Facial Expression Recognition Algorithm Based on Deep Convolution Neural Network

Leonid Ivanovsky, Vladimir Khryashchev, Anton Lebedev
P.G. Demidov Yaroslavl State University
Yaroslavl, Russian Federation
leon19unknown@gmail.com, vhr@yandex.ru, lebedevdes@gmail.com

Igor Kosterin
Fire and Rescue Academy
Ivanovo, Russia
kosteriniv@gmail.com

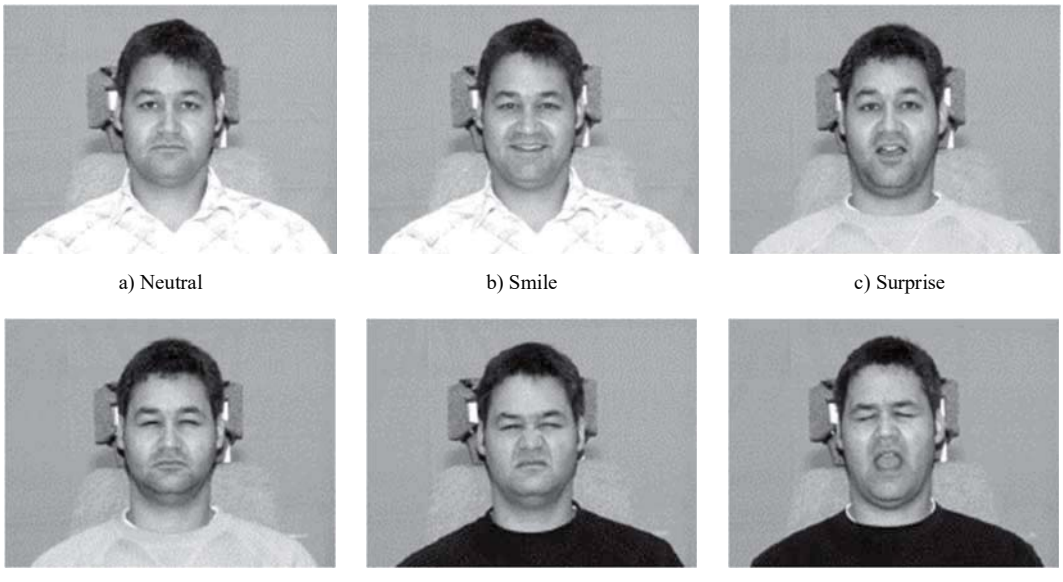
***Abstract*—This paper presents algorithms for smile detection and facial expression recognition. The developed algorithms are based on the implementation of a relatively new approach in the field of deep machine learning - a convolutional neural network. The aim of this network is to classify facial images into one of the six types of emotions. The studying of algorithms was carried using face images from the CMU MultiPie database. To accelerate the neural network operation, the training and testing processes were performed parallel, on a large number of independent streams on GPU. Fo r developed models there were given metrics of quality.**

I. Introduction

The purpose of this paper is to develop and study the robust algorithm of facial expression recognition and smile detection on face image based on deep convolutional neural networks. It is well known that a person is able to solve this task quite easily without effort or delay, regardless of gender, age, race of the analyzed face, as well as the presence or absence of additional factors such as eyeglasses, jewelry, mustache or beard.

An automatic solution of this problem can be used in clinical psychology, psychiatry, neurology, human-computer interaction interfaces (for example, to analyze the impact of advertising), in lie detectors. In addition, in recent years, the use of such algorithms is becoming popular in the field of video analytics applications, such as evaluating front office staff performance or in security systems and systems of terrorist identification. The combination of facial expression and speech can be used for systems of multimodal person identification [1]. In practical applications, additional complications arise due to the face rotation, the presence of optical obstacles, the presence of noise, blurring of the face, insufficient lighting of the scene image resolution, etc. [2], [3].

The human face is characterized by a fairly wide range of facial expressions in everyday life. In the situation of developing an algorithm based on machine learning, it is necessary to identify certain classes of facial expression [4]. In most works, 6 or 7 such classes are distinguished: neutral, smile, surprise, squint, disgust, scream and anger. The examples of images of such classes are shown on Fig. 1.



a) Neutral b) Smile c) Surprise

d) Squint e) Disgust f) Scream

Fig. 1. Six classes of facial expression for images from the CMU Multi-Pie Face Database

The task of facial expression recognition has some difficulties. The facial expressions could be identified by observing face signals. But some types of emotions differ from each other in a few discrete facial features. This factor is also depends on individual differences of subjects, such as degree, frequency or rate of expressions. Individual peculiarities in these characteristics are an important part of human's identity. In addition, facial expressions can be spontaneous. An ideal facial expression analysis system has to address all these factors to output accurate recognition results [5].

Thus, the task of the algorithm is to detect a person's face in the video stream, highlight the necessary characteristics and classify it into one of these categories using machine learning algorithms. It must perform automatically and in real-time for all stages.

At the stage of face detection, various modern algorithms can be used: from classical Viola-Jones based on a set of rectangle features, to computationally effective PICO, or algorithms based on deep convolutional neural networks. In most practical situations, such algorithms are able to detect more than 90% of the observed faces in the video stream. In this task there are two possible practical approaches: face detection on each frame or face detection on the first frame and further tracking it [6].

There are two approaches to the identifying characteristics for the task of emotion recognition: the use of geometric features or features based on appearance. Geometric features are the shape and location of the mouth, eye, eyebrows, nose and the presence of a mustache or beard. Features based on appearance, such as Gabor wavelets, are applied to specific areas on the face image to extract the vector function [5], [7].

The final classification can be carried out in each video frame or on the basis of the video fragment analysis. The most modern approach to solving this problem is the use of a deep convolutional neural network, which makes it possible to exclude from consideration the stage of the formation of features, since they are formed in this case automatically in the process of training the neural network.

The use of this approach allows obtaining the best results of classification in many problems [1], [4], [5]. So in the article [9] the different types of the deep convolutional neural network architecture are given. These types of architecture allow solving the problems of predicting the gender and age of a person using person's face image. The paper [10] describes the use of the convolutional neural network in the problem of classifying objects of different types in an image. In paper [11] there is presented two-stage classifier to recognize one of seven facial expressions. At the first stage, support vector machines is used for the pairwise classifiers, each SVM was trained to recognize two emotions. At the second stage, due to the approach of multinomial logistic ridge regression, there is the transformation of the representation produced by the first stage into values of probabilities. The best performance was 91.5%. The paper [12] describes a hybrid system in which deep convolutional neural network and Logistic regression classifier are combined. At the first stage, the network recognizes face images. At the second stage, the regression classifier is used to classify the features learned at the previous stage. The accuracy was 86.06%. In paper [13] there is also presented a two-stage classifier to distinguish one of seven facial expressions. In the first place, a neural network is used to classify neutral and nonneutral types of emotions. Then Gaussian mixture models are used for the rest types of emotions. The best performance was 71%. The paper [14] describes the architecture of convolution neural network for the face recognition problem. At first, facial features are extracted using feature extraction algorithm. At second, in images there are obtained that have high probability of feature occurrence using a Gaussian model. And finally, the convolutional neural network extracts the information about the expressions based on the image features. The accuracy was around 85%. In paper [15] the developed convolutional neural network is used for the task of facial expression recognition. This network is trained with a combination of raw pixel data and Histogram of Oriented Gradients featrures. The best performance was 80.5%.

This article consists of five parts. The first part is devoted to the process of facial expression recognition and its practical applications. The second is devoted to the standard test image databases applied to the problem of facial expression recognition. The third section is devoted to the convolutional neural network. It describes the peculiarities of using this approach in the field of deep machine learning, as well as about the tools for building and training a classifier of this type. Also this section describes the network's architecture that was used in the task of smile detection and facial expression recognition on images, as well as the peculiarities of classifier training. The forth section shows the results of testing of the developed convolutional neural network. For both tasks, the accuracy value of the classifier is indicated, a confusion matrix and some metrics of the network's operation quality are given. In addition, for the task of smile detection there was built ROC curve and calculated AUC-ROC value. Then, in conclusion you can find summarizing and suggestions about the possible improvement of the quality of the proposed classifier.

II. FACIAL EXPRESSION DATABASES

A standardize database of images is the important part for learning, efficiency evaluation and comparative analysis of different methods for facial expression recognition. Nowadays, there are some available databases of images and videos.

The Cohn–Kanade AU-Coded Face Expression Database (Cohn–Kanade) consists of gray level images of two views of faces (frontal view and 30-degree view) in 210 adults between the ages of 18 and 50 years. They were 69% female, 31% male, 81% Euro-American, 13% Afro-American, and 6% other ethnical groups [16]. The examples of images from Cohn–Kanade Database are shown on Fig. 2.

Fig. 2. Examples of images from the Cohn–Kanade AU-Coded Face Expression Database

Fig. 3. Examples of images the from the Japanese Female Facial Expression Database

Fig. 4. Examples of images from the MMI Facial Expression Database

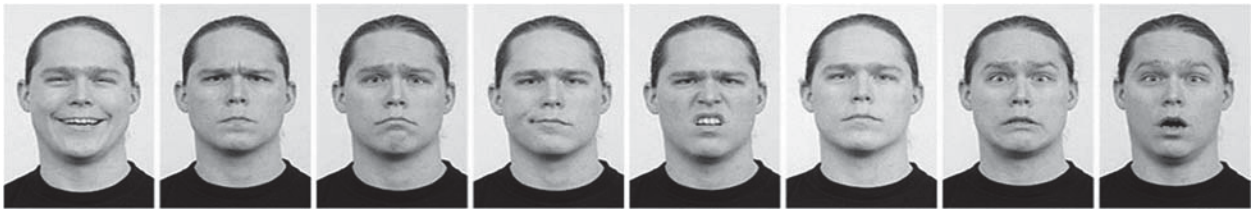


Fig. 5. Examples of images from the Radboud Faces Database

Fig. 6. Examples of images from the CMU Multi-Pie Face Database

The Japanese Female Facial Expression (JAFFE) Database contains 213 gray level images of 7 basic facial expressions (neutral, sadness, surprise, happiness, fear, anger, and disgust) posed by 10 Japanese female subjects [17]. The examples of images from JAFFE Database are shown on Fig. 3.

The MMI Facial Expression Database (MMI) contains more than 1500 samples of both static images and photo sequences of faces from 19 subjects in frontal and profile views displaying various facial expressions [18]. The examples of images from MMI Database are shown on Fig. 4.

The Radboud Faces Database (RaFD) consists of more than 8000 color images of 8 facial expressions (neutral, sadness, contempt, surprise, happiness, fear, anger, and disgust) from 5 angles of view (180°, 135°,90°, 45°,0°) from 67 different subjects [19]. The examples of images from RaFD are shown on Fig. 5.

The CMU Multi-PIE Face Database (Multi-Pie) contains more than 750,000 color images of 6 facial expression (neutral, smile, surprise, squint, disgust, scream) from 337 subjects. The images were made on different angles (less than 90°) with different lighting of the scene [20]. The examples of images from Multi-Pie Database are shown on Fig. 6.

III. Modeling of a Convolutional Neural Network

The paper presents a developed algorithm, based on convolutional neural network - a special architecture aimed at the rapid and qualitative recognition of various objects, as well as their effective classification [9]. Convolutional neural networks are related to the algorithms of deep machine learning. The peculiarity of convolutional neural networks is that image descriptors are formed due to a two-dimensional convolution operation, while the convolution filters themselves are formed during the learning process [8].

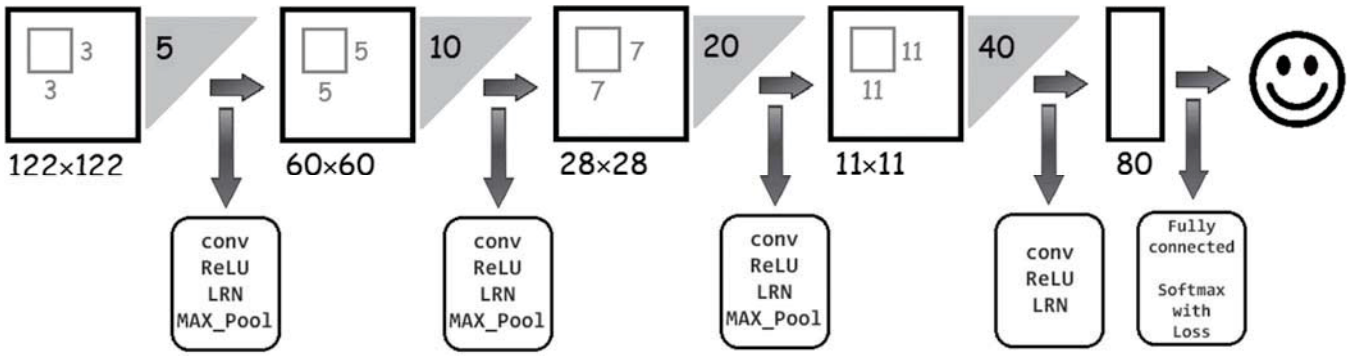


Fig. 7. Convolutional neural network architecture

According to results, in most modern problems of classification, this approach of deep machine learning shows the best result [1].

The network's architecture was implemented using Caffe framework [21] based on the deep convolutional neural network proposed in [9]. This library allows to describe the convolutional neural network and its parameters for the launch in 2 files of the prototxt format. Caffe framework can be integrated into projects written in Python. Also this library allows to use implemented and ready algorithms of machine learning. Nowadays, Caffe framework is used to solve problems on the prediction of gender or age of a person by face images [1], as well as the detection of various objects from satellite images [8].

As shown on fig. 7, the developed neural network consists of 4 convolutional layers (conv), 4 layers with the ReLU activation function, 4 layers realizing the local normalization process (LRN), and also from 3 layers describing the process of sampling using the maxpooling operation, one fully connected layer and one softmax layer. The size of every layer, their type and the order in this network were selected in virtue of the size of input image and the deep convolution neural network proposed in [9].

Patches of the size 122 × 122 × 5 pixels that are randomly cut from grayscale images of resolution 128 × 128 pixels are fed at the input of the neural network. On the first convolutional layer, 10 kernels of the size 3 × 3 × 5 are applied to the patches with a pixel step of 1. After applying the ReLU activation function, local normalization and maxpooling operation with a 2 × 2 filter and pixel step of 2, the features will have 10 × 60 × 60 size. A similar sequence of operations is alternately performed one by one for 20 kernels of 5 × 5 × 10 size, 40 kernels of 7 × 7 × 20 size and 80 kernels of 11 × 11 × 40 size. The values of features on every maxpooling layer of developed convolution neural network for the example of input image are shown on Fig. 8. After all the operations done, a vector of 80 features is obtained. After that it is transferred to the fully connected layer. This layer allows to receive the vector of features, which consists of probability values belong to each class. This vector allows correlating the original image with some class in accordance with the values of the counted features by means of softmax layer. In other words, the neural network classifies the input image to the facial expression detected on the person's face.

The approach based on deep convolutional neural network has great computational complexity. To accelerate the neural network operation, the training and testing processes were performed on a large number of independent streams on GPU. For this, the parallel computing technology NVIDIA CUDA was used. CUDA allows solving the problem of high resource consumption of algorithms by parallelizing the computations on GPU. This technology is cross-platform and is supported by all modern NVIDIA graphics cards.

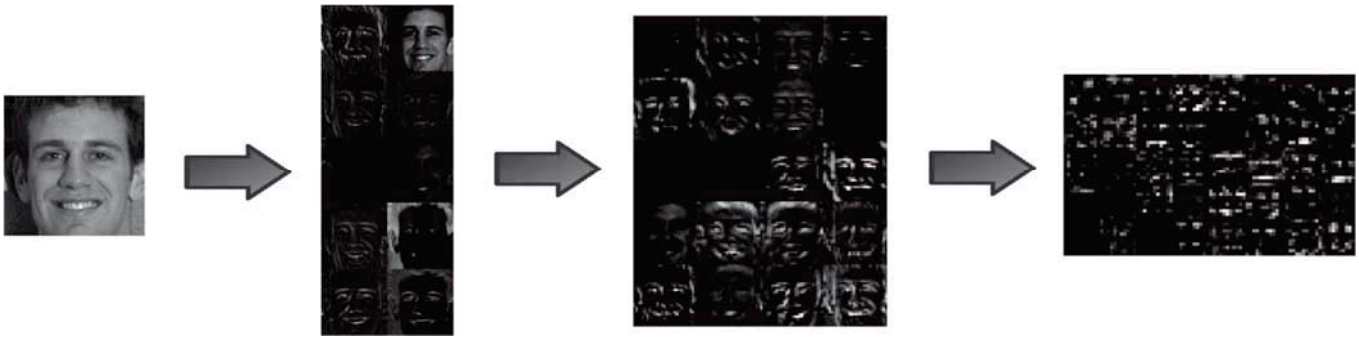


Fig. 8. Values of features on maxpooling layers of convolution neural network for the input image

In addition to the file describing the structure of a deep convolutional neural network, a file describing the values of some parameters is also needed to train the model developed on the Caffe. The developed convolutional neural network was launched on the graphic processor of the video card. The learning rate was fixed and set equal to 10^{-2}. As a numerical optimization algorithm, a stochastic gradient descent (SGD) was chosen using momentum equals to 0.9. For the regularization of the model, the weight updating rule was applied in the learning process, as a result with a weight decay equals to 5×10^{-4} was added. The classifier ended training after completing 60,000 iterations.

The developed convolutional neural network was used for smile detection and facial expression recognition by face images.

IV. Results

Numerical experiments of efficiency evaluation for developed algorithm were performed for the images of Multi-Pie database [20]. For an experiment there was prepared an uniform sampling of 30000 randomly selected images (5000 images for each of 6 facial recognition presented in Multi-Pie database) with different lighting of scene and angle of view less than 45°. These images were labeled according to one of 6 facial expression. On each picture from the sampling there was cut the face image of 128 × 128 size and transformed into black-and-white mode. Such conversion was implemented by means of computationally effective PICO algorithm.

This set of images was divided on train and test samples in the ratio 80/20. Train and test samples did not have same pictures. Labels of images from samples held in text file. There were considered 2 classes (smile and non-smile) in the case of smile detection and 6 classes (neutral, smile, surprise, squint, disguist, scream) in the case of facial expression recognition accordingly.

The learning and testing processes for the deep convolution neural network was carried out on the remote server with eight-core processor AMD FX 8320, 16 Gb RAM, the videocard NVIDIA GeForce GTX 980 and lasted around 5-7 minutes. The results of numerical experiments cite in Table I.

TABLE I. Testing results of convolution neural network

Task	Accuracy (A)
Facial expression recognition	84,98%
Smile detection	94,73%

An accuracy (A) of classifier was calculated according to the following formula:

$$A = \frac{P}{N},$$

where P is a quantity of right classified images and N is the size of test sample. For every task there were given confusion-matrices, which allow to evaluate the quality of classifiers based on deep convolution neural network.

TABLE II. Confusion-matrix for smile detection

Classes		Real class	
		Smile	**Non-smile**
Predicted class	**Smile**	2948	264
	Non-smile	52	2736

As shown in Table II, for smile detection, false-positive and false-negative examples are insignificant. Type 2 errors are more than type 1 errors. The metrics of classifier's quality in Table III, precision (P), recall (R) and F-score (F), for two classes, smile and non-smile, were calculated by formulas

$$P = \frac{TP}{TP + FP}, \qquad R = \frac{TP}{TP + FN}, \qquad F = 2\frac{P \times R}{P + R},$$

where TP is the value of true-positive examples, FN is type 1 error and FP is type 2 error.

TABLE III. Error analysis for smile detection

Error analysis		Metrics		
		Precision (P)	**Recall (R)**	**F-score (F)**
Classes	**Smile**	0,98	0,92	0,95
	Non-smile	0,91	0,98	0,94

Also in the case of smile detection there was built ROC curve and calculated its quantitative interpretation - AUC value. To construct ROC-curve there is operated with the fraction of true-positive results (TPR), the fraction of false-positive results (FPR) and the fraction of false-negative results (FNR), which are calculated by formulas

$$TPR = \frac{TP}{TP + FN}, FPR = \frac{FP}{FP + TN}, FNR = \frac{FN}{TP + FN},$$

where TN is the value of true-negative examples [22]. The ROC curve chart is shown on Fig. 9.

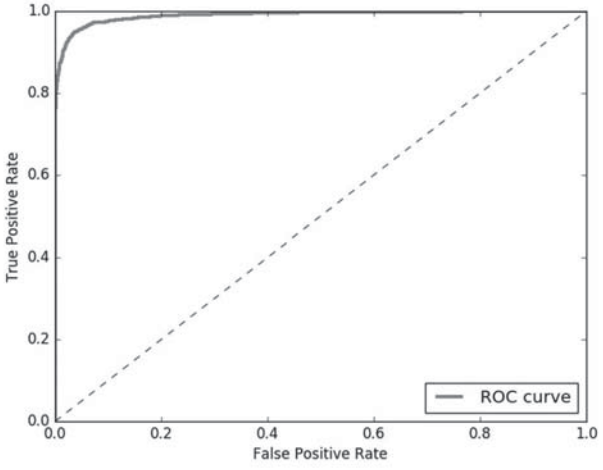


Fig. 9. ROC-curve for smile detection

Thus, the learned convolution neural network copes with smile detection by face image well enough: the value of accuracy is around 95% and AUC-ROC value is approximately equal to 0,98.

This classifier works a bit worse for facial expression recognition. It's confirmed by values of metrics, such as precision, recall and F-score, which were given due to Table IV after the testing process for deep convolution neural network.

The values of metrics from Table V were calculated by following formulas:

$$A_c = \frac{N_c}{N}, \quad P_c = \frac{A_{c,c}}{\sum_{i=1}^{n} A_{c,i}}$$

$$R_c = \frac{A_{c,c}}{\sum_{i=1}^{n} A_{i,c}}, \quad F_c = 2\frac{P_c \times R_c}{P_c + R_c},$$

where A_c is a fraction of right classified examples for class c, $A_{i,j}$ is an element of confusion-matrix of $n \times n$ size and P_c, R_c, F_c are values of precision, recall and F-score for class c accordingly.

TABLE IV. CONFUSION-MATRIX FOR FACIAL EXPRESSION RECOGNITION

Classes		Real class					
		Smile	**Surprise**	**Disgust**	**Squint**	**Scream**	**Neutral**
Predicted class	**Smile**	944	32	34	87	0	75
	Surprise	15	931	0	0	3	2
	Disgust	10	2	856	314	6	50
	Squint	5	0	78	534	1	27
	Scream	2	26	7	13	990	2
	Neutral	24	9	25	52	0	844

TABLE V. ERROR ANALYSIS FOR FACIAL EXPRESSION RECOGNITION

Error analysis		Metrics		
		Precision (P)	**Recall (R)**	**F-score (F)**
Classes	**Smile**	0,94	0,81	0,87
	Surprise	0,93	0,98	0,95
	Disgust	0,86	0,69	0,77
	Squint	0,53	0,83	0,65
	Scream	0,99	0,95	0,97
	Neutral	0,84	0,88	0,86

As shown in Table IV, V, the worst classified facial expression is "Squint". The classifier often confused them with class "Disgust". It can be explained as follows: in Multi-Pie the images with these types of facial expressions were hardly distinguishable, as shown on Fig. 10.

a) Disgust b) Squint

Fig.10. Example of images from the sampling with hardly distinguishable facial expressions

However, this learned deep convolution neural network copes with facial expression recognition by images well: the value of accuracy is around 85%.

V. CONCLUSION

In the end, it is needed to notice, that this introduced algorithm is quite simple for its implementation. This model allows to detect smiles and recognize facial expressions. The testing process shows quite good results: the accuracy of classifier is 84,98% and 94,73% according to the task. For smile detection, type 1 and 2 errors are insignificant. For facial expression recognition 4 from 6 types of emotions (smile, surprise, scream and neutral) classify quite well: the values of F-score from each of these classes are more than 0,86.

The realized algorithm can be improved due to more qualitative samples, using other databases, adding new layers or performance analysis of already existing layers in deep convolution neural network. Thereafter it is planned, that this algorithm will be plugged in cameras and tested on real images.

ACKNOWLEDGMENT

This work was supported by Russian Foundation for Basic Research grant №15-07-08674 and UMNIK-NTI "Development of algorithms for predicting individual behavior based on visual recognition of emotions", contract №0033562, agreement №11320GU/2017.

REFERENCES

[1] I. Goodfellow, Y. Bengio, A. Courville, *Deep Learning*, The MIT Press, 2016, 800 p.

[2] V. Khryashchev, I. Nenakhov, A. Lebedev, A. Priorov, "Evaluation of Face Image Quality Metrics in Person Identification Problem", *Proceedings of the 19thConference of Open Innovations Association FRUCT'19*, Jyvaskyla, Finland, 7-11 November 2016, pp. 80–87.

[3] V. Khryashchev, L. Shmaglit, A. Shemyakov, "The Application of Machine Learning Techniques to Real Time Audience Analysis System". *In: Favorskaya M., Jain L.C. (eds.) Computer Vision in Control Systems-2*, Intelligent Systems Reference Library, vol. 75, Springer International Publishing, Switzerland, 2015, pp. 49–69.

[4] Y. Lewenberg, Y. Bachrach, S. Shankar, A. Criminisi, "Predicting Personal Traits from Facial Images using Convolutional Neural Networks Augmented with Facial Landmark Information", *Proceedings of the 13th AAAI Conference on Artificial Intelligence*, 2016, pp. 4365–4366.

[5] Y. Tian, T. Kanade, J.F. Cohn, "Facial Expression Recognition", *In Li S.Z., Jain A.K. Handbook of Face Recognition. Second Edition*, Springer-Verlag London Limited, 2011, pp. 487–519.
[6] A. Lebedev, V. Pavlov, V. Khryashchev, O. Stepanova, "Face Detection Algorithm Based on a Cascade of Ensembles of Decision Trees", *Proceedings of the FRUCT'18*, Saint-Petersburg, Russia, 18-22 April 2016, pp. 161–166.
[7] P. Flach, *Machine Learning: The Art and Science of Algorithms That Make Sense of Data*, Cambridge University Press, 2012, 409 p.
[8] S. Raschka, *Python Machine Learning*, Packt Publishing Ltd., 2015, 454 p.
[9] Z. Niu, M. Zhou, L. Wang, X. Gao, G. Hua, "Ordinal Regression with Multiple Output CNN for Age Estimation", *2016 IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 4920–4928.
[10] S. Paisitkriangkrai, J. Sherrah, P. Janney, A. Van-Den Hengel, "Effective Semantic Pixel labeling with Convolutional Networks and Conditional Random Fields", *2015 IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2015, pp. 36–43.
[11] G. Ford, *Fully automatic coding of basic expressions from video. Technical Report INC-MPLab-TR-2003.03*, Machine Perception Lab., Institute for Neural Computation, University of California, San Diego, 2002.
[12] H. Khalajzadeh, M. Mansouri, M. Teshnehlab, "Face Recognition using Convolutional Neural Network and Simple Logistic Classifier", *In: Snášel V., Krömer P., Köppen M., Schaefer G. (eds) Soft Computing in Industrial Applications. Advances in Intelligent Systems and Computing*, vol. 223. Springer, Cham, 2013, pp. 197 – 207.
[13] Z. Wen, T. Huang, "Capturing subtle facial motions in 3d face tracking", *Proceedings of the 9th IEEE International Conference on Computer Vision*, 2003.
[14] H. Singh, R. Patel, "Facial Expression Analysis using Deep Learning", *International Research Journal of Engineering and Technology*, vol. 4, issue 10, 2017, pp. 66 – 69.
[15] S. Alizadeh, A. Fazel, "Convolutional Neural Networks for Facial Expression Recognition", *CoRR*, abs/1704.06756, 2017.
[16] The Cohn–Kanade AU-Coded Face Expression Database. Web: http://www.pitt.edu/~emotion/ck-spread.htm.
[17] The Japanese Female Facial Expression (JAFFE) Database. Web: http://www.kasrl.org/jaffe.html.
[18] The MMI Facial Expression Database. Web: https://mmifacedb.eu.
[19] The Radboud Faces Database. Web: http://www.socsci.ru.nl:8180/RaFD2/RaFD?p=main.
[20] The CMU Multi-PIE Face Database. Web: http://www.cs.cmu.edu/afs/cs/project/PIE/MultiPie/Multi-Pie/Home.html.
[21] Caffe Framework. Web: http://caffe.berkleyvision.org.
[22] Scikit-learn. Web: http://scikit-learn.org/stable/.