Journal of Modeling and Optimization 2020;12(1):51-59
https://doi.org/10.32732/jmo.2020.12.1.51

Deep Learning Based Human Emotional State Recognition in a Video

A.A. Moskvin, A.G. Shishkin

Faculty of Computational Mathematics and Cybernetics, MSU - Moscow State University, Moscow, 119991, Russia
E-mail: alalmoskvin@gmail.com, shishkin@cs.msu.ru

Received: 28 April 2020; Accepted: 18 May 2020; Available online: 10 June 2020

Abstract: Human emotions play significant role in everyday life. There are a lot of applications of automatic emotion recognition in medicine, e-learning, monitoring, marketing etc. In this paper the method and neural network architecture for real-time human emotion recognition by audio-visual data are proposed. To classify one of seven emotions, deep neural networks, namely, convolutional and recurrent neural networks are used. Visual information is represented by a sequence of 16 frames of 96 × 96 pixels, and audio information - by 140 features for each of a sequence of 37 temporal windows. To reduce the number of audio features autoencoder was used. Audio information in conjunction with visual one is shown to increase recognition accuracy up to 12%. The developed system being not demanding to be computing resources is dynamic in terms of selection of parameters, reducing or increasing the number of emotion classes, as well as the ability to easily add, accumulate and use information from other external devices for further improvement of classification accuracy.
Keywords: Artificial neural networks; Deep learning; Emotion recognition; Video; Speech signal.

1. Introduction

Currently, artificial intelligence technologies are actively used in everyday life, including a wide range of software and hardware systems being able to change their behavior during operation. Among the areas where such technologies play an important role is the recognition of specific object features in a given data stream.

Feature extraction based on deep learning often solves the given problem more effectively than a human, especially when it comes to large amounts of data, where it is necessary to detect non-linear relationship between variables. One of such examples is a recognition of faces and their individual details in images.

There is a significant interest in the solution of these problems due to the diversity of their applications in such areas as security systems, user verification, newsgroups, computer games, medicine, e-learning etc. The implementation of human emotional state recognition by video sequence would help people suffering from neuropsychiatric disorders in correctly determining emotionally colored speech patterns and would provide the opportunity to distinguish the faked emotion from the true one.

The process of determining emotions in a video consists of two major steps: face detection in images, and analysis of video data in order to determine the type of emotion. At the moment when searching for faces in a video, a number of various methods are used, among them the principal component analysis [1], Viola- Jones AdaBoost classifier [2], template matching [3], convolutional neural network [4].

In this paper a system based on deep neural networks has been developed. It allows recognizing human emotions in real-time with limited computing resources from a video sequence in which both visual and speech information is present. The advantage of the system is its robustness, since the system was mainly trained on raw examples, and not on selected and prepared in a special way video.

2. Related work

In recent years a number of interesting results have been obtained in recognition of various face emotions both in still pictures and in video images. For example, software presented at the Future Decoded 2014 conference in the UK was able to analyze a human emotion by his face expression [5]. The method proposed determines the presence in a separate snapshot of such basic emotions as anger, contempt, disgust, fear, happiness, tranquility, sadness and surprise. The results presented in a numeric format belong to the interval from 0 to 1 and represent the probability of a certain emotion, where 0 is the absence of emotion, and 1 is a pronounced emotion. Emotion recognition is also a part of large Microsoft Project Oxford [6]. In addition to the emotion recognition in an image, it allows check of English text spelling and speech recognition.

The results of the competition INTERSPEECH 2009 Emotion Challenge [7] showed that the accuracy of human emotion recognition using only audio channel was far from ideal one. The following results were obtained: the average efficiency for two emotion classes was 71.86%, for five emotion classes - 58.83%. As a result of low efficiency of one-modality based emotion recognition different types of signals describing emotions were used. In [8], multimodal (image, speech and tactile signals) restricted Boltzmann machines (RBMs) are considered that are able to both generate and recognize six emotional states, which are represented as activation of the upper layer neurons. The developed model allows to reproduce the emotion corresponding to the missing modality on the basis of the other two modalities.

For the classification of emotional states in real time by video sequence, as a rule, convolutional neural networks are used. In [9], several data sets were used to train the neural network resulting in 57.1% accuracy of recognition. A significant drawback of the approach in [9] is a small size of the data sets used, besides video sequences are very similar to each other. In [10], with the help of convolutional networks and the Bayes classifier all images were divided into 3 classes. 6470 images were used for analysis. The database was compiled from pictures of various social events, such as a wedding for positive events, a meeting - for neutrals, and outcries — for negative emotional states. Accuracy on the test set was 64.68%. Analysis of the results showed that for some types of emotions, the Bayes method worked better, and for others, the convolutional neural network was preferable, however, their joint use always provided higher efficiency than each method did separately.

A hybrid approach to emotion recognition, when various methods are used for recognizing emotions based on audio and video signals, is presented in [11]. The data from video image are processed by convolutional neural networks, and audio signal is analyzed by deep belief networks. In addition, the authors extract emotional features around the mouth using K-means method and use autoencoder to describe the space-time information. At the prestigious competition EmotiW 2014, the developed model showed an accuracy of 47.67% in recognizing seven basic emotions. In the 2018 paper [12] authors used only video channel to obtain information about the emotion class and as a result they obtained recognition accuracy of 45.51%. Excellent results - more than 99% accuracy - can be achieved by analyzing electroencephalography of the human brain [13].

Paper [14] focuses on a system of recognizing human's emotion from a detected human's face. The analyzed information is conveyed by the regions of the eye and the mouth into a merged new image in various facial expressions pertaining to six universal basic facial emotions. The methodology uses a classification technique of information into a new fused image which is composed of two blocks integrated by the area of the eyes and mouth, very sensitive areas to changes human's expression and that are particularly relevant for the decoding of emotional expressions. Finally, was used merged image as an input to a feed-forward neural network trained by back-propagation. Such analysis can detect emotion with 76 percent of accuracy on Cohn-Kanade (CK) database.

3. Proposed method

The developed method is based on the application of deep neural networks to the video and audio channels of the video stream. According to the result of analyzing information from each channel, a decision is made about the video record belonging to one of the seven classes - a neutral state, anger, sadness, fear, disappointment, happiness, surprise. Processing each of the data channels is a separate large subtask. The following main steps can be distinguished for a video channel:

1) video sequence splitting into separate frames,
2) face detection in images,
3) pre-processing of images,
4) use of the received images as input data for a neural network.

The first stage is the splitting of each video into the sequence of 16 images, each of which must contain a face. At this step, the main problem is the choice of the image in which the object is not in motion, i.e. the face is not blurred. If the image is blurred, it is preferable to use one of the next frames in a video sequence, where the boundaries of the object are clearer. Next, using the AdaBoost method [2], a search and segmentation of a face in the image is performed. At the preprocessing stage, all faces are transformed to a common format, and images that cannot be used for one reason or another are excluded, for example, their size is too small. The last stage is to take an image as an input of a previously pre-trained neural network, which has 7 outputs corresponding to the probabilities of belonging of this sequence of images to each of the given classes.

The processing stages of the audio channel are described below:

1) splitting the audio signal into separate frames,
2) extraction of audio features from each frame,
3) use of the features as input data for the recurrent neural network.

In our case, the audio signal was divided into separate frames based on the Welch periodogram method [15] - the vector of signal samples was divided into overlapping segments (as a rule, 50% overlap was used), after which each segment was multiplied by a weighting function and a discrete Fourier transform (DFT) was calculated. Since

the localization of parameters is important for us, in addition to the DFT, the discrete cosine transform was also used. For practical reasons, the window length was chosen to be 20ms. For each window, a large number of audio features were calculated in both the time and frequency domains. Next, the most significant features resistant to external noise and allowing to accurately describe audio segment were selected. The final stage of audio preprocessing is similar to the final stage of video channel processing, the only difference is in the architecture of the neural network used.

4. Data

One of the important steps in developing emotion recognition system is a search and retrieval of data sets suitable for analysis of human emotions. Almost in all existing databases the emotional label of video images with people's faces, is not known, and the audio channel is either absent or too noisy. As a result, various options for collecting a dataset were considered, namely, retrieving data from video-sharing websites, cutting segments from films, and use of existing databases.

Video-sharing websites such as YouTube contain a lot of video and, in addition, allow to search by tags. Unfortunately, most relevant and authorized for download videos are of insufficient quality.

The next considered option was the use of fragments of feature films (that are free or with copyright expired) where a character exhibits specific emotion. The advantage of this approach is high image quality and the ability to obtain a large amount of required data. However, the need to manually set the boundaries of the video sequence, within which specific emotion is present, significantly slows down the process.

The existing emotion databases are the most suitable option. As a rule, they have been many times used to solve various problems of pattern recognition and consist of labeled examples. However, freeware datasets are usually quite small in volume.

As a result, we decided to use fragments of feature films, complementing them with the most suitable for our problem TR-CS-11-02 Acted Facial Expressions In The Wild Database [16], that was preprocessed in above-mentioned manner. This database contains video sequences divided into 7 classes; however, some video sequences were unsatisfactorily short (<1s). The total number of video sequences used was neutral - 207, anger - 197, sadness - 178, fear - 127, disappointment - 114, happiness - 113, surprise - 120.

5. Image selection from video sequence

To use a video channel for emotion recognition, one must select a sequence of images where a face is present in each image, and all images are of good quality. Then all video sequences must be splitter up into equal number of frames and a single frame must be chosen among them. In case of random or according to some rule choice, due to the movement of objects, there is a possibility that the selected image contains a fuzzy ("blurred") face, which further affects the quality of recognition. Hence, it is necessary to select a frame that does not contain blurred face.

5.1 Collecting a dataset of blurry and clear images

To determine whether an analyzed face in an image is clear or blurred, a separate dataset was collected, which included face images of both types. The dataset contained random frames from video sequences. Since the manual separation of random images into clear and blurry ones is a time-consuming process, we have developed a special method to do this automatically. One of the matrix filters was applied to each image: low frequency

$$L = \frac{1}{14}\begin{bmatrix} 1 & 2 & 1 \\ 2 & 2 & 2 \\ 1 & 2 & 1 \end{bmatrix}$$ (labeled as blurry images)

and high frequency

$$H = \begin{bmatrix} 0 & -1 & 0 \\ -1 & 5 & -1 \\ 0 & -1 & 0 \end{bmatrix}$$ (labeled as clear image) [17].

Since only a moving object is blurred, the filter should be applied only to the selected part of the image that contains the face, and not to background. The filter boundaries were chosen empirically, depending on the fraction of the image that face occupied. Before applying each filter, the center of the face was located by AdaBoost method [2], and the coordinates of face center were served as a center of square region where the filter was applied. The size of the region was set randomly in the range from 30 to 60% of the entire image size.

5.2 Image type determination

To determine the type of an image - blurred or clear one - a feedforward neural network with 6 fully connected layers was developed. For each grayscale image in the training set, we made histograms, with range of values from 0 to 255, which were used as inputs to the neural network. The rather simple neural network architecture was chosen to satisfy the requirements of real-time solution. However, the results were unsatisfactory. Therefore, two more layers were added to add noise to the input data. To further improve results image histograms were built to visualize and analyze the data as well. In Fig. 1 each picture shows the comparison of the number of pixels of a certain value in a randomly chosen blurred (light gray) and clear (dark gray) images.

Finally, Sobel filters in each direction and Laplace filter were applied to the analyzed images. The histograms of these images for each color channel (3 × 256 vectors) were made and used as an input to the feedforward neural network. We obtained the accuracy over 98%. This model was later used in the selection of the required frame from the video sequence.

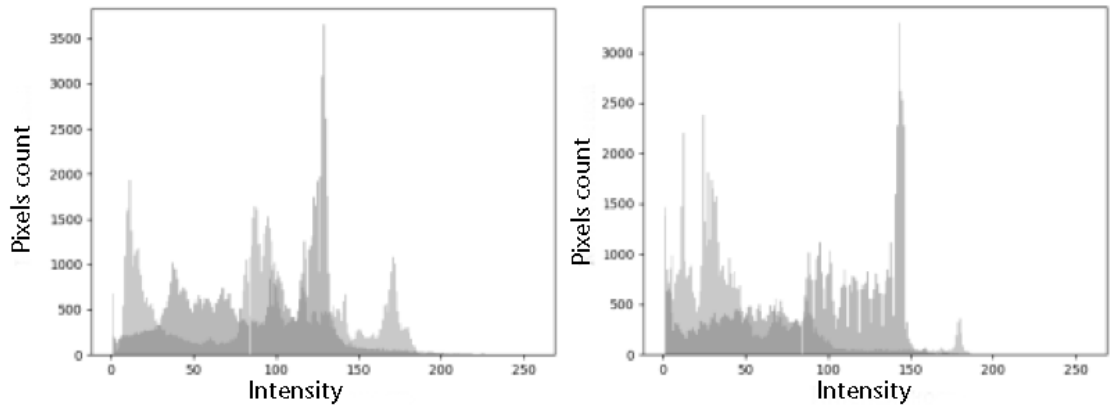

Fig. 1. Comparison of the number of pixels of a certain value in randomly chosen blurred (light gray) and clear (dark gray) images.

6. Audio features selection

The amplitude of the speech signal varies significantly over time. In particular, the amplitude of unvoiced speech segments is significantly less than the amplitude of voiced segments. As a result, it is necessary to extract features from quite short segments of speech signal.

We divided speech signal into separate frames with a length of 20ms. Then for each frame the following features were extracted by OpenSmile software [18].

1) Mel-frequency cepstral coefficients (the first 15 coefficients). As is known, speech generation depends, first of all, on the form of the vocal tract, which determines which sounds are pronounced at the moment. The shape of the vocal tract is reflected, for example, in the envelope of the short-term power spectrum of speech signal. Mel-frequency cepstral coefficients are able to represent this envelope with the help of a small amount of data.

2) Linear prediction coefficients (8 first coefficients).

3) Zero-cross rate that has low values for voiced speech and high values for unvoiced [19].

4) Speech fundamental frequency.

5) Average value of the spectrum of the analyzed speech signal.

6) Normalized average values of the spectrum.

7) Intensity, amplitude, energy, window number.

Also, statistical characteristics and moments for the above parameters were added: arithmetic average of the signal envelope, standard deviation of envelope values, maximum value, minimum value, delta coefficients, third-order moment, fourth-order moment, fraction of time when the value exceeds 75% of the maximum, the proportion of time when the magnitude value exceeds 90% of the maximum, the first three quartiles.

Finally, 562 parameters were obtained for each frame. Due to the large number of features, it was necessary to reduce their number, selecting only the most significant ones.

6.1 Parameter covariance

To select the most significant features describing an audio channel, we used a following method of parameter covariance:

1) Select all pairs of speech features whose covariance coefficient is greater than a certain threshold value.

2) Of all the pairs, select only those that are found in more than a half of audio signals that were analyzed.

3) Build a connected graph and select connected components in it.

4) From each component, select any parameter that will later be in the result set.

In speech signals for all 7 classes, pairs of audio features were found, the covariance coefficient of which was greater than the value of C, for different values of C ∈ [0.50, 0.55, 060, 065, 0.70, 0.75, 0.80, 0.85, 0.90]. The numbers of different feature pairs in speech signals of a certain class were analyzed, and it can be concluded that the emotion classes do not have the specific properties of having too large or too few covariance pairs with respect to others, that is, the described algorithm does not depend on the emotion class.

A total of 1136 speech signals were analyzed, of which pairs of features were selected that occur in more than a half of signals. The connected components were found for each value of the threshold C. In Fig. 2 the dependencies of the size of the obtained set of covariation pairs and connected components on the covariance coefficient are shown.

To minimize the dimensionality and maximize the covariance, a covariance coefficient of 0.90 was chosen. This decision was made on the basis of Fig. 2, from which it follows that with a 90% coincidence, more than half of the features disappear, which is quite enough for solving the problem.

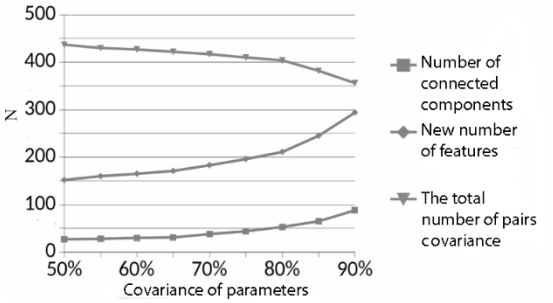

Fig. 2. The dependence of N on covariance of parameters, where N is the number of connected components (shown by squares), the new number of features (diamonds) or the total number of covariation (triangles).

6.2 Autoencoder

Autoencoder is a neural network consisting of two logical parts - encoding and decoding. Encoding is some function $g: g(x) = x', \dim(x) > \dim(x')$ and decoding is a function $f: f(x') = x'', \dim(x'') = \dim(x)$, where x' is a vector of new parameters received. It is required to choose the functions f and g so that the error function $L(x, x'') \to 0$. For the problem of reducing the number of audio features, a sequence of 30 vectors with 562 parameters in each of them was used as an input to autoencoder network. When encoding the original features, their temporal order does not influence on their compression; therefore, the autoencoder did not contain recurrent layers, but only fully connected ones, and had the form presented in Table 1.

Table 1. Autoencoder structure

Layer Type	Number of Inputs	Number of Outputs
Fully connected 1	562	400
Fully connected 2	400	256
Fully connected 3	256	400
Fully connected 4	400	562

After analyzing various loss functions, the following error function was chosen:

$$\frac{1}{N}\sqrt{\sum_i \left|\log x_i'' - \log x_i\right|^2}.$$

7. The structure of neural networks

To solve the problem of human emotional state recognition by audiovisual data, various types of deep neural networks were used, namely, convolutional and recurrent networks. Convolutional networks were used for image processing, and recurrent networks were used for audio data analysis.

For the speech signal, a recurrent neural network consisting of nine layers and receiving as an input a matrix of 37 × 256 was implemented. Here 256 denotes the number of features in each frame of the audio signal and 37 is a number of frames. The output is a vector of 7 real values. Each value ranging from 0 to 1, denotes the probability that input features belong to one of the emotional classes. The structure of the recurrent network is shown in Fig. 3.

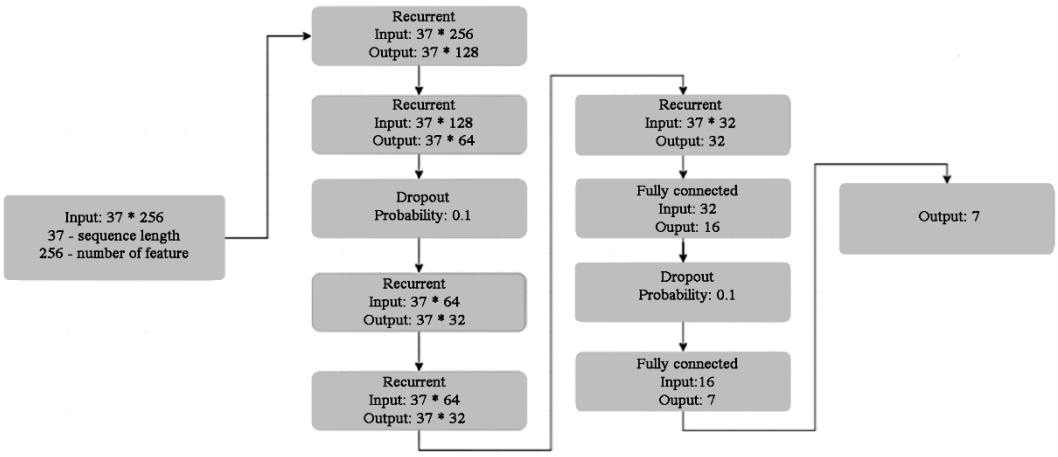

Fig. 3. Audio channel neural network structure.

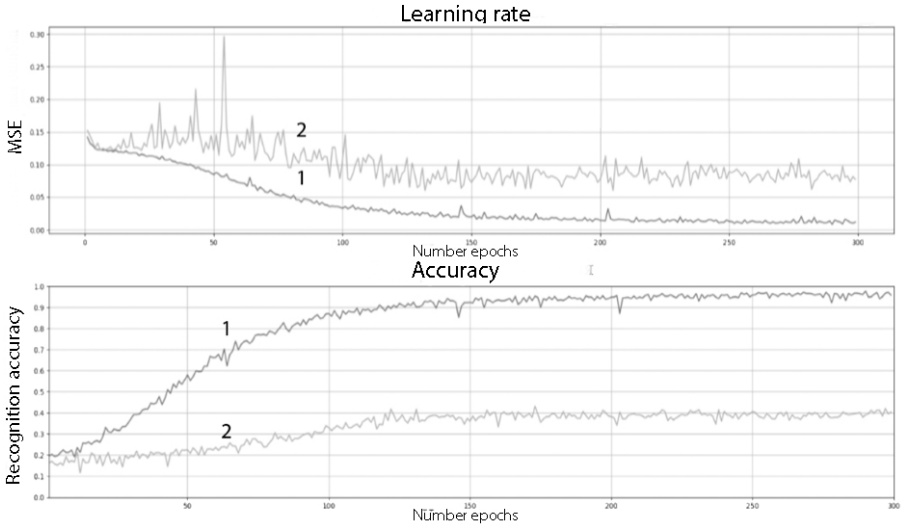

Fig. 4. Train (1) and validation (2) learning of the neural network for an audio channel.

With the help of this model, as can be seen from Fig 4, by the 120th epoch an accuracy of 39% was achieved on validation set, and later it did not change. At the same time the accuracy on the training set was almost 100%. This indicates the overfitting - the state of the network, when the model uses features that are typical for the training set and is not capable of generalization. To overcome the overfitting, the K-fold cross-validation and dropout were used [20].

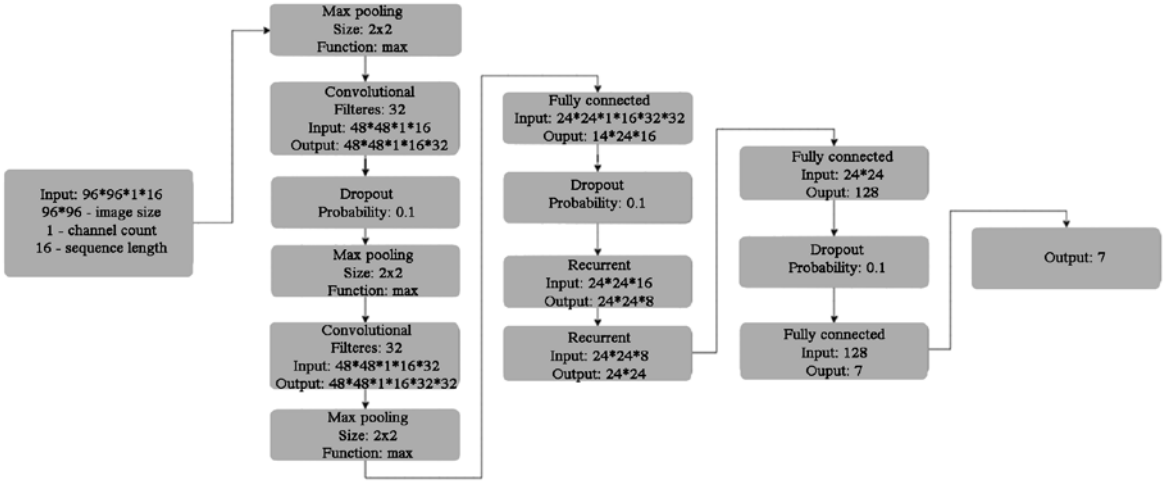

Fig. 5. Video channel neural network.

When analyzing the video, a more complex model shown in Fig. 5 was used. It used 16 grayscale images of 96x96 pixels as an input to neural network that consisted of 16 layers, including 2 convolutional layers, 3 pooling layers, 3 dropout layers, 3 recurrent layers and 4 fully connected layers.

The training of the model was faster than in the case of the audio channel, and by the 70th epoch maximum accuracy was achieved on the training set (Fig. 6). After that, small fluctuations were observed on the validation set ranging from 45 to 50%. As a result, a peak with 49% accuracy was chosen as a moment to stop training.

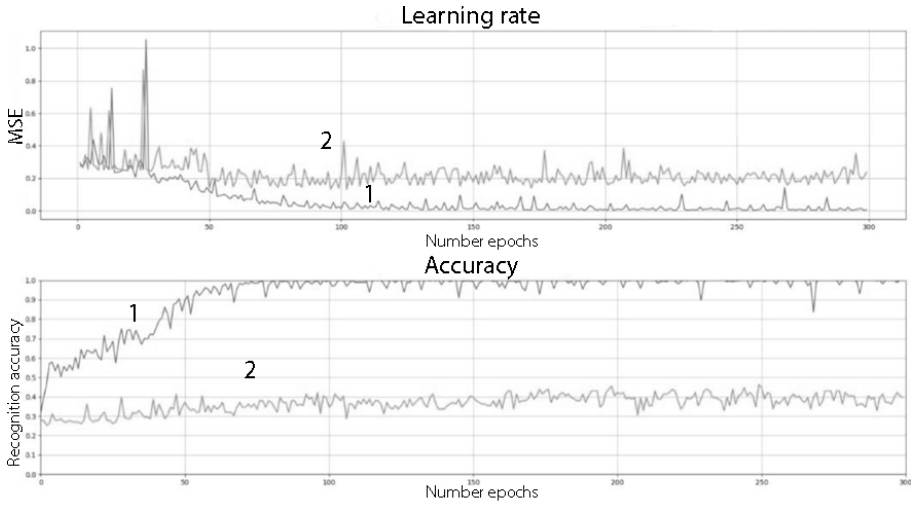

Fig. 6. Train (1) and validation (2) learning of the neural network for a video channel.

8. Analysis of the results

To estimate the performance of the developed model the ROC curve and area under this curve (AUC) were used. The ROC curve represents the dependence of true positive classifications rate on the false one's rate at various threshold values and is one of the most popular tools to measure the efficiency of solving binary classification problems [21]. In order to extend these concepts to non-binary classification problems, a pairwise comparison can be used, i.e. converting the problem into binary one by one-against-all approach. In Fig. 7 the ROC curves for emotion recognition when using both video and audio channels, as well when using either video or speech are shown. It can be seen that the recognition of the emotional state in the case of all available information is more effective comparing to the use of individual channels. To fuse the results of two neural networks, an elementary neural network consisting of two neurons was used, taking two parameters as an input, namely the results of recognition of the audio and video neural network, and having one output parameter-emotion class.

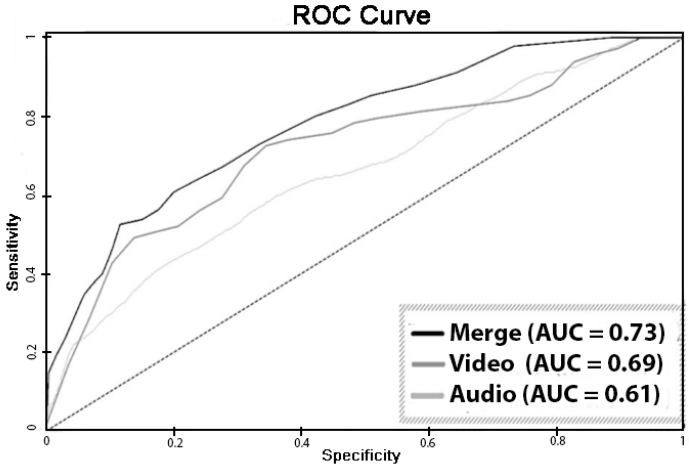

Fig. 7. ROC curve.

It can be seen from Fig. 8 where the train and test results of the final neural network for both video and audio channels are presented, that after 20 epochs, an accuracy of 69% was obtained. On the validation set, the accuracy reached a maximum value of - 59%.

Analysis of the error matrix presented in Fig. 9, showed that most errors occurred in the case of emotions that are physiologically similar. So, for example, in 17% of cases, anger it was recognized as a disappointment, and in 26% of cases fear was defined as a disappointment.

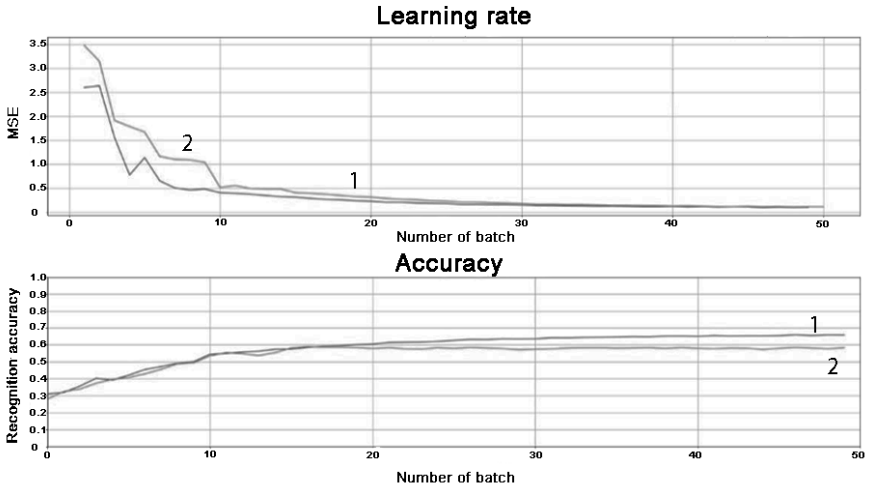

Fig. 8. Train (1) and validation (2) learning of the neural network for a merge channel.

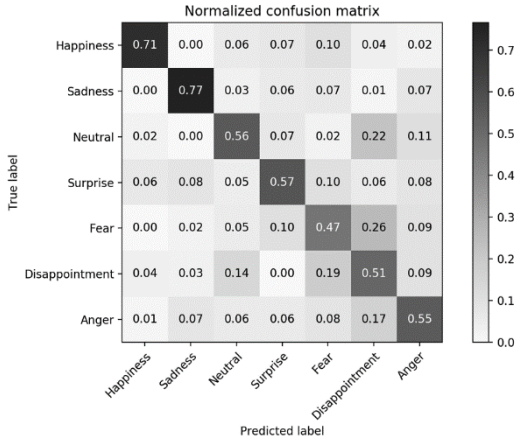

Fig. 9. Normalized confusion matrix.

9. Conclusion

In this paper, the problem of determining the human emotional state from a video is considered. To recognize one of the seven given emotions, the model consisting of deep neural networks, namely, convolutional and recurrent ones was developed.

One of the advantages of the system is its robustness, since the system was mainly trained on raw unprocessed videos.

Additional difficulty to the problem gives the fact that the emotional state of a person can change during the same video sequence. For the training of neural networks, a database of fragments of feature films was collected, as well as the freely distributed TR-CS-11- 02 dataset acted facial expressions in the wild database [16] was used. Due to the movements of people in video, their faces in individual frames may be blurred. To select the clear images, we developed a special neural network with very modest computing requirements.

Automatic data structuring and bringing them to a unified format is implemented, which includes scaling images to one size, removing image parts not affecting the recognition, creating train and test sets. To determine the probability of an emotional state from one of seven possible classes, deep neural network has been developed, consisting, among other, of convolutional and recurrent layers. To avoid the overfitting of neural network, a method of evaluating the analytical model and its behavior on independent data was used, besides dropout layers were added to the deep neural network.

Based on a series of experiments, the optimal hyperparameters of the neural network as well as parameters of data processing were chosen. As a result the emotion classification model achieved a final accuracy of 59%. In a

computer with an Intel® Core™ processor i5-6200U with 2.30GHz and 8 GB RAM data processing time was less than 0.5s, which makes it possible to solve the problem in a real time. It should be noted that the use of audio data, in addition to visual one, allows significantly increase the accuracy of correct recognition of the human emotional state.

10. References

[1] Hemachandran K, Ningthoujam SD. Face recognition using principal component analysis. International Journal of Computer Science and Information Technologies. 2014;5(5): 6491-6496.
[2] Huang X, Acero A, Hon HW. Spoken language processing: a guide to theory, algorithm, and system development. Prentice Hall PTR; 2001.
[3] Ashok V, Balakumara T, Gowrishankar C, Vennila ILA, Kumar AN. The fast haar wavelet transform for signal & image processing. Arxiv Preprint ArXiv. 2010;7(1):126–130.
[4] Yang W, Zhou L, Li T, Wang H. A face detection method based on cascade convolutional neural network. Multimedia Tools and Applications. 2019;78(17):24373-24390.
[5] Chronaki G, Hadwin JA, Garner M, Maurage P, Sonuga-Barke EJ. The development of emotion recognition from facial expressions and non-linguistic vocalizations during childhood. British Journal of Developmental Psychology. 2015;33(2):218-236.
[6] Ho CC. Azure machine learning walkthrough. 2016. DOI: 10.13140/RG.2.1.3171.2247.
[7] FanY, Lam JCK, Li VOK. Multi-region ensemble convolutional neural network for facial expression recognition. arXiv:1807.10575 [Retrieved 13 Aug 2018].
[8] Horii T, Nagai Y, Asada M. Emotion recognition and generation through multimodal restricted Boltzmann machines. In:2015 IEEE/RSJ International Conference on Intelligent Robots and Systems Workshop on Grounding Robot Autonomy: Emotional and Social Interaction in Robot Behaviour. 2015. Hamburg, Germany.
[9] Duncan D, Shine G, English Ch. Facial emotion recognition in real time. http://cs231n.stanford.edu/reports/2016/pdfs/Report.pdf.
[10] Fan Y, Lam JCK, Li VOK. Multi-region ensemble convolutional neural network for facial expression recognition. arXiv:1807.10575. [Retrieved 2017].
[11] Kahou SE, Bouthillier X, Lamblin P, Gulcehre C, Michalski V, Konda K, Jean S, Froumenty P, Dauphin Y, Boulanger-Lewandowski N, Ferrari RC. EmoNets: Multimodal deep learning approaches for emotion recognition in video. Journal on Multimodal User Interfaces. 2016;10(2):99-111.
[12] Sun MC, Hsu SH, Yang MC, Chien JH. Context-aware cascade attention- based RNN for video emotion recognition. In: 2018 1st Asian Conference on Affective Computing and Intelligent Interaction, ACII Asia. 2018.
[13] Moon SE, Jang S, Lee JS. Convolutional neural network approach for EEG-based emotion recognition using brain connectivity and its spatial information. In:2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE. 2018.p. 2556-2560.
[14] Rázuri JG, Sundgren D, Rahmani R, Cardenas AM. Automatic emotion recognition through facial expression analysis in merged images based on an artificial neural network. In: 2013 12th Mexican International Conference on Artificial Intelligence. 2013. IEEE. p. 85-96.
[15] Rabiner LR, Juang BH. Fundamentals of speech recognition. Prentice Hall, Englewood Cliffs, NJ; 1993.
[16] McDuff D, Kaliouby R, Senechal T, Amr M, Cohn J, Picard R. Affectiva-MIT facial expression dataset (AM-FED): Naturalistic and spontaneous facial expressions collected "In-the-Wild". In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops 2013. p. 881-888.
[17] Bolon P. Two-dimensional linear filtering. Digital Filters Design for Signal and Image Processing. 2010. p. 233–260.
[18] Lukianitsa AA, Shishkin AG. Automatic detection of changes in emotional states via speech signal. Speech Technologies. 2009(3): 60-76.
[19] Eyben F, Weninger F, Wollmer M, Schuller B, Munchen T. Open-source media interpretation by large feature-space extraction. TU Munchen, MMK. 2016.
[20] Goodfellow I, Bengio Y, Courville A. Deep learning. MIT Press. 2016.p. 499-522.
[21] Ferri C, Hernández-Orallo J, Salido MA. Volume under the ROC surface for multi-class problems. In: European Conference on Machine Learning 2003. Springer, Berlin, Heidelberg. 2003. p. 108-120.