notes in Computer Science, vol. 4714, G. Alonso, P. Dadam, M. Rosemann, Eds. Springer, Berlin, Heidelberg, 2007, pp. 328-343. doi: 10.1007/978-3-540-75183-0_24

[8] "Fuzzy Miner," June 17, 2009. Accessed on: Jan. 29, 2019. [Online]. Available: http://www.processmining.org/online/fuzzyminer

[9] A. Batyuk and V. Voityshyn, "Process Mining: Applied Discipline and Software Implementations," KPI Science News, no. 5, pp. 22-36, 2018. doi: 10.20535/1810-0546.2018.5.146178

[10] A. Batyuk and V. Voityshyn, "Software Architecture Design of the Information Technology for Real-Time Business Process Monitoring," ECONTECHMOD, vol. 7, no. 3, pp. 13-22, 2018.

[11] A. Burattin, A. Sperduti, and W.M.P. van der Aalst, "Heuristics Miners for Streaming Event Data," Computing Research Repository, Dec. 2012. Accessed on: May 07, 2018. [Online]. Available: https://arxiv.org/abs/1212.6383

[12] A. Burattin, Process Mining Techniques in Business Environments. Cham: Springer, 2015. doi: 10.1007/978-3-319-17482-2

[13] A. Burattin, A. Sperduti, and W.M.P. van der Aalst, "Control-flow discovery from event streams," 2014 IEEE Congress on Evolutionary Computation (CEC), Beijing, 2014, pp. 2420-2427. doi: 10.1109/CEC.2014.6900341

[14] S.J. van Zelst, B.F. van Dongen, W.M.P. van der Aalst, "Event stream-based process discovery using abstract representations," Knowledge and Information Systems (KAIS), vol. 54, no. 2, pp. 407–435, Feb 2018. doi: 10.1007/s10115-017-1060-2

[15] J.Oliveira, Business Process Discovery in Real-Time, Técnico Lisboa. Accessed on, Apr. 10, 2020. [Online]. Available: https://fenix.tecnico.ulisboa.pt/downloadFile/395142792129/Extended%2520Abstract.pdf

[16] N. Marz and J. Warren, Big Data: Principles and best practices of scalable realtime data systems 1st ed. Shelter Island, NY: Manning Publications, 2015.

[17] R. P. Jagadeesh Chandra BoseWil, M. P. van der Aalst, and I. Žliobaitė, M. Pechenizkiy, "Handling Concept Drift in Process Mining," in Advanced Information Systems Engineering. CAiSE 2011. Lecture Notes in Computer Science, London, UK, 2011, vol. 6741, pp. 391-405. doi: 10.1007/978-3-64221640-4_30

[18] A. Rozinat, ProM Tips - Which Mining Algorithm Should You Use?, Fluxicon, Oct. 18, 2010. Accessed on: Jan. 29, 2019. [Online]. Available: https://fluxicon.com/blog/2010/10/prom-tips-mining-algorithm/

[19] A. Batyuk and V. Voityshyn, “Distributed software system with web interface for automated business process discovery,” Bulletin of Lviv Polytechnic National University: Information Systems and Networks, vol. 5, pp. 70-77, 2019. doi: 10.23939/sisn2019.01.070

[20] "Road Traffic Fine Management Process", [Online]. Available: https://data.4tu.nl/repository/uuid:270fd440-1057-4fb9-89a9-b699b47990f5. [Accessed: 14 Apr 2018]

[21] O. Mulesa, F. Geche, and A. Batyuk, "Information technology for determining structure of social group based on fuzzy c-means," 2015 Xth International Scientific and Technical Conference "Computer Sciences and Information Technologies" (CSIT), Lviv, 2015, pp. 60-62. doi: 10.1109/STC-CSIT.2015.7325431

[22] 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, 2017, vol. 689. Shakhovska N. et al. Eds. Cham: Springer, 2017, pp. 538-557. doi: 10.1007/978-3-319-70581-1_38

IEEE Third International Conference on Data Stream Mining & Processing
August 21-25, 2020, Lviv, Ukraine

Calibration of Low-Cost IoT Sensors in Streams

Nataliia Manakova
Department of Media System Technology
Kharkiv National University of Radio Electronics,
SoftSetve Inc.
Kharkiv,Ukraine
natalymanakova@gmail.com

Anna Vergeles
Oracle
Kharkiv, Ukraine
annie.vergeles@gmail.com

***Abstract*— This study examines the limitations and requirements of exploiting, as well as the practical aspects of calibrating low-cost sensors in IOT systems. The primary attention is paid to overcoming the features of the streaming nature of data in real systems equipped with processing devices with not high capacity. The leading technological solution is the use of low computational cost mathematical methods of and mini-batch streaming processing. Two-phase approach for (1) initial calibration and (2) monitoring for automatic detection of a deviation (failure/bias/drift) of a sensor (s) were developed and examined on real-life datasets. During the initial calibration phase, the response function is built as a step-wise regression on low-cost sensors stream with corresponding ground-truth sensor stream. On next phase, probability convergence as a control of the correspondence of calibrated values of the particular sensor in adjacent time windows let us identify differences between the expected values from the sensor and the received ones, which will inform the system about the need to calibrate the sensor or about the possible incident. In conclusion, limitations and future work directions are listed.**

Keywords*— *failure detection; sensor calibration; smart farming; smart city; IoT; stream data

I. INTRODUCTION

To improve the quality of life in its various aspects, such as the functionality of cities, public safety, increase the productivity of various industrial and farming enterprises, the deployment of large-scale IoT applications that use sensors and actuators (things) is required. IoT and related technologies, such as artificial intelligence and all aspects of decision making, are essential and depend on the accuracy of the data. With respect to the primary of data quality of data quality, it should be realized that the second critical component of the IoT systems' effective exploitation is their economic viability. Budgetary constraints force a balanced approach to equipping systems, which leads to the widespread use of low-cost sensors. It must be taken into account that such sensors have disappointing characteristics such as frequent failures that can be 10-15% cases as well as susceptibility to bias and/or drift of generated values. Besides, the sensors always have some noise - small random deviations or meaningless background information. These features of low-cost sensors raise concerns regarding the reliability of decisions made on uncertain data. The following are a few examples where the influence of sensors is significant, to name a few:

• In chemical or food production, where the control of the final product composition depends on the readings of sensors, such as pressure and gas mixtures, incorrect readings lead to a decrease in the quality of the product.

• In a smart city, faulty light sensors may provoke an inappropriate on-off mode of street lighting.

• Non-calibrated sensors that measure the concrete humidity can cause problems in building a skyscraper and critical danger of its exploitation.

• An improperly calibrated airport camera may not notice anyone with signs of a contagious disease.

• Refrigerators in retail stores that transmit incorrect temperatures may mean that the drinks are not sufficiently cooled, resulting in lower sales.

Currently, there are many studies in the field of processing and improving the sensory data accuracy through calibration, that are presented in recent review papers [1] and a lot of research papers, some of them described below.

For instance, paper [2] proposes an interesting two-stage calibration approach based on controlled algorithms, including multidimensional linear regression(MLR) in the first stage, and more computationally complex approaches in the second, such as random forests(RF), support vector machines (SVM), and artificial neural network (ANN). In the papers [3] and [4], various algorithms of artificial neural networks (ANN) are studied along with multidimensional linear regression (MLR) to improve the quality of the output data.

In other studies, the need to develop approaches taking into account the specifics of large and streaming data данных [5], including generated by sensors of IoT systems, is emphasized. So, in an investigation [6] using an example of a real sports tracking system, IoT poses new challenges and requires real-time data analysis services that can cope with vast amounts of data.

Despite ongoing studies on automated calibration using sophisticated computational methods, the real situation is that low-cost sensors exploitation still requires significant calibration efforts.

In other words, in order to ensure effective decision-making regarding fully utilizing the low-cost sensors potential, it is necessary to plan on-field calibration that should be performed by staff trained and equipped with high-quality (and high-expensive) or recently calibrated reference sensors.

Manual individual calibration can take up to 30% of the maintenance time of the whole system and make up a significant part of the budget. Nevertheless, the planned

978-1-7281-3214-3/20/$31.00 ©2020 IEEE

manual calibration of inexpensive sensors as a tool of the data accuracy improving cannot be neglected, since training AI on incorrect low-quality data can lead to useless or even erroneous results, which is a much larger and more expensive problem than calibration process. To reduce the cost of regular full manual calibration, the authors propose to perform two-stage calibration: (1) pre-deployed initial calibration by the true ground sensor and (2)selective calibration by real-time monitoring the stream values generated by sensors.

It should be highlighted that an essential challenge in the process of handling data generated by sensor systems is the streaming nature of the data that requires to process the received data in real-time, which imposes restrictions on the calibration time. Besides, it must be taken into account that the devices included in the handling pipeline usually have not either powerful processors or storage devices, so they are not designed to perform large-scale computing operations.

Therefore, the authors of this research focused on developing a practical approach that will allow:

- perform preliminary (initial, pre-deploying) calibration of low-cost sensors, taking into account the streaming nature of the generated data, which means avoiding high computational cost methods,
- carry out real-time monitoring as a means of reducing efforts and time costs for full manual scheduled calibration by conducting selective point calibration of sensors with detected deviations.

Thus, the goal of our work is to develop mathematical tools (approach) for calibration and monitoring as a means of improving the quality of data from sensors and significantly reducing the amount of manual calibration by automatically detecting a non-standard mode (failure/bias/drift) of a sensor(s).

The rest of the article is structured as follows. The prerequisites of the practical approach and the algorithms for its implementation are given in Section 2. Section 3 describes the architecture for implementing the presented approach. Section 4 describes an experiment to verify the effectiveness of the proposed approach, as well as the data sets used to carry it out. Section 5 concludes the work and also includes possible directions for future work.

II. METHOD

There are *n* low-cost high-noise sensors and one expensive but accurate sensor «zero» to be used as an ethalon with ground truth values. Amount of sensors *n* is assumed to be big enough, moreover sensors may be located at significant distances and/or in hard-to-get places, which implies their regular manual calibration to be difficult or even impossible. As it was specified above in description of state-of-the-art approaches in sensor calibration, usage of artificial intelligence techniques such as neural networks or other complexed computational methods may not be applicable to streaming data due to the nature of streaming and low processing power in sensor system. In order to avoid getting over budget during deployment and further exploatation of such a system, the proposed method is focused on a low processing cost. As it will be shown below, the proposed technique allows to reach quality of calibration while taking into account constraints of processing streaming data in real-time.

A. Two-stage calibration

Suggested approach of sensor data processing consists of two stages:

Initial calibration can be performed offline, as well as in real-time, using streaming data that includes telemetry from the high-cost ethalon sensor «zero». This stage is to be performed before system deployment and also in case of any issue that requires sensor replacement or prior to deploying additional set of sensors. All *n* sensors plus zero-sensor are placed together in the same environment and are supposed to have similar values at the same moment of measurement.

Monitoring during operation (in-field calibration) is performed in real-time, using streaming data that comes from sensors installed in their regular location. This is a usual operational mode extended with monitoring to detect either external issues or, in some cases, sensors that require replacement due to malfunctioning.

Both stages work with single data stream that comes from sensors and can be described as: streams from low-cost sensors $s_1, s_2, \ldots, s_n$ for all sensors $i=1,..,n$ to be calibrated, and high-cost ethalon sensor s_0, simultaneously send telemetry data $x_1, x_2, \ldots, x_m^i$ where $x_j \in R$ in random moments $t_1, t_2, \ldots t_m^i$ for all $j=1,..,m^i$ with different length for each sensor:

$$S_{i\,m}{}^{i} = \{(x_1^i, t_1^i), (x_2^i, t_2^i), \ldots, (x_m^i, t_m^i)\},\ i=0..n, \quad (1)$$

that forms mixed queue of events out of all n+1 sensors:

$$S = \{S_{i\,m}{}^{i}\} = \{\{(x_j^i, t_j^i)\}\}\ j=1..m^i;\ i=0..n. \quad (2)$$

Data order and signal periodicity are random, but in a controlled environment it does not take much of an effort to keep periodicity close to certain value with minor deviations. Data can consist of valid values and seldom outliers or missing data.

Main approach here is based on assumption that only sensor itself, not an environment, is a source of value fluctuations. Thus, all measurements within a small time window will show various observations of the same random value. Under this condition, all observed values are supposed to be grouped within a certain range and outliers beyond this interval are a low-probability event.

Pre-processing of the mutisensor stream (2) includes:

(1) Splitting common stream into a keyed stream by sensor id, resulting in a keyed set of separate streams from sensors (1).

(2) Splitting xji for each sensor into batches is performed by non-overlaping windows $T_k = \{T_k^{start}, T_k^{end}\}$ where all $t_j^i \in [T_k^{start}, T_k^{end}]$. Each batch of values is further processed one-by-one in streaming or offline mode.

(3) Missing values in this case can be defined and unequal lengths of $S_{i\,m}{}^{i}$ and $S_{0\,m}{}^{0}$. To deal with missing values it is suggested to run one-by-one pairing sensor values with ethalon values with either dropping all non-paired values for the case when a low-cost sensor is a shorter serie, or applying last observerd value correction if an ethalon, not a low-cost sensor, is a shorter serie out of two.

(4) Since initial calibration mode is a supervised mode, it is unlikely that series will differ in length by more than a couple of observations.

B. Initial calibration

First calibration stage - initial calibration - as mentioned above, is a process when all sensors 1:n and an ethalon sensor 0 are combined at the same location (same envorinment) and are simultateously running. Gathered and processed stream allows to findcalibration coefficient β as a coefficient for response function $Y = f(\beta, X) + \varepsilon$, where Y is set of values form ground-truth sensor 0, X is set of values from each senor under calibration, ε is noise with normal distribution N(0,1). Calibration coefficient for linear response function is a vector $\beta=<\beta 0, \beta 1>$ where vector elements are offset and gain respectively. Further in paper we use an offset only, calculation of gain is the subject of future study. During initial calibration stage it is suggested to keep a calculated γ (gamma) value that shows worst seen deviation of corrected low-cost sensor value from corresponding value of ethalon.

ALGORITHM 1. The pseudo code for initial calibration

INPUT: Accumulated queue S from Si sensors $i=0,..,n$ with time points $j=1,..,m^i$, where S_0 is ground -truth (ethalon)
OUTPUT: Initial calibration reference as $< \beta_i , \gamma_i >$

```
FOR i = 0 to n
    calculate mean for each stream M_i
END
FOR i=1 to n
    calculate offset β_i = M_i — M_0
END
FOR i =1 to n
    FOR j=1 to m^i
        calculate residuals Δ_j^i = S_i m^i - β_i - S_i m^0
        find chosen Xth percentile γ_j^i = Δ_pX^j
        γ_i = max(γ_j^i)
    END
    Store β_i and γ_i
END
```

Fig. 1. Initial calibration algorithm

Observations from sensor corrected with $f(\beta, X)$ are assumed to converge in probability to corresponding observations from ethalon. This means that seldom events when $Y_i - f(\beta, X_i) > \gamma$ are so rare that $P(Y_i - f(\beta, X_i) > \gamma) \rightarrow 0$. Coefficiens β and γ are kept as a reference values in a stateful stream and further are used for stream monitoring (in-field calibration).

Here is the algorithm for simplified β and γ calculation.

1. Accumulate queue from sensors S for N time periods, where each period has enough values for percentile to be applied (see step 5 below). That is near 100 values for 99th percentile to make sense or 20 values for 95th percentile to make sense.

2. Using whole length of accumulated keyed streams calculate mean for each stream $\mathbf{M_i}$, $\boldsymbol{i = 0,..,n}$.

3. For all low-cost sensors calculate offset $\boldsymbol{\beta_i = M_i — M_0}$, $\boldsymbol{i = 1,..,n}$.

4. For all low-cost sensors' stream values calculate residuals $\Delta_j^i = \boldsymbol{S_{i\,m^i} - \beta_i - S_{i\,m^0}}$ while keeping data regarding time period.

5. Within each time period find chosen Xth percentile $\boldsymbol{\gamma_j^i} = \Delta_{pX}^i$.

6. Store $\boldsymbol{\beta_i}$ and $\boldsymbol{\gamma_i} = \max(\boldsymbol{\gamma_j^i})$ as initial state.

The pseudo code for initial calibration algorithm presented in Fig.1.

To measure quality of calibration at this stage one can see the ratio γ_i/M_i. If significance of approximation suits the task and corresponds to domain, the low-cost sensor can be further used «in field».

C. Monitoring during operation (in-field calibration)

Monitoring during operation (in-field calibration) works with a stateful streaming sensor data (2). It undergoes same pre-processing steps plus application of caibration coefficient. Both actions have to be performed in a real-time which leads to using mini-batches, that is windowing function with small enough window length. Exact window range has to be predefined depending on required time to response and can vary from several seconds to several minutes of even hours, however, it should contain enough values to apply percentile function.

ALGORITHM 2. The pseudo code for monitoring sensors stream to choosen in-field calibration

INPUT: Accumulated queue S from Si sensors $i=1,..,n$ with time points $j=1,..,m^i$,
Initial calibration reference as $< \beta_i , \gamma_i >$
Length L of time windows $k = 1,..,m^i/L$
OUTPUT: Set of sensors to be re-calibrated

```
FOR i =1 to n
    FOR j=1 to m^i
        calculate new values Ŝ_j^i = S_i m^i - β_i
    END
END
FOR i =1 to n
    FOR k=1 to m^i/L
        calculate mean of corrected values M_i k
        FOR j=1 to m^i
            Find difference Δ_j^i = Ŝ_j k^i - M^i(k-1).
        END
        find Xth percentile for the k-th window γ_k^i = Δ_pX^i
    END
    If γ_k^i >> γ^i the sensor put in list for re-calibration)
END
```

Fig. 2.Monitoring for in-field calibration algorithm

When in regular working mode there is unexpected difference between same sensor values in consecutive moments of time, it may help detect following situations:

- drift of this sensor
- smooth transition to another mode
- sudden deviation (anomaly)

It is suggested to detect difference using probability convergence approach in order to control calibrated values for the same sensor i in consecutive time windows k, $k+1$. If significant difference is being detected, system will alert about possible incident or suggested of sensor re-calibration.

Below is a phase II (in-field) algorithm, see pseudo code on Fig. 2.

1. For all low-cost sensors" stream values calculate new values $\hat{\boldsymbol{S}}_j^i = \boldsymbol{S_{i\,m^i} - \beta_i}$

2. Calculate mean of corrected values $M_{i\,k}$

3. Find difference $\Delta_j^i = \hat{\boldsymbol{S}}_{j\,k}^i - M^i{}_{(k-1)}$.

4. Find Xth percentile within the k-th window $\boldsymbol{\gamma_k^i} = \Delta_{pX}^i$

5. If $\boldsymbol{\gamma_k^i >> \gamma^i}$ the sensor needs attention (incident or re-calibration).

III. ARCHTECURE

To keep it simple, there are several open-source tools that help process sensor data in real-world.

Those tools were used for experiment.

Apache NiFi helps gathering data from sensors and arranges streams for further processing. Apache Flink performs calculations in a keyed stream (sensor id is a key) while saving status for each batch. In our case status equals to the number of events per sliding window that have exceeded threshold value in order to be or not to be considered a random outlier. High-level deployment diagram suggested for stream processing is shown below on Fig.3.

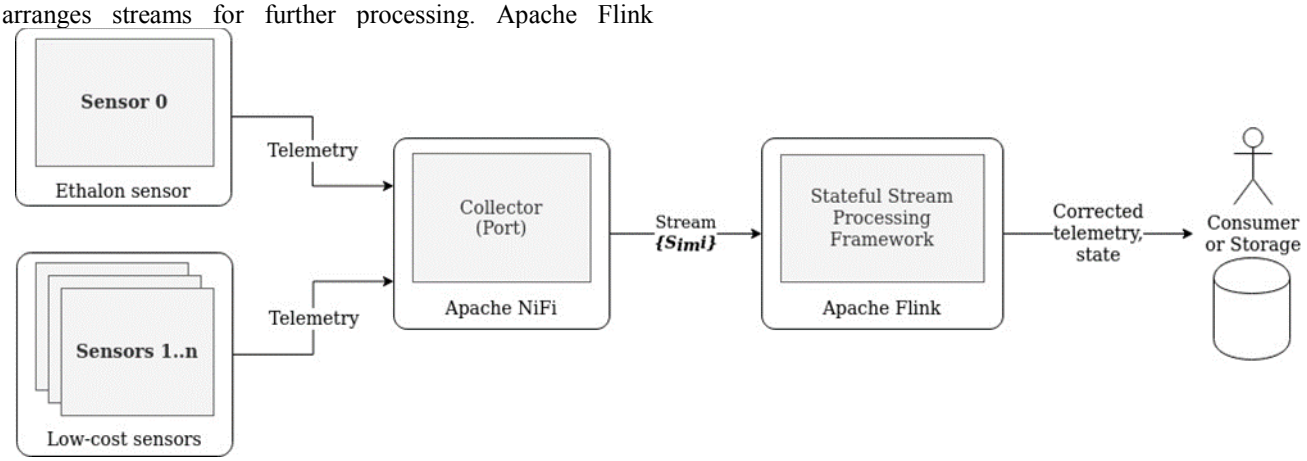


Fig.3. High-level deployment diagram

IV. EXPERIMENTAL STUDY

To conduct an experiment authors have chosen two datasets, the first one is a frequently used dataset from open repository [7], the second one (real-life dataset) has been generated by one of authors.

First dataset [8] which includes 100 recordings of a sensor array under different conditions in a home setting: background, wine and banana presentations. The array includes 8 MOX gas sensors, and humidity and temperature sensors.

Based on the dataset an experimntal stream was emulated using only a part of data from the dataset (Temperature for several subsets of ids).

Figure 4 below shows raw and corrected series. First 1000 values belong to the first calibration phase. Calculated confidence interval is $\gamma = \pm 0.396°C$. Moments when corrected value does not fit into γ interval are so rare that in our concept of convergence in probability such events are not considered as significant but rather as outliers.

On the next picture, figure 5, same series but second phase, situation is different. At some moment there is a shift in signal and at corrected value constantly drifts from previous time window near observation number 2000.

Constant drift cannot be ignored as rare event, thus an alert will be risen to check a corresponding sensor.

The next two types of sensors were used for generating real-life dataset:

Sensors used for calibration: The DHT11 - ultra low-cost digital temperature and humidity sensor wich works in next ranges: 20-80% humidity readings with 5% accuracy, and 0-50°C temperature readings ±2°C accuracy[9].

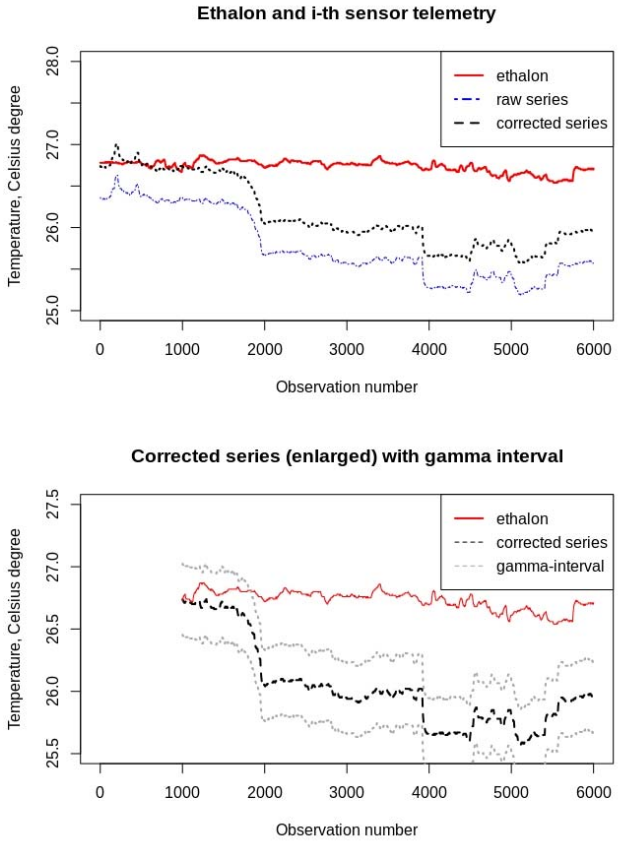


Fig. 4-5. Experimental series

Sensor used as reference (ground truth) device: BME 280 5B I2C - sensor from Bosch is the best low-cost sensing solution for measuring humidity with ±3% accuracy, barometric pressure with ±1 hPa absolute accuraccy, and temperature with ±1.0°C accuracy [10].

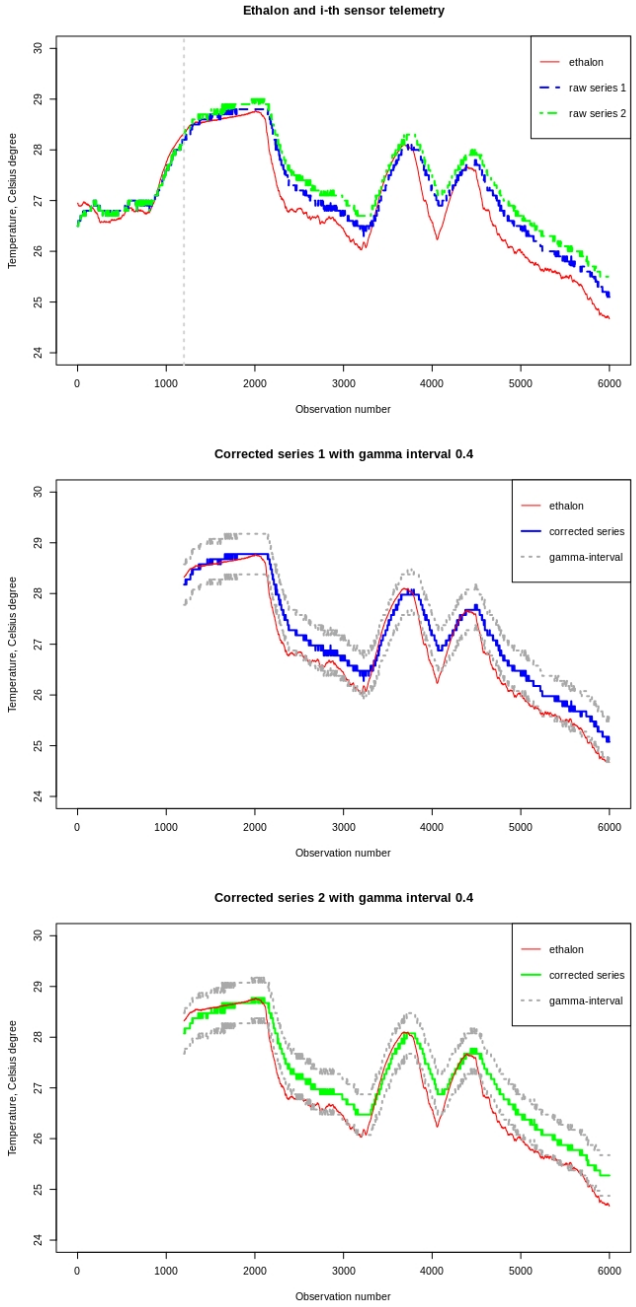


Fig. 6-7-8. Real-life experiment. Situation with rather big value for gamma

During experiment several sensors were arranged in the same location. Figures 6-7-8 and 9-10-11 show result for two low-cost sensors and one ethalon observed at different time.

On the figure 7 and 8 corrected series remained within $\gamma = \pm 0.4°C$ interval if compared to ethalon. On the figure 10-11 $\gamma = \pm 0.14°C$ correspondingly which in both cases is significantly more accurate than ±2°C accuracy by design.

Low efforts to achieve more accuracy is the main idea of the proposed approach.

Adoption of a gamma value, that is too small, may lead to more intensive alerting. Ticks at the bottom on figures 10 and 11 display moments when delta between current value and mean from the previous sliding window exceeded estimated gamma value (phase II algorithm). Moments when amount of events got to be not consecutive are supposed to be considered as moments that require attention due to drift or significant changed in series' behavior.

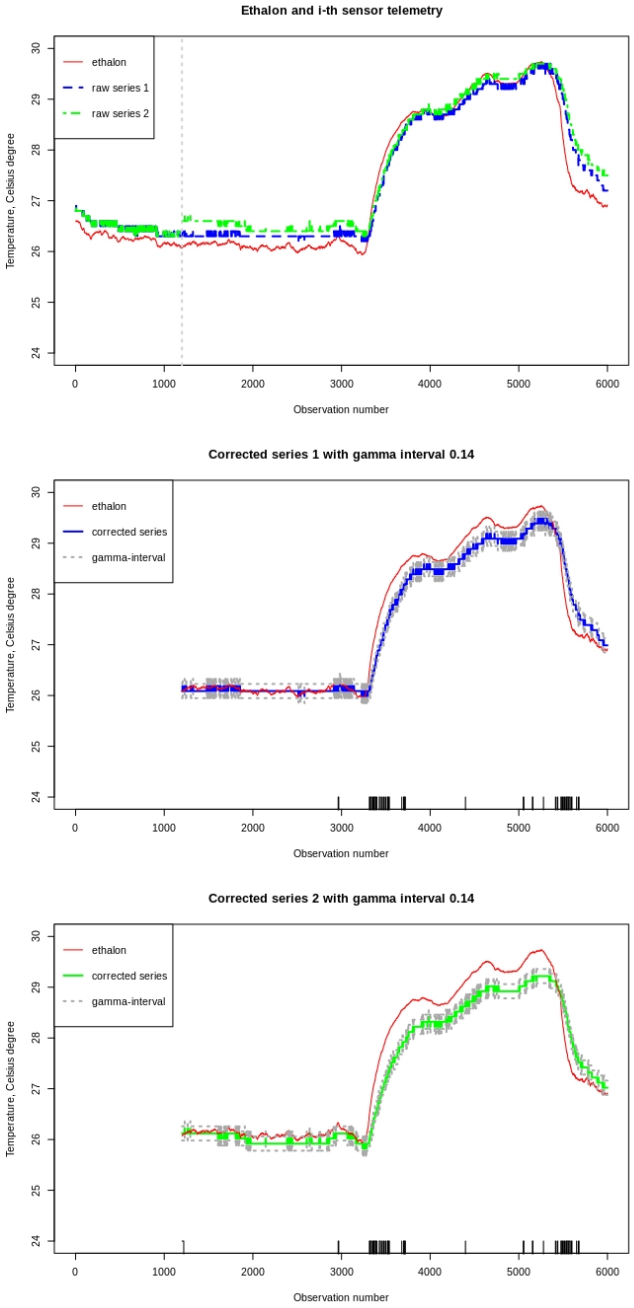


Fig. 9-10-11. Real-life experiment. Situation with rather small value for gamma

V. CONCLUSIONS AND FUTURE WORK DIRECTION

Traditional calibration methods have disadvantages, such as high computational cost and capacity requirements, which do not fit exploited devices and requirements for handling real-time streaming data. To solve these challenges, this study examined ways to improve the quality of streaming data received from sensors, as well as reduce the cost of scheduled if-field calibration.

To deal with the described above requirements, we propose a multi-sensor calibration approach in both of the most applied real-life scenarios of initial calibration as well as monitoring for in-field calibration scheduling, taking into account the features of stream processing. To conduct both experimental and real-life studies, a conceptual architecture with heterogeneous multi-sensor data streams was developed.

It should be noted that this study has some significant limitations of applicability associated with the following.

Firstly, when developing the method, an assumption was made about the presence of a high-precision sensor, which we use as a ground-truth reference.

Furthermore, it is crucial to consider that high likely sensors will behave diversely under substantially dissimilar conditions that can occur in different places. For instance, the sensor is calibrated against the pedestrian historical center of the city and then deployed in heavy traffic or an industrial zone. Here we emphasize that the sensors should be used under conditions similar to the conditions in which they are calibrated in order to maintain appropriate error estimates.

Finally, we considered only one-dimensional calibration, that is, sensors that generate complex data, as is often the case for air analyzers and require multidimensional calibration, are not taken into account.

Thus, the method applicability's essential limitations are ground-truth reference sensor presence, one-dimensionality of the sensors, and the similarity of the calibration and deployment conditions of the sensors.

Overcoming the mentioned limitations applies to work further directions, as well as a further improving effectiveness and flexibility of stream sensor monitoring. This last-mentioned direction of our research interest lies in analyzing the deviation for sensors belonging to the same cluster in a certain sense, such as adjacent sensors or sensors installed near objects of a similar impact.

REFERENCES

[1] J. M. Barcelo-Ordinas, M. Doudou, J. Garcia-Vidal, and N. Badache, "Self-calibration methods for uncontrolled environments in sensor networks: A reference survey," Ad Hoc Networks, vol. 88, pp. 142–159, 2019, doi: 10.1016/j.adhoc.2019.01.008.

[2] J. M. Cordero, R. Borge, and A. Narros, "Using statistical methods to carry out in field calibrations of low cost air quality sensors," Sensors and Actuators B: Chemical, vol. 267, pp. 245–254, 2018, doi: 10.1016/j.snb.2018.04.021.

[3] D. B. Topalović, M. D. Davidović, M. Jovanović, A. Bartonova, Z. Ristovski, and M. Jovašević-Stojanović, "In search of an optimal in-field calibration method of low-cost gas sensors for ambient air pollutants: Comparison of linear, multilinear and artificial neural network approaches," Atmospheric Environment, vol. 213, pp. 640–658, 2019, doi: 10.1016/j.atmosenv.2019.06.028.

[4] R. Wei, K. Ouyang, X. Bao, X. Gao, and C. Chen, "High-precision smart calibration system for temperature sensors," Sensors and Actuators A: Physical, vol. 297, p. 111561, 2019, doi: 10.1016/j.sna.2019.111561.

[5] J. Samosir, M. Indrawan-Santiago, and P. D. Haghighi, "An Evaluation of Data Stream Processing Systems for Data Driven Applications," Procedia Computer Science, vol. 80, pp. 439–449, 2016, doi: 10.1016/j.procs.2016.05.322.

[6] R. Cortés, X. Bonnaire, O. Marin, and P. Sens, "Stream Processing of Healthcare Sensor Data: Studying User Traces to Identify Challenges from a Big Data Perspective," Procedia Computer Science, vol. 52, pp. 1004–1009, 2015, doi: 10.1016/j.procs.2015.05.093.

[7] UCI Machine Learning Repository. [Online]. Available: http://archive.ics.uci.edu/

[8] F. Huerta and R. Huerta, Gas sensors for home activity monitoring Data Set. [Online]. Available: http://archive.ics.uci.edu/ml/datasets/Gas+sensors+for+home+activity+monitoring

[9] DHT11 basic temperature-humidity sensor + extras. [Online]. Available: https://www.adafruit.com/product/386

[10] Adafruit BME280 I2C or SPI Temperature Humidity Pressure Sensor. [Online]. Available: https://www.adafruit.com/product/2652

IEEE Third International Conference on Data Stream Mining & Processing
August 21-25, 2020, Lviv, Ukraine

The Timer Compression of Data and Information

Ruslan Skuratovskii
Department of Informatics and computer science.
NTU "KPI" named by I.Sikorskii and IAPM.
Kiev, Ukraine
0000-0002-5692-6123

Yevgen Osadchyy
National University named by Taras Shevchenko
Kiev, Ukraine
skeleton@univ.kiev.ua

Volodymyr Osadchyy
Faculty of Informatics and Computer Engineering
Igor Sikorsky Kyiv Polytechnic Institute
Kyiv, Ukraine
vo_ua@hotmail.com

***Abstract*— The ability to find short representations, i.e. to compress data, is crucial for many intelligent systems. This paper is devoted to data compression and a transform-based quantitative data compression technique involving quick enumeration in a unary-binary time-based numeral system (NS). The symbols comprising the alphabets of human-computer interaction languages (HCIL), which are used in an informational message (IM), are collected in primary code tables, such as the ASCII table. The statistical-oriented data compression method using unconventional timer encryption and encoding information are proposed by us. It was constructed probability - discrete model of the set of character sequences and characterized some probabilistic algorithms associated with the recovery of text by its public key and its cipher. We find the possibility of parallel implementation of this method by building a block of timer tags. The necessary estimations of complexity are obtained. The method can be used to compress SMS messages. Probabilistic-statistical analysis and evaluation of their effectiveness are obtained.**

Keywords— Block coding, archiving, data compression algorithm, timer coding

I. INTRODUCTION

The main tasks of character-coding are the creation of methods and tools for storage, compression, protection, etc. To date, many theoretical binary coding limits have already been practically achieved, but there are unused opportunities. First of all, they contain the so-called "timer" encoding. It has already been used in a binary Central Committee due to a partial implementation - the introduction of a timer. The aim of the work is to prove that the process of restoring text using a statistical text generator *Gs*, which operates according to a deterministic algorithm, can be unambiguously mapped onto the process of recording timer marks with minimal complexity. Thus, we one-to-one associate a text, or some numerical sequence, for example, a generalized Fibonacci sequence [1], over the alphabet X and its code in the timer coding system, with the alphabet Y, which, in particular, is the alphabet for unit calculus system, which is actually set by timer marks [1,2]. The set of all tuples of lengths n will be denoted by X^n [7]. The effectiveness of statistical analysis with subsequent data compression is proved.

II. REVIEW OF THE RESULTS OF PREDECESSORS

As shown by studies of data compression programs based on classical compression methods [thirteen], their efficiency is already close to the compression limits defined by the classics of coding theory. Using concrete examples, [1, 2, 3, 4, 15] showed that compression after Shannon–Fano methods and their successors [3, 4] repeated lossless encoding compresses by several times the magnitude of a number properly corresponding to the CC of the incoming IM, while lossy encoding yields compression by tens of times. The researchers did not address the efficiency of BD compression by these methods, especially when the IM's primary coding has entropy. But based on our preliminary analysis of classical compression methods' capabilities, the achieved compression/decompression ratio of such IMs will not improve.

A probabilistic-statistical analysis and estimation of the efficiency of a statistically oriented generator is compared with the over-combinatorial generator *Gt* described in [1–3]. Coding with labels on the tape of a machine record was started by A. Turing [3]. To compress and protect information, F. Bardachenko in [4] in the form of timer masquerading. We focused on the results and evidence of the effectiveness of the statistically trained text generator we proposed [2, 4, 8, 10, 13]. In this work we continue the investigation of author [10, 14, 15].

III. TIMER CODING IN LANGUAGE COMMUNICATION

In accordance with the provisions of [7, 8, 13], the quantitative value of a binary number in one bit, displayed through the direct access memory in ASCII, determines the sequence of quantitative values from 1 ... to $256_{(10)}$ records that can be used using 8 bits (cells). A similar number of cells will be needed when using a binary-single midrange $NS_{(2-1)}$. This is shown in the results table. It will be a combination of characters, but from 8 records for 32, the number of one bit memory cells is: 000000000000000000000000000000000000000000000 000000000000000000000101000000000000000000000 000000. The algorithm for generating such a sequence is not binary complexity, but shift "1". The sequence of actions should be increased up to 8 times.

IV. MAIN RESULT

To control a generation of texts it is used regular binary counter. This is a regular binary counter, the frequency of which is synchronized with the reference time, and the readings are displayed in the alphanumeric notation of the language of time intervals. It is used for: timer interruptions, computational paralleling, processing distributions, and so on. Today, for the creation of blockchain, crypto currency, electronic money, internet of things and many other things we

978-1-7281-3214-3/20/$31.00 ©2020 IEEE

are trying to solve the problems of distributed and remote processing of information.

Definition. The text generator statistically oriented (*Gs*) is a generator that implements the necessary series of frequency distribution $\wp$, where $f_i > f_j$, characters with ordering of numbers $n_i < n_j$, where n_i — number of the symbol in the received order $\wp$.

Probabilistic-statistical analysis and estimation of the efficiency of a statistically oriented generator *Gs* are compared with the over-combinatorial generator described in [1,2]. The calculated speed of archiving data is the largest among the world's analogs. The time complexity $O(n^2)$ from which, taking into account the frequency of the generator, it is V easy to obtain a snooze time.

Definition. The text generator with alphabetical ordering (*Gp*) is a generator that sequentially implements texts, generating characters in alphabetical order, that is, not taking into account the frequency of appearance of these symbols in a certain kind of texts.

Theorem 1. There is a bijective mapping, and for the construction of the mapping, no sequential generation of all preceding texts is necessary for the given text, between the set of binary numbers and the set of timestamps of these numbers.

Indeed, the timestamp of a given binary number *a,* and hence of the corresponding text, can be obtained from formula

$$t = \frac{N_2(a)}{V}. \qquad (2)$$

Where $N_2(a)$ is a binary number that corresponds to the code of the given text (the weight of the text for a given symbol generation order); V — frequency of the generator,

$|N_2(a)| = Wh_p(T)$,

where $Wh_p(T)$ is the complexity of the generation (weight) of the text T for a certain linear ordering. For a statistically oriented generator with a symbolic order $\wp$ the value of the timestamp is given by

$$t(\mu_0) = \frac{Wh_s(T)}{V}. \qquad (3)$$

Knowing the weight of the text, you can divide it into k parts. For each part, calculate the weight and calculate the value of the timestamp $t(\mu_i)$ $1 \le i \le k$. Thus, a timestamp system $y = t(\mu_\Sigma)$ by calculating the sum of the series, rather than generating a list of multiple hyperwords in accordance with the lexicographic ordering obtained $\wp$.

Elementary random events will be considered symbols of the text that reach values from the alphabet A, and since the product, sum, the composition of random variables remains a random variable, then we consider a random variable

$$T = \prod_{i_j=1}^{n} \xi_{i_j} = \prod_{j=1}^{n} a_j$$

A random variable is a measurement function χ, which maps from in $\mathbb{R}$. That is, it is necessarily a really significant amount. Therefore, we introduce a one-to-one correspondence of the entire alphabet with a subset of natural numbers $\mathbb{N}$ capacity $|X|$.

Also we can consider the generated text as a random element - a measurable function that acts with Ω in "Any abstract space." Then "random event ξ " is: $\{ \xi \mid \chi(\xi) <$ Const $\}$.

Definition. The number of the symbol $s \in A$ for a given ordering $\wp$ there is his decent number $n_s \in \mathbb{N}$ in this ordering, which is carried out for a statistical generator *Gs* according to the probabilities characteristic of the symbols of the selected text. Note that the weight of the symbol will be its number n_i in the frequency ordering, such, that if $f_i > f_j$, then $n_j > n_i$.

The text symbols have relative frequencies f_s, $s \in A$, where A is alphabet of symbols with T by which we order them by the drop in frequency. Recall that the generator generates symbols in a given descending order sequentially by bit. The time and result of the work is uniquely determined by the timestamp.

Definition.The text generator statistically oriented (*Gs*) is a generator that implements the necessary series of frequency allocations $\wp$, where $f_i > f_j$, characters with ordering of the numbers $n_i < n_j$, where n_i is the symbol number in the received order $\wp$.

Note that in the case of a parallel block generation of a whole text of length n simultaneously, we have the complexity $Wh(T) = \sum_{i=1}^{n} n_i |X|^{n-i}$ and the corresponding value of the **timestamp** calculated using the formula:

$$\varphi(a_5N^5 + a_4N^4 + a_3N^3 + a_2N^2 + a_1N + a_0) =$$
$$= \varphi(c_2N^2 + c_1N^1 + c_0) \circ \varphi(d_2N^2 + d_1N + d_0),$$

$$a_{n+m}N^{n+m} + \ldots + a_1N + a_0 = (c_nN^n + \ldots + c_0)(d_mN^m + \ldots + d_0),$$
$$a_{n+m+1} = c_n,\ a_{n+m} = c_{n-1},\ \ldots,\ a_m = d_m, \ldots, d_1 = a_0.$$

The text symbols f_s, $s \in A$ have relative frequencies, where A is the alphabet of symbols with T by which we order them by the frequency drop. Recall that the generator generates symbols in a given descending order sequentially by bit. The time and result of the work is uniquely determined by the timestamp.

Generating text is as follows: first, it sorts through all the texts of length 1, then length 2, then 3, etc. And when the timer counter works, it stops the work without using the previous texts. Using block coding [3], we generate texts of length n at a time, without wasting time on generating and sorting out the texts of smaller length.

Definition. The timer generator G_T, in the general sense, is an instrument that implements the count (increment) in the selected system of counting (s.ch.) with a constant frequency and sampling. In the quality of system of notation can be a non-positional system of notation, and for a timestamp a unary code can be used, similar to the Rice code, which is based on the non-position unitary calculus system [4].

G_T will be used to stop the generator code text at the right time, when its output will be restored to the initial text.

Let $T=(a_1...a_n)$ is a random text with fixed the distribution of probability $p\left(a_i=s_{i_j}=f_j, j\in(1,...,N)\right), \sum_{j=1}^{N} f_j=1, f_j\geq 0$. Then the conditional mathematical expectation of the complexity of restoring the text is denoted by $E\left(Wh_S(T)\right)$.

Let $ET=\sum_{i=1}^{n} Wh(s_1)P(a_1=s_1)$ is the expected weight of the text from this branch of knowledge.

Theorem 2. The expected complexity of the work of a statistically oriented text generator is less than the expected complexity of the simple test generator found in [2].

We establish the formula of the type of conditional mathematical expectation and try to derive the general formula

$$Wh\left(s_{i_1}\right)=\left(n_{i_1}-1\right)(N)^{n-1}+\sum_{i_2=1}^{N} P\left(a_2=s_{i_2}\right)Wh\left(s_{i_2}\mid a_1=s_{i_1}\right)$$

.

Where

$$Wh\left(s_{i_2}\left|a_1=s_{i_1}\right.\right)=\left(n_{i_2}-1\right)(N)^{n-2}+$$
$$+\sum_{i_3=1}^{N} P\left(a_3=s_{i_3}\right)Wh\left(s_{i_3}\mid a_1=s_{i_1}, a_2=s_{i_2}\right)$$

and n_{i_2} (n_{i_2}) is the number of symbol, standing on the first (second) position in the desired text, which is determined in the order that the text generator has. This is the mathematical expectation of the weight for the variant, when the first character is already set correctly and the generator is looking for the correct variant for the 2nd character. Note that on the right side we have the conditional weight (conditional mathematical expectation) of the *i*-th symbol, which depends on the previous symbols. Here recursively defined functions are used.

$$Wh\left(s_{i_3}\left|a_1=s_{i_1}, a_2=s_{i_2}\right.\right)=\left(n_{i_3}-1\right)(N)^{n-3}+$$
$$+\sum_{i_4=1}^{N} P\left(a_4=s_{i_4}\right)Wh\left(s_{i_4}\mid a_1=s_{i_1}, a_2=s_{i_2}, a_3=s_{i_3}\right) \quad (4)$$

And so on we do recursively transformations for symbols s_{i_4} at known a_3 and the same for the subsequent. The last expression is $Wh\left(s_{i_n}\mid a_1=s_{i_1},...,a_{n-1}=s_{i_{n-1}}\right)$ is a mathematical expectation of the weight of the remainder of the text. It is easy to understand that $Wh\left(s_{i_n}\mid a_1=s_{i_1},...,a_{n-1}=s_{i_{n-1}}\right)=n_{i_n}$, furthermore $Wh\left(s_{i_n}\mid a_1=s_{i_1},...,a_{n-1}=s_{i_{n-1}}\right)\leq n_{i_n}$ since in addition to the serial number, the generator still knows the previous symbols from T, therefore having the frequencies of diagrams and trigrams T the generator will decide on the choice of the correct symbol on the nth position on average before

$$Wh\left(s_{i_{n-1}}\left|a_1=s_{i_1},...,a_{n-2}=s_{i_{n-2}}\right.\right)=\left(n_{i_{n-1}}-1\right)N+$$
$$+\sum_{i_n=1}^{N} P\left(a_n=s_{i_n}\right)n_{i_n}=\left(n_{i_{n-1}}-1\right)N+\sum_{i_n=1}^{N} f_{i_n}n_{i_n}.$$

Moving on in the reverse order, we obtain and collapse the formulas using previous formula:

$$Wh\left(s_{i_{n-2}}\left|a_1=s_{i_1},...,a_{n-3}=s_{i_{n-3}}\right.\right)=\left(n_{i_{n-2}}-1\right)N^2+$$
$$+\sum_{i_{n-1}=1}^{N} P\left(a_{n-1}=s_{i_{n-1}}\right)\times$$
$$\times\left(\left(n_{i_{n-1}}-1\right)N+\sum_{i_n=1}^{N} f_{i_n}n_{i_n}\right)=\left(n_{i_{n-2}}-1\right)N^2+$$
$$+\sum_{i_{n-1}=1}^{N} f_{i_{n-1}}\left(\left(n_{i_{n-1}}-1\right)N+\sum_{i_n=1}^{N} f_{i_n}n_{i_n}\right)=$$
$$=\left(n_{i_{n-2}}-1\right)N^2+N\sum_{i_{n-1}=1}^{N} f_{i_{n-1}}n_{i_{n-1}}-$$
$$N\sum_{i_{n-1}=1}^{N} f_{i_{n-1}}+\sum_{i_{n-1}=1}^{N} f_{i_{n-1}}\left(\sum_{i_n=1}^{N} f_{i_n}n_{i_n}\right)=$$
$$=\left(n_{i_{n-2}}-1\right)N^2+N\sum_{i_{n-1}=1}^{N} f_{i_{n-1}}n_{i_{n-1}}-$$
$$-N+\sum_{i_n=1}^{N} f_{i_n}n_{i_n}=$$
$$=\left(n_{i_{n-2}}-1\right)N^2+N\sum_{i_{n-1}=1}^{N} f_{i_{n-1}}n_{i_{n-1}}+$$
$$+\sum_{i_n=1}^{N} f_{i_n}n_{i_n}-N.$$

And, consistently, we substitute:

$$Wh\left(s_{i_{n-3}}\left|a_1=s_{i_1},...,a_{n-4}=s_{i_{n-4}}\right.\right)=\left(n_{i_{n-k}}-1\right)N^3+$$
$$+\sum_{i=1}^{N}(f_{i_n-2}((n_{i_{n-2}}-1)N^2+N\sum_{i_{n-1}=1}^{N} f_{i_{n-1}}n_{i_{n-1}}+\sum_{i_n=1}^{N} f_i n_{i_n}-N)$$
$$=(n_{i_{n-3}}-1)N^3+N^2\sum_{i=1}^{N} f_{i_{n-2}}(n_{i_{n-2}}-1)+$$
$$+N\sum_{i=1}^{N}(f_{i_{n-2}}\sum_{i=1}^{N} f_i n_{i_{n-1}})+\sum_{i=1}^{N} f_{i_{n-2}}\sum_{i_n=1}^{N} f_i n_{i_n}-N\sum_{i_n=1}^{N} f_{i_{n-2}}=$$
$$=(n_{i_{n-3}}-1)N^3+N^2\sum_{i=1}^{N} f_{i_{n-2}}(n_{i_{n-2}}-1)+$$
$$+N\sum_{i=1}^{N} f_i n_{i_{n-1}})+\sum_{i_n=1}^{N} f_i n_{i_n}-N=$$
$$=(n_{i_{n-3}}-1)N^3+N^2\sum_{i=1}^{N} f_{i_{n-2}}n_{i_{n-2}}+$$
$$+N\sum_{i=1}^{N} f_i n_{i_{n-1}}+\sum_{i_n=1}^{N} f_i n_{i_n}-N-N^2.$$

Let's analyze the compression ratio. Since the value of expression $\sum_{i=1}^{N} f_i n_i$ is minimal when the frequencies $f_1\geq f_2\geq...\geq f_N$ answer ordered in descending weight $n_1\leq n_2\leq...\leq n_N$, then according to Chebishov's inequality [17], this weight is the smallest. It is clear that this value is less than the complexity of the work $Wh(T)=\sum_{i=1}^{n} n_j N^{n-j}$ a conventional serial text generator, which proves the effectiveness of a statistically-oriented generator.

The end of the proof.

We denote by L_{out} this is the sum of the lengths of the codes of words (or blocks of words) in the new alphabet, that

is $L_{out} = \sum_{i=1}^{N} l\left(C\left(w_i\right)\right)$. The sum of the lengths of the text characters in the original alphabet, in the case of block coding not symbolic but will be denoted by $L_{in} = \sum_{i=1}^{N} l\left(w_i\right)$, where $l\left(w_i\right)-$ is the length of the *i-th* block of a text. The formula for the compression index gives the following result

$$k = \frac{\sum_{i=1}^{N} l\left(w_i\right) p_i}{\sum_{i=1}^{N} l\left(C\left(w_i\right)\right) p_i} = \frac{N \cdot n \cdot p_i}{N \cdot p_i} = \frac{n}{1} = n.$$

Thus, the compression index of timer coding is equal to *n*.

V. DATA COMPRESSION THROUGH METHODS OF IMPROVED QUANTITATIVE COMPRESSION IN A UNARY-BINARY NUMERAL SYSTEM (NS)

In order to more fully demonstrate the benefits of advanced quantitative compression methods for BD, we will begin a comparison with classical methods, using the quantities of the simplest integers as examples. There is a bijective mapping which is established in **Theorem 1**, and for the construction of the mapping, no sequential generation of all preceding texts is necessary for the given text, between the set of binary numbers and the set of timestamps of these numbers. Before comparing their quantities, we check the state of memory cells (MC). It is known that memory is formatted before read/write operations are performed on a range of memory. Any kind of memory (M) can be reduced to the simplest form: linear memory. The state of M and its constituent MCs can be defined as existent or non-existent. Programmatically, this looks like a corrupted MC, which is avoided via formatting/not included in M. In a binary DC, it is assumed that a single-bit MC may be in one of two states: "0 / 1". Their combinations give rise to all the quantitative variety of the IM, which can fit into the ranges of its M. However, it is known that an MC's "0" state does not actually have a magnitude, but instead is only the magnitude's indication a number's lower-order digits. Therefore, a MC of binary M potentially contains a superfluous quantity. A question arises. Why store something superfluous in M? Indeed, a single-bit MC potentially open to padding with a number of "0/1"s can always be unambiguously updated with them. Using the reasoning above, we attempt to confirm the need for the existence of $NS_{1\text{-}2}$, and not solely $NS_{2\text{-}1}$ of arithmetic and any other numeral system designed to identify the magnitude of the positional number. We will show its advantage in compressing the magnitude of the simplest numbers in Table 1.

TABLE I. ABSOLUTE VALUATION OF A NUMBER IN BITS (POSITIONS) OF LINEAR M

Size memory in bits M	Absolute valuation of a number		
	$NS_{2\text{-}1}$	**$NS_{1\text{-}2}$**	**NS_{10}**
One MC	0 1	- 1	0 1
Two MCs	00 01 00 01 10 00 01 10 11	10 110 1110	00 00 00 00 01 00 00 00 01 02 00 00 01 02 03

As can be seen from Table 1, as a number's bit-length increases, the compression ratio of the representation of the same quantity of symbols in numbers in $NS_{1\text{-}2}$ changes more quickly than it does for numbers in $NS_{2\text{-}1}$ and NS_{10}. In subsequent iterations of compression, this behavior continues thanks to a more complex implementation of the counting algorithms. Accordingly, we restrict ourselves to the result of representation of the second iteration of compression for $NS_{1\text{-}2}$. This result is presented in Table 2.

TABLE II. THE SECOND ITERATION OF COMPRESSION FOR $NS_{1\text{-}2}$

Bit width in M	**Representation of the magnitude of a number**	**No. in NS_{10}**
	$NS_{1\text{-}2}$	
One MC	- 1	0 1
Two MCs	10 *10 **10	01 02 03

Here, the * symbol denotes a MC that can be filled with any digit/symbol in $NS_{2\text{-}i}$, where $1 \le i \le 2^n$. For clarity, in Figure 1 we present a graphic representation of the relationships identified in Tables 1 and 2.

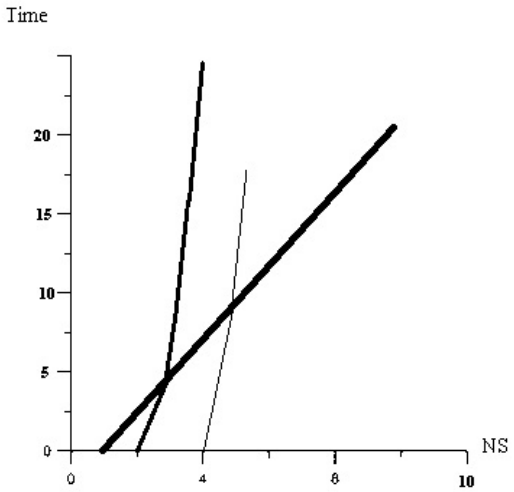


Fig. 1. Efficiency of advanced quantitative compression methods (thick line corresponding to $NS_{1\text{-}2}$) and classical methods for $NS_{2\text{-}i}$, where $1 \le i \le 2^n$. A thin line corresponds to compression using a decimal number system

Compression using a single binary system ($NS_{1\text{-}2}$) is more effective because any number is recorded with a single timer label. Compression performance using a timer label based on $NS_{1\text{-}2}$ is greater since the number of iterations in a single number system and, therefore, speed is 20% higher than that of a binary number system due to the fact that there is no transfer of to the higher order digit. In a binary counter, the time for counting unit shifts, where is the digit in the number. In the 2-number system, there is a transfer of the discharge to the senior digit when a maximal number of bits, i.e. number of the form 11111 is reached.

VI. RESULTS

The ability to compress information to the size of a timer mark has been obtained.

VII. CONCLUSION

For quantification, a data compression index was applied. The presented results of the study are sufficient to understand the proposed compression methods. It should only be noted that the possibility of iterative data compression through a continuous algorithm may lead to compressing data to the size of the TS, which can be a single CC. The time required to restore a compressed IM is a secondary objective. It can be accelerated, for example, using parallel high-speed $NS_{(1-2)}$ counters implemented as processors. The main objective of this publication is to study the possibility of compressing information using an alternate (improved) quantitative method based on a transform-based timer-based coding technique.

REFERENCES

[1] R. Skuratovskii and O. Trembovetska, "Application of discrete structures and numerical sequences in block codes", Naukovie Visti KPI, n.6, 2014, pp. 68-75.

[2] R.V. Skuratovskii, "The method of fast timer encoding of texts". Cybernetics and System Analysis, 49 (1), 2013, pp. 154-161.

[3] D. Lind and B. Marcus, An introduction to symbolic dynamics and coding, Cambridge University Press 1995.

[4] V. F. Bardachenko, "Analysis of the Characteristics of Time-Masked Information", USM, no. 3, 1994, pp.16-29.

[5] N. Koblitz, Algebraic aspects of cryptography, vol. 3, Algorithms and Computation in Mathematics, Springer-Verlag, Berlin, 2004.

[6] A. Danek, "Application of the Burrows-Wheeler Transform for Searching for Approximate Tandem Repeats", Springer Springer-Verlag US, 2012, pp. 256-256.

[7] Y.O. Osadchyy, O.Y. Osadchyy and R.V. Skuratovskii, "Numerical regularities and timer coding information", Artificial intelligence, №3, 2017, pp. 1-22.

[8] V.M. Tereshchenko and Y.O. Osadchyy, "Pransforming technology of coding information in the computer of the von Neumann architecture", Research topics: International scientific youth school "Systems and means of artificial intelligence (AIIS`2017), 2017, pp. 210-214.

[9] A. M. Turing, "On Computable Numbers, with an Application to the Entscheidungsproblem. A Correction", Proceedings of the London Mathematical Society, vol. s2-43, Iss. 6, 1938, pp. 544–546. doi:10.1112/PLMS/S2-43.6.544

[10] R. Skuratovskii, "Parallel solution in fast methods of timer coding of information", High Performance Computing Kyiv, October 22-23, 2018. Source: http://hpc-ua.org/hpc-ua-18/participants/

[11] A.A. Bolotov, S.B. Gashkov, A.B. Frolov and A.A. Chasovsky, An Elementary Introduction to Elliptic Cryptography – CompBook, vol. 2 2006.

[12] Y.I. Volkov and N.M. Voynalovich, Elements of Discrete Mathematics. Kirovograd Central Ukrainian State Pedagogical University, 2000.

[13] Y.O. Osadchyy, "An Approach to Improvement of Timer Methods of Information Security", Vis. TANG, Economic and mathematical modeling, № 1, 1999, pp. 30–35.

[14] R. Skuratovskii, "The Derived Subgroups of Sylow 2-Subgroups of the Alternating Group and Commutator Width of Wreath Product of Groups", Mathematics, Basel, Switzerland, № 8(4), 2020, pp. 1-19.

[15] R.V. Skuratovskii and V. Osadchyy, "Order of Edwards and Elliptic Curves Over Finite Field", WSEAS Transactions on Mathematics, vol. 19, 2020, pp. 253-264.

[16] Shannon-Fano algorithm, URL: [https://studfiles.net/preview/5368369/page:4/].

[17] V.A. Gnatyuk, "Mechanism of laser damage of transparent semiconductors", Physica B: Condensed Matter, 2001, pp. 308-310.

IEEE Third International Conference on Data Stream Mining & Processing
August 21-25, 2020, Lviv, Ukraine

Procedure for Determination of Professional Competence of a Higher Education Institution Graduate

Anatolii Shtymak
Department of Systems Analysis and Optimization Theory
Uzhhorod National University
Uzhhorod, Ukraine
ORCID 0000-0003-4602-5353

Pavlo Mulesa
Department of Cybernetics and Applied Mathematics
Uzhhorod National University
Uzhhorod, Ukraine
ORCID 0000-0002-3437-8082

Mykola Malyar
Department of Cybernetics and Applied Mathematics
Uzhhorod National University
Uzhhorod, Ukraine
ORCID 0000-0002-2544-1959

***Abstract*—The formation of professional competence of a higher education institution graduate is ensured by mastering various educational components (disciplines, pieces of training, qualification work, etc.), which are combined into cycles of preparation of an educational program. This paper proposes a procedure for determining the level of professional competence of a graduate based on assessments obtained during training in a higher education institution, consisting of the following stages: determining the level of proficiency by training cycles and determining the level of competence of the graduate as a whole. Methods in this area allow to obtain a quantitative assessment of decisions made through their qualitative descriptions. The structure of the fuzzy system is described based on the dependence of the fuzzy output value (the level of competence of the graduate) on the fuzzy input values (estimates).**

Keywords—competencies, competency, competency-based approach, fuzzy set, linguistic variable, fuzzification, aggregation, defuzzification.

I. INTRODUCTION

The dynamic changes taking place in the modern world determine specific tasks of the system of assessment of educational achievements of future specialists at each stage of training. They are determining the level of competence of future specialists who study at higher education institutions, the level of ability of professionals to apply their knowledge throughout life. One of the most urgent tasks in the field of education is the problem of assessing and monitoring the competence of participants in the educational process in higher education institutions, which is associated with the development of adequate measurement methods and algorithms for competence analysis.

Traditionally, the goals of higher education were determined by the set of knowledge, skills, and abilities that a graduate must possess. Today, such an approach has proved insufficient. Society needs graduates who are ready to participate in further professional activities, able to solve the life and professional problems set before them practically. This fact largely depends not only on the acquired knowledge, skills, and abilities but on some additional qualities to define which the concepts of "competencies" and "competence" are used, which today are more in line with the understanding of modern educational goals. [1] Competences are personal characteristics, the ability to perform certain functions, types of behavior, and social roles. Under competence, we mean integral attributes of the individual, which characterize the willingness to solve problems arising in the process of life and professional activity, using knowledge, experience, individual abilities. Competence is not reduced to a set of competencies and a sum of knowledge, skills, and abilities. It also includes motivational, social, and behavioral components. It characterizes the integrated qualities of graduates of higher education institutions, namely the result of learning. The approach that considers the graduate with competencies, namely what he can do, what ways of activity he has mastered, what he is ready for - is called the competency approach. [2]

II. ANALYSIS OF PREVIOUS RESEARCH

Competence is regarded as a professional literacy, the degree of qualification of a specialist, as the level of development of personality and human culture. The lack of a unified approach leads to ambiguous interpretation and creates difficulties in classification. In the educational process, students form and develop professional competencies that determine readiness for professional activity. The most general classification includes three major classes of competencies:

subject (special) competencies, essential for the - implementation of professional tasks;

super professional (sometimes called basic) competencies needed to work effectively in the organization;

key competencies that determine the successful socialization of each graduate.

There are no contradictions between these classes. Furthermore, they should be considered complementary, allowing to detail educational tasks. [3]

In practice, there are two types of learning outcomes of students of higher education: general competences (interchangeable skills) and competences that correspond to the subject (theoretical, practical, knowledge, and skills relevant to the subject). Therefore, there is a need to evaluate these results according to the applicable criteria (parameters). [4] The level of competence of the graduate depends on several interrelated factors. These factors include knowledge of the theoretical foundations of the disciplines being studied, the ability to apply the acquired knowledge in practice, the ability to self-education, personal characteristics, etc.

An insufficient level of development of any factor influencing the formation of competencies ultimately reduces

978-1-7281-3214-3/20/$31.00 ©2020 IEEE

the level of competence of the specialist. Therefore, competencies should be assessed comprehensively according to several criteria, i.e., the task is multi-criteria. When solving such problems, different methods of convolution of criteria into one generalized (integral) criterion are often used.

III. The Purpose of the Article

The purpose of this article is to describe approaches to determine the level of competence of a graduate of a higher education institution based on assessments obtained by students for different types of work in the learning process, using fuzzy set theory and information processing methods in the vaguely given information.

IV. The Main Part

A. Determining the Level of Competence of the Graduate

The integrated nature of the graduate's competence requires the development of a holistic system of measuring instruments. Among the methods of measuring and assessing the components of competence can be divided into two major groups: test technologies and methods of expert assessment. At this stage, the field of applied mathematics is actively developing fuzzy modeling, which allows you to more adequately model various objects and processes of educational activities, taking into account the human factor. This fact explains the fact that fuzzy modeling can give results more productive and useful in education than the results of explicit modeling.

The implementation of fuzzy modeling methods is based on the theory of fuzzy sets and fuzzy logic based on it, which allows taking into account the nature of human thinking to a greater extent. This explains the fact that an effective competency analysis requires an approach in which accuracy, rigor, and mathematical formalism are not necessary, and a vague methodological scheme is used. The peculiarity of this approach is that it uses linguistic variables, instead of numerical variables, or in addition to them. Simple relationships between variables are described using fuzzy expressions, and fuzzy algorithms define complex relationships. Methods in this area allow obtaining a quantitative assessment of decisions made through their qualitative descriptions. [5]

Traditionally, the quality of vocational education is judged by the assessments of students' performance, which they received as a result of their control procedures: tests and exams. Each teacher is responsible only for the content of his subject, the organization of its assimilation and academic control. The main forms of organization of the process of its assimilation are also artificial, which are specially invented as forms of the educational process and outside it is rare or almost non-existent. The introduction of the competency approach not only changes the results and goals of education, criteria, and procedures for diagnosing the level of real achievements. Besides, it also changes the type of learning along with other, adequate goals, criteria and procedures content, forms, methods, tools, organization of the educational environment, and activities for those who teach and learn.

The formation of professional competence of university students is influenced by various disciplines, which are combined into training cycles. Besides, its formation is influenced by such factors as participation in their own projects, scientific conferences and seminars, subject competitions, etc. [6]

To determine the level of competence of the graduate, consider the procedure of assessment of competence, which consists of several stages: determining the level of competence by training cycles and the level of competence of the graduate of higher education.

Thus, to solve the problem of determining the level of competence of a graduate of a higher education institution, we consider the structure of a fuzzy system based on the dependence of the initial value (level of competence) on several input values (grades). The functionality of such a system is determined by the following steps:

- fuzzification: conversion of clear input variables (estimates) into fuzzy values, i.e. determining the degree of correspondence of inputs to each of the fuzzy sets;
- aggregation of fuzzy inputs in initial values according to training cycles;
- aggregation of fuzzy inputs in the general initial value;
- defuzzification: conversion of a fuzzy initial value into a clear one.

At the first stage, we determine the level of competence and the degree of belonging to the corresponding fuzzy set, formed as a result of studying the disciplines of each cycle of training according to the curriculum: a cycle of social sciences, a cycle of fundamental disciplines, cycles of general and practical training, obligatory and optional courses. The level of competence directly depends on the grades that the student got as a result of studying the disciplines of the cycle, as well as on the number of credits allocated for the study of these disciplines. In the future, all estimates are considered as fuzzy linguistic variables, with corresponding term sets. Term sets are the levels of formation of the graduate's competence. [7]

In the second stage, using the levels of competence obtained for each cycle in the first stage, with the help of aggregate convolution, we obtain an integrated assessment of the competence of a graduate of a higher education institution. The weighting of the convolution, in this case, is chosen depending on the total number of credits provided for the study of each of the training cycles: the higher the number of credits, the greater the value of the weighting factor.

To form the competence of the graduate, we use the term set of the linguistic variable through the following levels:

- low (receptive-productive) level;
- average (reproductive) level;
- sufficient (constructive-variable) level;
- high (creative) level.

First, we determine the level of competence of the graduate of the institution of higher education in each of the cycles of training, using the assessment of disciplines and the table of correspondence of the results of control on the ECTS scale.

B. Fuzzification of Inputs

Fuzzification is the conversion of clear input values (in this case - the value of estimates on the ECTS scale) to fuzzy sets. [8]

We consider a fuzzy linguistic variable $\widetilde{X}$ – "level of competence", which is defined by a set of four linguistic terms K_1 – "law level", K_2 – "average level", K_3 – "sufficient level", K_4 – "high level", that is $\widetilde{X} = \{K_i \mid i = 1,2,3,4\}$. Each of the terms is a fuzzy set

$$K_i = \{(x, \mu_{K_i}(x)) \mid x \in X, 0 \le \mu_{K_i}(x) \le 1, i = 1,2,3,4\}.$$

Belonging functions $\mu_{K_i}(x)$, $i = 1,2,3,4$ of a linguistic variable $\widetilde{X}$ define in space $X \in [1..100]$ using a scheme:

$$\mu_{K_i}(x) = \frac{\text{point} - (\text{lower limit of the range} - 1)}{\text{gradation range}}. \qquad (1)$$

Therefore, when fuzzifying a clear input (evaluation) x_j, $j = 1,2,...,s_t$, where s_t – the number of assessments of the relevant training cycle, determines the degree of its belonging to one of the four linguistic terms K_i, with the corresponding membership functions $\mu_{K_i}(x_j)$, $i = 1,2,3,4$, $j = 1,2,...,s_t$. These degrees are the values of the membership function $\mu_{K_i}(x)$ at the point $x = x_j$, $j = 1,2,...,s_t$.

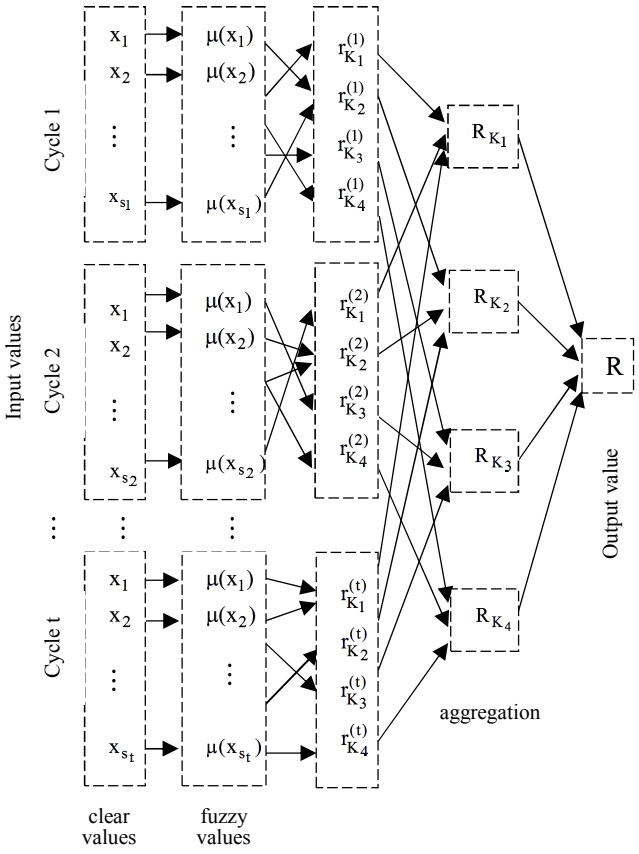


Fig. 1. The scheme of assessment of the level of competence of the graduate.

Thus,

– linguistic term K_1 – "low level", the meaning of which belongs to the range $[1..59]$ corresponds to the following membership function

$$\mu_{K_1}(x) = \frac{\text{point}}{59}; \qquad (2)$$

– linguistic term K_2 – "average level", the meaning of which belongs to the range $[60..73]$ corresponds to the following membership function

$$\mu_{K_2}(x) = \frac{\text{point} - 59}{14}; \qquad (3)$$

– linguistic term K_3 – "sufficient level", the meaning of which belongs to the range $[74..89]$ corresponds to the following membership function

$$\mu_{K_3}(x) = \frac{\text{point} - 73}{16}; \qquad (4)$$

– linguistic term K_4 – "high level", the meaning of which belongs to the range $[90..100]$ corresponds to the following membership function

$$\mu_{K_4}(x) = \frac{\text{point} - 89}{11}. \qquad (5)$$

The scheme of fuzzy assessment of the level of competence of the applicant for higher education can be implemented as shown in Fig.1.

C. Aggregation

Let's consider m - cycle of training and disciplines included in it. Let for each j - discipline a student gets x_j points on a 100-point scale. Thus, the student's results for this cycle can be represented as a vector $(x_1, x_2, ..., x_{s_m})$, where s_m – the number of estimates of the cycle. Replace the grades from each discipline with fuzzy values with membership functions according to (2) - (5) and determine the degree of belonging to one of the four linguistic terms that corresponds a certain level of competence of the graduate on the considered m - cycle of training using the formula

$$r_{K_i}^{(m)} = \sum_{j \in J_{K_i}^{(m)}} p_{ij}^{(m)} \mu_{K_i}(x_j), \; i = 1,2,3,4, \qquad (6)$$

where $J_{K_i}^{(m)}$ – the total of points that refers to i - linguistic term m - cycle, and $p_{ij}^{(m)}$ – is a weighting factor j - discipline i - linguistic term in m - cycle, which is selected depending on the number of credits allocated for study j - discipline in the considered cycle. In addition, the weights will be chosen so that $\sum_{j=1}^{s_m} p_{ji}^{(m)} = 1$, $i = 1,2,3,4$.

A similar approach is applied to all training cycles provided by the curriculum. As a result of using the convolution (6), for each m - cycle we obtain fuzzy numerical values that characterize degree of belonging to one of the four linguistic terms, i.e. vector

$(r_{K_1}^{(m)}, r_{K_2}^{(m)}, r_{K_3}^{(m)}, r_{K_4}^{(m)})$, where $r_{K_i}^{(m)}$, $m = 1,2,...,t$, $i = 1,2,3,4$ – weighted fuzzy corresponding value i - level of competence of the graduate on m - cycle of preparation.

After calculating the fuzzy values that indicate the degree of belonging to one of the four linguistic terms for each of the cycles, we move on to the second stage: determining fuzzy numerical values that characterize the degree of belonging to the relevant term sets and correspond to a certain level of competence of a graduate as a whole, ie find the vector $(R_{K_1}, R_{K_2}, R_{K_3}, R_{K_4})$, where R_{K_i}, $i = 1,2,3,4$ – weighted fuzzy value. The values of these values are calculated by the formula

$$R_{K_i} = \sum_{m=1}^{t} d_m r_{K_i}^{(m)}, i = 1,2,3,4 \ , \qquad (7)$$

where d_m – weighting factor m - cycle of training, which is determined depending on the number of credits allocated for the study of this cycle, and $\sum_{m=1}^{t} d_m = 1$.

Other approaches can be used to generate fuzzy inputs. [9].

D. Defuzzification

Defuzzification of a fuzzy initial value to a clear one can be performed by one of the known methods (centroid method, first maximum, altitude defuzzification, etc.). Depending on the chosen method, we will have one or another clear value of the desired value. [10].

V. CONCLUSION

Thus, the use of the described approach makes it possible, based on the assessments obtained for each discipline in the learning process, to monitor the process of preparation of a graduate of a higher education institution. As well as to monitor the implementation of the competence model of the specialist; move to determine the level of competence of a graduate of a higher education institution. Which in turn, will ensure the creation of qualitatively new standards of higher education, expand the field of the future professional activity of a graduate and ensure its competitiveness in the labor market, and enable employers to choose the appropriate level of professional competence.

REFERENCES

[1] I. A. Zimniaya, "Key competencies – a new paradigm of education result", Higher education, no. 5, 2003, pp. 34-42.

[2] K. S. Vasylyeva and O. A. Shherbyna, "Evaluation of students in a competency approach", Management of complex systems, vol. 16, 2013, pp. 159-163.

[3] Y. H. Tatur, "Competence in the model structure of the quality of training specialists", Higher education today, no. 3, 2004.

[4] G. A. Larionova, Competence in the training of university students, Chelyabinsk: ChGAU, 2004.

[5] M. M. Malyar and A. Yu. Shtymak, "Model of formation of rating of knowledge of a graduate", "INFOTEH-2013" – IT and information security in science, technology and education: Proceedings of the International Scientific and Practical Conference, Sevastopol: SevNTU, 2013, pp. 99-100.

[6] M. M. Malyar and A. Yu. Shtymak, "Block diagram of the evaluation of the student's competence", "INFOTEH-2011" – IT and information security in science, technology and education: Proceedings of the International Scientific and Practical Conference, Sevastopol: SevNTU, 2011, p. 216.

[7] L. A. Zadeh, Concept of linguistic variable and its application to the adoption of approximate solutions, Moskow: Mir, 1976.

[8] N. N. Malyar and A. Yu. Shtymak, "Modeling of the rating of knowledge of graduates", Problem of computer Intelectualization, 2012, pp. 340-344.

[9] O. M. Poleschuk, "On the application of fuzzy sets in problems of constructing leveled gradations", Lesnoy vestnik, no. 4(13), 2000, pp. 143-146.

[10] Yu. P. Zaychenko, Fuzzy models and methods in intelligent system. Kyiv: Slovo, 2008

Author's Index

Chief Editors

Olena Vynokurova, Dmytro Peleshko

Editorial Board

N. Tsyura, P. Zhernova

August 21-25, 2020

Lviv, Ukraine

DSMP