Sergii Babichev · Dmytro Peleshko ·
Olena Vynokurova (Eds.)

Data Stream Mining & Processing

Third International Conference, DSMP 2020
Lviv, Ukraine, August 21–25, 2020
Proceedings

Springer

Contents

Hybrid Systems of Computational Intelligence

Hybrid Deep Convolutional Neural Network with Multimodal Fusion

Olena Vynokurova[1,2](✉), Dmytro Peleshko[1], and Marta Peleshko[3]

[1] GeoGuard, Kharkiv, Ukraine
vynokurova@gmail.com, dpeleshko@gmail.com
[2] Kharkiv National University of Radio Electronics, Kharkiv, Ukraine
[3] Lviv State University of Life Safety, Lviv, Ukraine
marta.peleshko@gmail.com

Abstract. The Hybrid Deep Convolutional Neural Network with Multimodal Fusion (HDCNNMF) topology for the multimodal recognition of the speech, the face, the lips, and human gestures behavior is proposed. 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. Also, the method for forming the audio modality feature is proposed. This method is based on eigenvectors of Mel frequency cepstral coefficients (MFCC) and Mel frequency energy coefficients (MFEC) self-similarity matrix and allows increasing informativeness of modality feature. The specific characteristics of proposed fusion operation is that the data of the same dimension without regard to the modality type are fed to the input of fusion subsystem. During the experiments, the high recognition efficiency was obtained both in cases of individual modalities and their fusion. The distinctive feature of proposed HDCNNMF topology is that the input set can be extended by new modalities types. This extension of modalities set should improve the quality of identification, segmentation or recognition in complex ambiguous visual scenes and simplify the task of affordance detection.

Keywords: Multimodal fusion · Hybrid system · Deep learning · Video stream · Recognition

1 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

S. Babichev et al. (Eds.): DSMP 2020, CCIS 1158, pp. 62–78, 2020.
https://doi.org/10.1007/978-3-030-61656-4_4

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, HDCNNMF 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 HDCNNMF 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 HDCNNMF 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.

2 Related Works

The mainstream of understanding the dynamic scenes is the answer to the question "know what can be done with such objects?" [11,12]. That is, the category of the object is determined by the action, and not the visual appearance [4]. This means that semantic segmentation or recognition is a preprocessing stage in affordance detection. In the case of simple scenes, when the illumination is unchanged, the scale does not change, all projections of attention objects are separated, the pointing gestures is unique and the other, segmentation and recognition are high quality and affordance detection is quite successful [20]. In the case of complex scenes when the pointing gestures is ambiguous, projections of attention objects are not separated, the subjects of the scene are weakly identified and others, the quality of objects (processes) segmentation or recognition is weak. Accordingly, the definition of affordance detection is complicated or impossible at all.

In such cases, the segmentation and recognition systems should be a multimodal using the fusion operation of the chosen type. This approach is due to three reasons. The first of these is to improve the solution quality of the segmentation, identification or recognition tasks. An additional modality has additional useful informativeness, which can increase the accuracy of classification tasks.

The second reason - the lack of a signal in one of the modalities can be neglected by the presence of a signal in other modalities. An appropriate classification problem with sufficient accuracy will be solved.

The third reason - when new tasks related to the scene understanding or the appearance of new scenes, there is a need to expand the set with new modalities. It should be noted here that the existing most used fusion approaches require re-training throughout the system. In the vast majority of modern researches, only two modalities have been considered. These modalities are obtained from audio and video streams. Sometimes three type of modalities are considered: audio, video, NLP [17].

Processing audio modality is a speech recognition. 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 recognition. Among the results obtained in the field of recognition of gestures, it is should be noted researches based on the using shallow networks [1–3,5] and deep or 3D networks [10,13,24,29,32].

In [5], 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].

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 the following:

- 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.

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 [8,10,13,22,24,28,31,32].

In [28], a model for dynamic gesture recognition is presented based on Convolutional Long-Term Memory Recurrent Neural Network (CNNLSTM). In [13], the recognition of gestures is performed by means of a 3D separable convolutional neural network. The main idea of the research - at the network input data are grouped and each group is a separate array of inputs (a separate dimension). In this case, the convolution is carried out between groups.

In [29] multimodal emotion recognition (facial expressions) and speech are considered. The visual representation of emotions in different people is different. Therefore, the main problem of recognizing emotions is the lack of strict timelines for identifying emotions.

The system architecture in [29] consists of three networks: CNN for analysis of audio modality, ResNet for visual modalities analysis, and LSTM for fusion operations on outputs of audio and video modalities. The feature of the proposed methodology is that all networks are trained together.

The synchronization modalities task is also considered in [23]. The proposed multimodal fusion strategy is very effective for the identification of speakers. However, it is a closed system. Fusion pattern is specialized for these modalities. He does not provide the use of other modalities.

Summarizing the analysis of the selected existed approaches, one can formulate common general problems:

- orientation is exclusively on two modalities;
- missing or complicated addition of new modalities;
- use of relatively simple visual scenes;
- impossibility of full or partial effective functioning in cases of lack of input data from one of the modalities.

Contrary to these results, in [27], the authors have proposed multimodal architecture, which is partially devoid of these disadvantages.

The Aim of the Research. In the research, the main problem is the development of the HDCNNMF topology with specified intermediate-type fusion subsystem that is based on multidimensional modalities. The main purpose of this architecture is to increase the quality of segmentation and recognition of objects and actions in complex dynamic scenes.

3 Hybrid Deep Convolutional Neural Network with Multimodal Fusion

3.1 HDCNNMF Topology

HDCNNMF 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 HDCNNMF 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 HDCNNMF, the size of the input, output, kernel, and stride of operations in Sequential models 1–4 are shown in Table 1 for video modalities and Table 2 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 Subsect. 4.3. Tables 3, 4 and 5 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.

Table 1. Size of operations in network topology for video modality processing (generating video modalities feature map)

Layer	Input-size	Output-size	Kernel	Stride
Conv1	$27 \times 64 \times 64 \times 1$	$24 \times 62 \times 62 \times 64$	$4 \times 3 \times 3$	1
Pool1	$24 \times 62 \times 62 \times 64$	$12 \times 31 \times 31 \times 64$	$2 \times 2 \times 2$	$2 \times 2 \times 2$
Conv2	$12 \times 31 \times 31 \times 64$	$10 \times 28 \times 30 \times 128$	$3 \times 4 \times 2$	1
Pool2	$10 \times 28 \times 30 \times 128$	$5 \times 14 \times 15 \times 128$	$2 \times 2 \times 2$	$2 \times 2 \times 2$
Conv3	$5 \times 14 \times 15 \times 128$	$4 \times 12 \times 14 \times 256$	$2 \times 3 \times 2$	1
Pool3	$4 \times 12 \times 14 \times 256$	$2 \times 6 \times 7 \times 256$	$2 \times 2 \times 2$	$2 \times 2 \times 2$
Conv4	$2 \times 6 \times 7 \times 256$	$1 \times 2 \times 7 \times 512$	$2 \times 5 \times 1$	1

In Tables 3, 4 and 5, the spatial size of the 2D or 3D convolution kernel for audio and video modalities has the following dimension $C_d \times H_d \times W_d \times K_d$, where C_d is the kernel size in the time dimension (channels), H_d and W_d are the kernel dimensions of the modality frame in height and width respectively, and K_d is the number of kernels (filters).

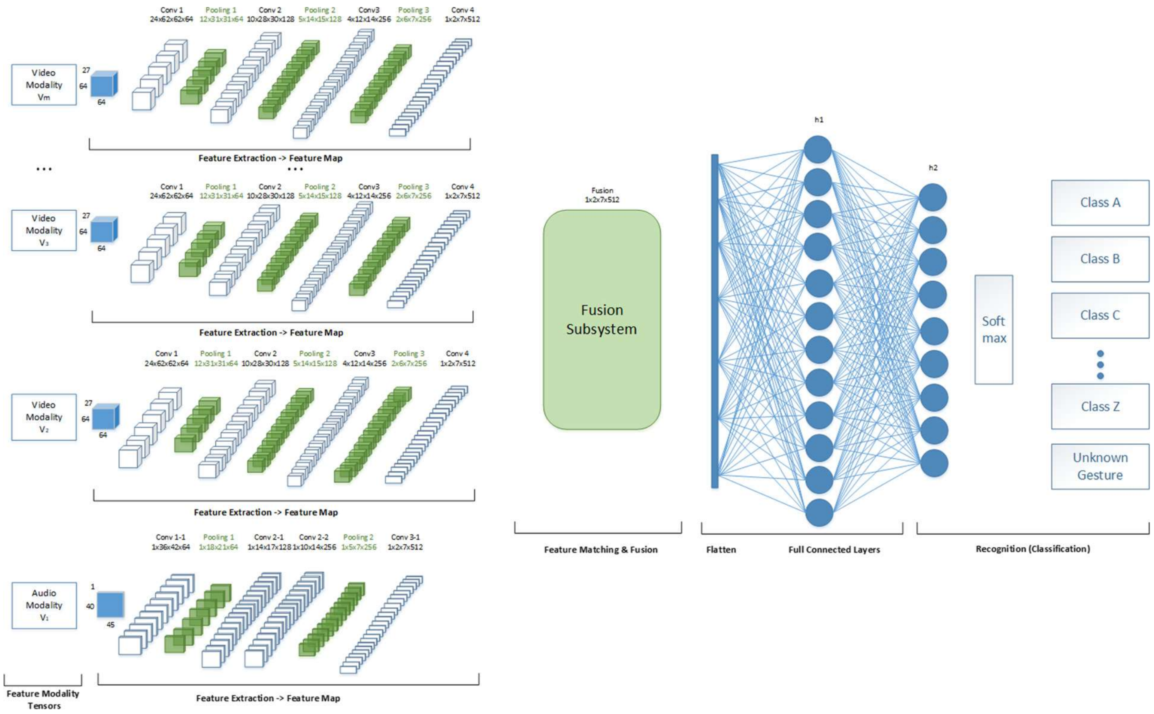

Fig. 1. Hybrid deep convolutional neural network with multimodal fusion topology

Table 2. Size of operations in network topology for audio modalities (generating audio modality feature map)

Layer	Input-size	Output-size	Kernel	Stride
Conv1	$1 \times 40 \times 45 \times 1$	$1 \times 36 \times 42 \times 64$	$1 \times 5 \times 4$	1
Pool1	$1 \times 36 \times 42 \times 64$	$1 \times 18 \times 21 \times 64$	$1 \times 2 \times 2$	$1 \times 2 \times 2$
Conv2-1	$1 \times 18 \times 21 \times 64$	$1 \times 14 \times 17 \times 128$	$1 \times 5 \times 5$	1
Conv2-2	$1 \times 14 \times 17 \times 128$	$1 \times 10 \times 14 \times 256$	$1 \times 5 \times 4$	1
Pool2	$1 \times 10 \times 14 \times 256$	$1 \times 5 \times 7 \times 256$	$1 \times 2 \times 2$	$1 \times 2 \times 2$
Conv3	$1 \times 5 \times 7 \times 256$	$1 \times 2 \times 7 \times 512$	$1 \times 4 \times 1$	1

Table 3. 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	$1 \times 2 \times 7 \times 512$	$1 \times 4 \times 7 \times 512$	–	–
	$1 \times 2 \times 7 \times 512$			
$\Theta 1$	$1 \times 4 \times 7 \times 512$	$1 \times 3 \times 7 \times 512$	$1 \times 2 \times 1$	1

Table 4. Size of operations in fusion subsystem of network topology for processing feature map of modalities (for 3 modalities)

Layer	Input-size	Output-size	Kernel	Stride
G1	$1 \times 2 \times 7 \times 512$	$1 \times 4 \times 7 \times 512$	–	–
	$1 \times 2 \times 7 \times 512$			
G2	$1 \times 2 \times 7 \times 512$	$1 \times 4 \times 7 \times 512$	–	–
	$1 \times 2 \times 7 \times 512$			
Θ1	$1 \times 4 \times 7 \times 512$	$1 \times 2 \times 7 \times 512$	$1 \times 3 \times 1$	1
Θ2	$1 \times 4 \times 7 \times 512$	$1 \times 2 \times 7 \times 512$	$1 \times 3 \times 1$	1
Ω	$1 \times 2 \times 7 \times 512$	$1 \times 4 \times 7 \times 512$	–	–
	$1 \times 2 \times 7 \times 512$			
Θ3	$1 \times 4 \times 7 \times 512$	$1 \times 3 \times 7 \times 512$	$1 \times 2 \times 1$	1

Table 5. 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	$1 \times 2 \times 7 \times 512$	$1 \times 4 \times 7 \times 512$	–	–
	$1 \times 2 \times 7 \times 512$			
G2	$1 \times 2 \times 7 \times 512$	$1 \times 4 \times 7 \times 512$	–	–
	$1 \times 2 \times 7 \times 512$			
G3	$1 \times 2 \times 7 \times 512$	$1 \times 4 \times 7 \times 512$	–	–
	$1 \times 2 \times 7 \times 512$			
Θ1	$1 \times 4 \times 7 \times 512$	$1 \times 2 \times 7 \times 512$	$1 \times 3 \times 1$	1
Θ2	$1 \times 4 \times 7 \times 512$	$1 \times 2 \times 7 \times 512$	$1 \times 3 \times 1$	1
Θ3	$1 \times 4 \times 7 \times 512$	$1 \times 2 \times 7 \times 512$	$1 \times 3 \times 1$	1
Ω	$1 \times 2 \times 7 \times 512$	$1 \times 6 \times 7 \times 512$	–	–
	$1 \times 2 \times 7 \times 512$			
	$1 \times 2 \times 7 \times 512$			
Θ4	$1 \times 6 \times 7 \times 512$	$1 \times 3 \times 7 \times 512$	$1 \times 2 \times 1$	$1 \times 2 \times 1$

Θ_j - Fusion Polling; Ω_j - Feature Fusion Union; G_j - Feature Map Union

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. 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 $= 1 \times 2 \times 7 \times 512$). Batch normalization and Dropout operation were used for all layers. Except for the last layer, all layers used the activation function Parametric ReLU (PReLU) [16]

$$\psi(o^{[i]}) = max(0, o^{[i]}) + a_i min(0, o^{[i]}) \tag{1}$$

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:

$$E(k) = -\sum_{j=1}^{Z} y_j(k) log(\hat{y}_j(k)) \tag{2}$$

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) learning algorithm [21], which can be written in the form

$$W(k+1) = W(k) - \frac{\eta}{\sqrt{S[g^2](k) + \epsilon}} g(k) \tag{3}$$

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

$$S[g^2](k) = \gamma S[g^2](k-1) + (1-\gamma) g^2(k). \tag{4}$$

Here, $g(k) = \nabla_W E(k)$ is the gradient of the optimization criterion by the tuned parameters W, $0 \le \gamma \le 1$.

3.2 Feature Extraction of Modalities

The stage of preprocessing is generation of the features for each modality which is shown in Fig. 2. Input features are formed in the tensor form. The sizes of data tensors of each modality are shown in Fig. 1 and Tables 1, 2, 3, 4 and 5.

Audio Modality. The input signal of audio modality V_1 is a discrete mono speech signal in the format of WAV with a sampling frequency of 16 kHz–40 kHz and a length of T ms. Such a signal may be obtained from the input device or extracted from the video. The voice region in the speech signal had a different start point. Therefore, a speech detector [19] was used to reduce noise in the input set of features.

The discrete speech signal $x = x(t), t = 0, 1, \ldots, T$ splits into frames x_j of length $l(x_j) = 20$ ms:

$$x_j = x(t), t \in I_{t,j} \tag{5}$$

where $I_{t,j}$ is j-th time section of length 20 ms in $[0, T]$.

These sections are the basis of a sample of audio modality features. The length of section is 20 ms because the speech signal in this interval has the pseudo-stationary property. The basic characteristic for forming the features of audio-modalities is MFEC [19] $C_{j,i}$ of the speech signal, which can be determined by the formula

$$S_{j,i} = \sum_{k=0}^{I_{t,j}-1} ln\|X_i(k)\|^2 H_j(k), 0 \leq i \leq I_{t,j} \tag{6}$$

where M is a coefficients number; $X_i(k)$ is a specter of speech signal in interval $I_{t,j}$ using the FFT and Hemming window; $H_j(k)$ is a set of filters for frame x_j.

Using $C_{j,i} = C_{j,1}, C_{j,2}, \ldots, C_{j,M}$ for each x_j we can obtain the self-similarity matrix

$$C_j = [c_{j,(m,n)}]_{m=1,\ldots,M}^{n=1,\ldots,M} \tag{7}$$

the elements of which can be written in the following way:

$$c_{j,(m,n)} = \frac{C_{j,n}}{C_{j,m}}, m = 1, \ldots, M, n = 1, \ldots, M \tag{8}$$

where M is a count of filters.

The matrix C_j is square matrix with dimension $M \times M$. Based this matrix we can define $a_{j,i}$ eigenvectors

$$C_j a_{j,i} = \lambda_{j,i} a_{j,i} \tag{9}$$

where $\lambda_{j,i}$ is the eigenvalues of the matrix C_j. In our case we take $M = 40$.

For further formation of the "cube" features we take the first vectors $a_{j,1}$. Based on the vectors $a_{j,i} = a_{j,(i,1)', \ldots, a_{j,(i,M)}}$ we can obtain matrix of frame features x_j

$$a_j = [a_{j,(m,n)}]_{m=1,\ldots,M}^{n=1,\ldots,N_a}. \tag{10}$$

Here, $a_{j,(m,n)}$ is a normalized value $a'_{j,(m,n)}$. In our case we take $N_a = 48$. The total length of the time span which is considered is $Na \times 20\,\mathrm{ms}$. If the specified duration was not enough, then the matrices a_j can be interpolated or extrapolated.

Video Modalities. For solving practical tasks, we use two topologies of proposed network, which is shown on Fig. 1. The first topology of network (TP_1) consists of two modalities: V_1 (the audio modality) and V_3 (the video modality for lips detection). The second topology of network (TP_2) consists of four modalities: V_1 and V_2, V_3, V_4 (V_2 - video modality for face detection, V_4 – video modality for gestures detection. Such types of modalities are the most widespread).

In the general network topology, which is shown in Fig. 1, three video modalities are considered. The first of these is V_2 - video modality for face detection, the second modality V_3 - video modality for lips detection and the third modality V_4 – video modality for gestures detection. Such types of modalities are the most widespread. Motion detection was performed using algorithm from [14]. Video streams were selected and processed by OpenCV. The feature extraction for V_2 was performed in accordance with the following algorithm:

1. Detecting a motion.
2. Using a background model based on a Gaussian set, we separate the background [15,26].
3. Selecting frames and converting them to grayscale using skvideo and OpenCV libraries.
4. Cropping an excessive part of the image in the frame.
5. Extracting ROI using the algorithm from [27] which is based on the dlib library.
6. Using the bicubic interpolation for forming ROI with size 64 × 64 pixels.

Features for modality V_3 were selected based on a known proportion: the region of the human lip is the third bottom part of the human head region. This approach allows detecting faces along with the extraction of ROI for V_2 modality.

Since this proportion may be slightly different for each person, using the Haar detector and Viola-Jones method [30], ROI of faces are periodically verified. The combined detection of the face and lips gives the opportunity to get the full synchronization of the two modalities. Using 27 sequential frames the tensors of input features for the modalities V_2 and V_3 (with dimensions of $27 \times 64 \times 64$ pixels) are formed. This dimension determines the existence a set of 27 frames. Very often it is difficult to provide such a set. The reasons for this may be short stream or inadequate performance of the feature detector.

To solve this problem, it is suggested to interpolate or extrapolate the required number of frames (that is, the number that is missing) based on the optical stream model. For this model, the value of brightness in each pixel is considered as the discrete value of some function. Accordingly, using the resampling operation based on the discrete Fourier transform [18], we obtain the function of

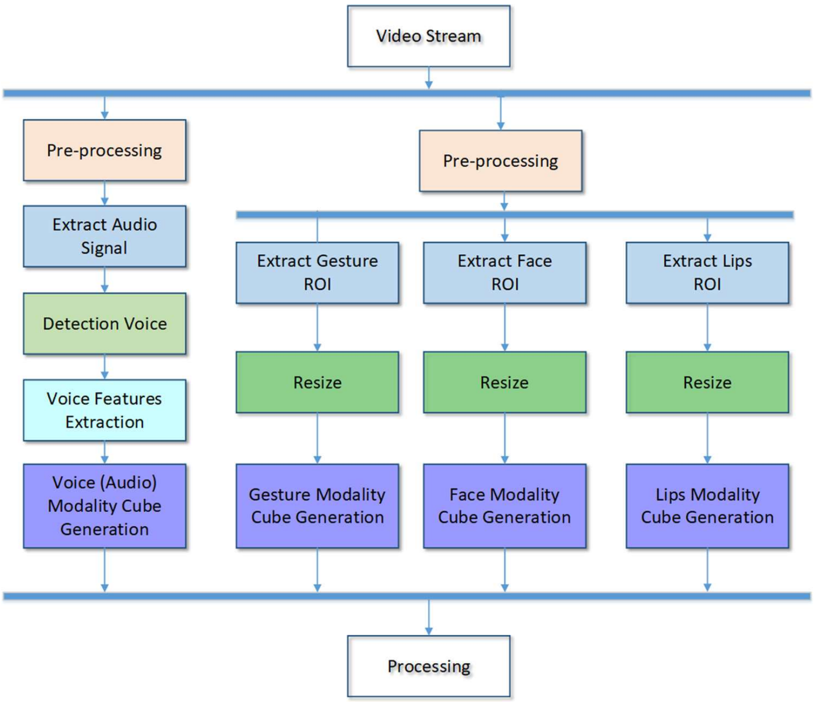

Fig. 2. Modality preprocessing

brightness in pixels of those frames that are missing. In the used dataset, 90% of the video streams were supposed to have one to six frames. Using this approach is satisfactory for a case where at least three ROIs in three different frames were detected in the video stream. In the absence of such frames, a zero data will be specified in the input set of attributes, which will be noise and will negatively affect the quality of work in the system. The features extraction for V_4 modality was performed in accordance with the following algorithm:

1. Detecting a motion.
2. Using a background model based on a Gaussian set, we separate the background [26].
3. Selecting frames and converting them to grayscale using skvideo and OpenCV libraries.
4. Cropping an excessive part of the image in the frame.
5. Increasing FPS up to 55 frames per second based on ffmpeg codec. In the case of the good-quality detection of ROI this operation can be omitted.
6. Extracting ROI for right and left hands based on algorithm from [7].
7. Concatenating the extracted regions into one array.
8. Using the bicubic interpolation for forming ROI with dimension 64×64 pixels.
9. In the case when the number of extracted ROIs were less than 27, the interpolation like for modalities V_2 and V_3 is used.

In the case that only one or two ROIs are detected for all modalities, you can select ROI on the other frames using the geometric characteristics of the detected ROI. This approach is possible when there is a sufficient number of frames in the video stream and the spatial shift of the object of attention is not very large. If the frames in the video stream are not enough, then they can be extrapolated according to the method described above.

4 Experiments, Results and Discussion

4.1 Preprocessing

The input tensors of audio and video modalities have been standardized and in addition, the input tensors of audio modality have been normalized in the interval [0, 1].

4.2 Metric for Performance Evaluation

The main metrics that have been used for performance evaluation are Equal Error Rate (ERR) (which is the point when false acceptance rate (FAR) and false recognition rate (FRR) are equal), Area Under Curve (AUC), Accuracy (Precision), Recall, F1 score. Also, the Micro and Macro Averaged metrics have been used for evaluation accuracy for multi-classes classification.

4.3 Evaluation on LWR Dataset

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 [6].

For comparison of the proposed approach with known ones, the HDCNNMF 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 $27 \times 64 \times 64$ (27 channel visual feature tensor with size 64×64 pixels) and, accordingly, an audio modality has $1 \times 40 \times 45$ dimension, which corresponds to 40 MFCCs, which were calculated on sections 20 ms without overlap on a full audio signal with a length of 900 ms. Then, a matrix of eigenvectors with a dimension of $1 \times \times 45$ was generated based on matrix with MFEC coefficients.

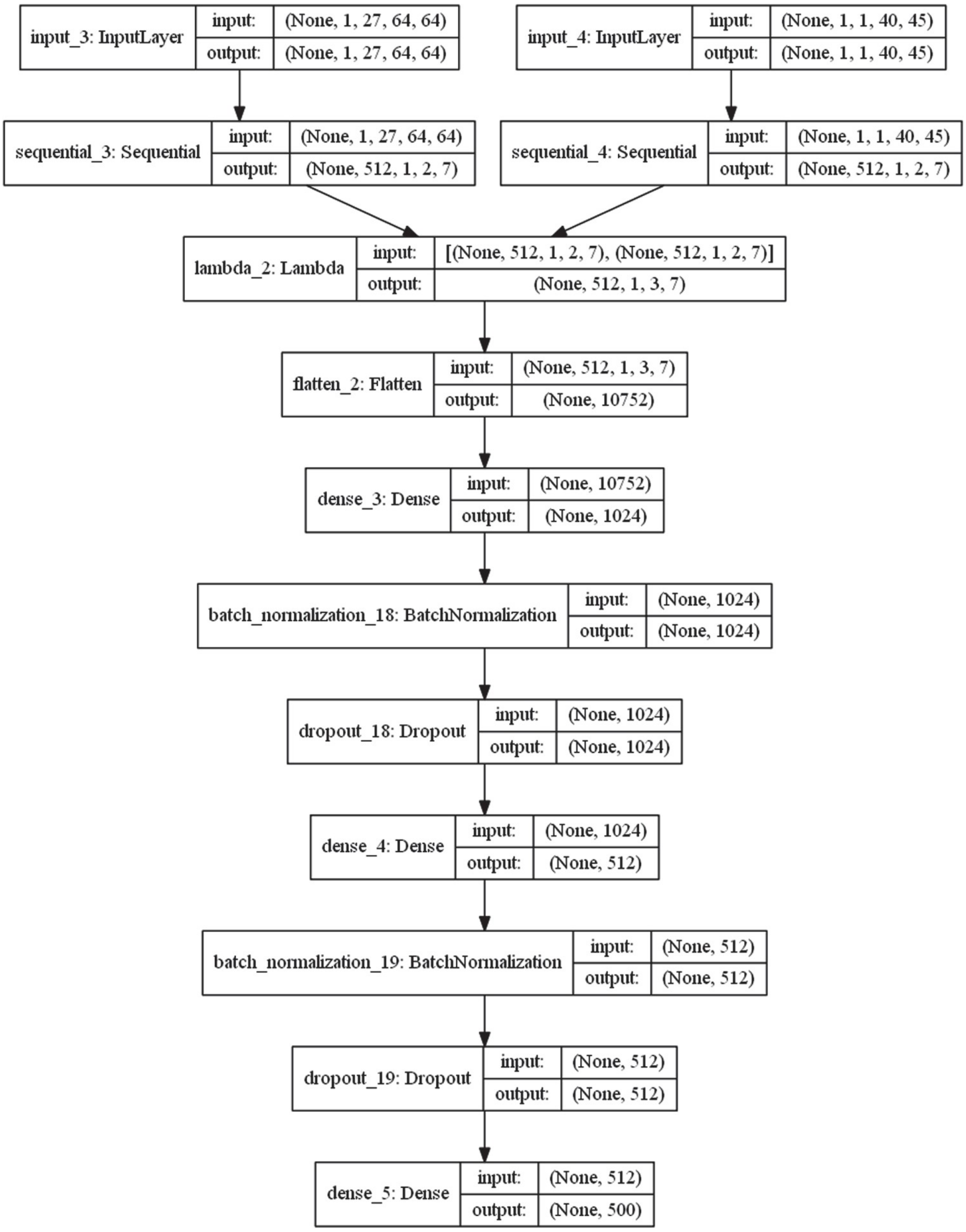

Fig. 3. Hybrid deep convolutional neural network with multimodal fusion topology (layer's dimensions)

Figure 3 and Tables 1, 2 and 3 show 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 [25], the authors have presented the main results for processing Oxford-BBC Lip Reading in the Wild (LRW) dataset based on different approaches. These approaches contain 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 from 15 to 20 epochs. All networks have approximately the same number of parameters (20M).

Table 6 shows the results of Oxford-BBC Lip Reading in the Wild (LRW) data set processing based on different approaches.

Table 6. Size of operations in fusion subsystem of network topology for processing feature map of modalities (for 3 modalities)

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.%
Proposed HDCNNMF	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.

5 Conclusions

As already mentioned the proposed HDCNNMF 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 HDCNNMF 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 HDCNNMF topology allows using the different patterns of fusion. In the process of developing HDCNNMF, 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. The results of experiments and their comparison with the existed approaches confirmed the high accuracy of HDCNNMF both in the case of individual modalities and in the case of their fusion. This fact allows using HDCNNMF in cases where some of the modalities are inaccessible or the modality stream has unsatisfactory quality.

Acknowledgments. 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. Bodyanskiy, Y., Dolotov, A., Vynokurova, O.: Evolving spiking wavelet-neuro-fuzzy self-learning system. Appl. Soft Comput. J. **14**(B), 252–258 (2014). https://doi.org/10.1016/j.asoc.2013.05.020
2. Bodyanskiy, Y., Setlak, G., Peleshko, D., Vynokurova, O.: Hybrid generalized additive neuro-fuzzy system and its adaptive learning algorithms. In: Proceedings of the 2015 IEEE 8th International Conference on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications, IDAACS 2015, vol. 1, pp. 328–333 (2015). https://doi.org/10.1109/IDAACS.2015.7340753
3. Bodyanskiy, Y., Vynokurova, O., Pliss, I., Peleshko, D., Rashkevych, Y.: Hybrid generalized additive wavelet-neuro-fuzzy-system and its adaptive learning. In: Zamojski, W., Mazurkiewicz, J., Sugier, J., Walkowiak, T., Kacprzyk, J. (eds.) Dependability Engineering and Complex Systems. AISC, vol. 470, pp. 51–61. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-39639-2_5

4. Castellini, C., Tommasi, T., Noceti, N., Odone, F., Caputo, B.: Using object affordances to improve object recognition. IEEE Trans. Auton. Ment. Dev. **3**(3), 207–215 (2011). https://doi.org/10.1109/TAMD.2011.2106782
5. Cheng, S.T., Hsu, C.W., Li, J.P.: Combined hand gesture-speech model for human action recognition. Sensors **13**, 17098–17129 (2013). https://doi.org/10.3390/s131217098
6. Chung, J.S., Zisserman, A.: Lip reading in the wild. In: Lai, S.-H., Lepetit, V., Nishino, K., Sato, Y. (eds.) ACCV 2016. LNCS, vol. 10112, pp. 87–103. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-54184-6_6. http://www.robots.ox.ac.uk/ vgg/data/lipreading/lrw1.html
7. Dibia, V.: Handtrack: A library for prototyping real-time hand tracking interfaces using convolutional neural networks (2017). https://github.com/victordibia/handtracking
8. Ding, R., Pang, C., Liu, H.: Audio-visual keyword spotting based on multidimensional convolution neural network. In: Proceedings of the 2018 IEEE International Conference on Image Processing, Athens, Greece, pp. 4138–4142 (2018). https://doi.org/10.1109/ICIP.2018.8451096
9. Favorskaya, M., Nosov, A., Popov, A.: Localization and recognition of dynamic hand gesture based on hierarchy of manifold classifiers. Int. Arch. Photogram. Remote Sens. Spat. Inf. Sci. **XL-5/W6**, 151–161 (2015). https://www.int-arch-photogramm-remote-sens-spatial-inf-sci.net/XL-5-W6/1/2015/isprsarchives-XL-5-W6-1-2015.pdf
10. Feichtenhofer, C., Pinz, A., Zisserman, A.: Convolutional two-stream network fusion for video action recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1933–1941 (2016). arXiv:1604.06573
11. Gibson, J.: The Theory of Affordances. Erlbaum, Hillsdale (1977)
12. Gibson, J.: The Ecological Approach to Visual Perception. Erlbaum, Hillsdale (1986)
13. Hu, Z., Youmin, H., Liu, J., Wu, B., Han, D., Kurfess, T.: 3D separable convolutional neural network for dynamic hand gesture recognition. Neurocomputing **318**, 151–161 (2018). https://doi.org/10.1016/j.neucom.2018.08.042
14. Jackson, L.: Motion-detection-python (2017). https://github.com/ic0n/Motion-Detection-Python
15. KaewTraKulPong, P., Bowden, R.: An improved adaptive background mixture model for real-time tracking with shadow detection. In: Remagnino, P., Jones, G.A., Paragios, N., Regazzoni, C.S. (eds.) Video-Based Surveillance Systems, pp. 135–144. Springer, Boston (2001). https://doi.org/10.1007/978-1-4615-0913-4_11. http://personal.ee.surrey.ac.uk/Personal/R.Bowden/publications/avbs01/avbs01.pdf
16. Kaiming, H., Xiangyu, Z., Shaoqing, R., Sun, J.: Delving deep into rectifiers: surpassing human-level performance on ImageNet classification. In: Computer Vision and Pattern Recognition (2015). arXiv:1502.01852
17. Kampman, O., Barezi, E., Bertero, D., Fung, P.: Investigating audio, video, and text fusion methods for end-to-end automatic personality prediction. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, Melbourne, Australia, pp. 606–611. Springer (2018). https://doi.org/10.1007/978-981-10-7560-5118
18. Lyons, R.: FFT interpolation based on FFT samples: a detective story with a surprise ending (2018). https://www.dsprelated.com/showarticle/1156.php

19. Majeed, S., Husain, H., Samad, S., Idbeaa, T.: Mel frequency cepstral coefficients (MFCC) feature extraction enhancement in the application of speech recognition: a comparison study. J. Theoret. Appl. Inf. Technol. **79**(1), 38–56 (2015)
20. Nguyen, A.: Scene understanding for autonomous manipulation with deep learning (2019). arXiv:1903.09761
21. RMSProp (2020). http://ruder.io/optimizing-gradient-descent/index.html#rmsprop
22. Rua, E., Bredin, H., Mateo, C., Chollet, A., Jimenez, D.: Audio-visual speech asynchrony detection using co-inertia analysis and coupled hidden Markov models. Pattern Anal. Appl. **12**, 271–284 (2009). https://doi.org/10.1007/s10044-008-0121-2
23. Sargin, M., Yemez, Y., Erzin, E., Tekalp, A.: Audiovisual synchronization and fusion using canonical correlation analysis. IEEE Trans. Multimedia **9**, 1396–1403 (2007). https://doi.org/10.1109/TMM.2007.906583
24. Simonyan, K., Zisserman, A.: Two-stream convolutional networks for action recognition in videos. In: Advances in Neural Information Processing Systems, vol. 1, pp. 568–576 (2014). arXiv:1406.21998
25. Stafylakis, T., Tzimiropoulos, G.: Combining residual networks with LSTMs for lipreading. In: INTERSPEECH, pp. 3652–3656 (2017)
26. Stauffer, C., Grimson, W.: Adaptive background mixture models for real-time tracking. In: Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 246–252 (1999). https://doi.org/10.1109/CVPR.1999.784637
27. Torfi, A., Iranmanesh, S., Nasrabadi, N., Dawson, J.: 3D convolutional neural networks for cross audio-visual matching recognition. Comput. Vis. Pattern Recogn. **PP**, 1396–1403 (2017). https://doi.org/10.1109/ACCESS.2017.2761539
28. Tsironi, E., Barros, P., Weber, C., Wermter, S.: An analysis of convolutional long short-term memory recurrent neural networks for gesture recognition. Neurocomputing **268**, 76–86 (2017). https://doi.org/10.1016/j.neucom.2016.12.088
29. Tzirakis, P., Trigeorgis, G., Nicolaou, M.A., Schuller, B., Zafeiriou, S.: End-to-end multimodal emotion recognition using deep neural networks. IEEE J. Sel. Top. Signal Process. **11**(8), 1301–1309 (2017). https://doi.org/10.1109/JSTSP.2017.2764438
30. Viola, P., Jones, M.: Robust real-time face detection. Int. J. Comput. Vis. **57**(2), 137–154 (2004). https://doi.org/10.1109/CVPR.1999.784637
31. Wu, P., Liu, H., Li, X., Fan, T., Zhang, X.: A novel lip descriptor for audio-visual keyword spotting based on adaptive decision fusion. IEEE Trans. Multimedia **18**(3), 326–338 (2016). https://doi.org/10.1109/TMM.2016.2520091
32. Yue-Hei Ng, J., Hausknecht, M., Vijayanarasimhan, S., et al.: Beyond short snippets: deep networks for video classification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4694–4702 (2015). https://doi.org/10.1109/CVPR.2015.7299101