Applied Mathematics & Information Sciences
An International Journal

http://dx.doi.org/10.18576/amis/130520

Improved Facial Expression Recognition with Xception Deep Net and Preprocessed Images

Maksat Kanatov [1,*], *Lyazzat Atymtayeva*[2] *and Mateus Mendes*[3]

[1] Kazakh-National Research Technical University after K.I.Satbayev, Almaty, Kazakhstan
[2] Suleyman Demirel University, Kaskelen, Kazakhstan
[3] Polytechnic Institute of Coimbra ESTGOH, Institute of Systems and Robotics of the University of Coimbra, Portugal

Received: 2 May 2019, Revised: 12 Aug. 2019, Accepted: 27 Aug. 2019
Published online: 1 Sep. 2019

Abstract: Automated Facial Expression Recognition (FER) is an important part of computer-human interaction. For decades, researchers and scientists have been trying to create a model of artificial intelligence that could think, learn, make decisions and act in a way similar to a real person. Among other skills, such model needs to recognise human facial expression to understand non-verbal language. The present paper describes a method to fine tune the FER process in images, using deep learning CNN model Xception, with preprocessing the images. The method has shown improved results when applied to different datasets.

Keywords: Convolutional neural networks, deep learning, computer vision, facial expression recognition

1 Introduction

Facial Expression Recognition (FER) is an actively researched area in computer vision for the last two decades. FER is a very important aspect in human-machine interaction. Humans demonstrate their emotional state using facial expressions, using the verbal method, gestures of hands and posture of the body. However, facial expressions are the most important part of these.

Research in facial expressions started hundreds of years ago [1], [2]. However, Automated Facial Expression Recognition (AFER) for computer vision appeared only in the last twenty years. At first there were works on the analysis of facial expressions [4], [6]. Then the first attempts were made to create AFER [7], [8], [9], [10]. Much earlier, French neurologist Duchenne de Boulogne conducted experiments when he was trying to find dependencies between facial muscles and emotions [1]]. Duchenne published his work with extraordinary photographs in the book "The Mechanism of Human Facial Expressions" in 1862. Figure 1 shows some examples of these photographs. Additionally, Charles Darwin worked on FE [2] at the same period of time. His work inspired other researchers like C. Izard and P. Ekman. Subsequently they contributed substantially to the body of acailable literature analyzing FE [3], [4]. Ekman and Friesen proposed six main universal emotions [5]. which can be common for all cultures [12]. They established the following facial expressions: disgust, fear, happiness, surprise, sadness and anger. This approach is widely used in Automated FER tasks. Using a discrete concept of emotions, Plutchik proposed his own model [13], which consists of eight basic emotions: joy, trust, surprise, anticipation, sadness, fear, anger and disgust. However, he adopted a new approach, where emotions can be mixed. He proposed that eight basic emotions could be mixed into twenty four new emotions. However, later Russel [11] proposed a new dimensional model, where emotions can be mapped into a two-dimensional space. The main idea of this model is demonstrated in Figure 2. This model consists of two features. The vertical line is arousal, which defines how sleepy or alert the emotion is. The horizontal line is a valence which defines if the emotion is negative or positive. The present work proposes results of experiments of an AFER system based on a fine-tuned Xception [32] architecture using image preprocessing procedures.

Section 2 describes available facial expression datasets, and datasets which are used in this work. Section 3 demonstrates the steps of preprocessing and data augmentation operations . Sections 4 and 5 demonstrate

* Corresponding author e-mail: maksatkanat@gmail.com

Fig. 1: Experiments of Duchenne de Boulogne, described in [1].

the construction of the used Xception architecture and the fine tuned hyperparameters of the optimizer. Section 6 demonstrates the results of training and validation steps. Section 7 concludes present works and describes perspectives and future work.

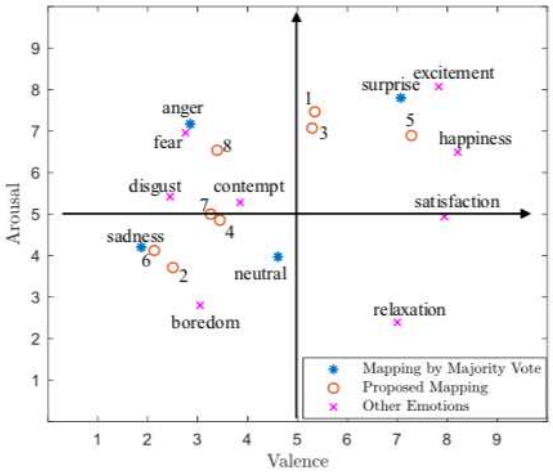

Fig. 2: Russel's dimensional emotion model [14].

2 Facial Expression Datasets

There are many available datasets for FER. These datasets can be divided into several categories and groups. First of all, datasets can be divided into three groups: RGB and GrayScaled (JAFFE [7], CK [16], CK+ [19], Oulus-CASIA [22], etc.), 3D (BU-3DFE [28]) and Thermal (Oulus-CASIA [22]). RGB and GrayScaled datasets are the most popular types, and they are widely used in the last two decades. These types can consist of posed or natural samples. Posed means that all samples are collected in the laboratory, using special camera position and distances for good images of facial emotions. Normally, in posed datasets images have a good illumination and resolution. Conversely, in the datasets which consist of natural samples, faces can be of any resolution, different illumination and at any angle relative to the camera. Regularly, natural images are collected from movies. Table 1 shows one list of datasets for FER, with some details.

Fig. 3: Samples from CK+ dataset.

Present work is developed using Extended Cohn-Kanade (CK+) [19] and Karolinska Directed Emotional Faces (KDEF) [30] datasets. The first version of Cohn-Kanade dataset was created by Kanade et al. [16]. Images for this database were collected from 210 adults between the ages eighteen and fifty years with different nationalities. Ten years later, Kanade et al. created a new database, which is named The Extended Cohn-Kanade dataset (CK+) [19]. The original distribution of posed facial expressions was extended from 486 Facial Action Coding Systems (FACS) sequences to 593 sequences, by adding another 113 sequences. Each sequence starts from neutral and goes to one of the 6 emotions. We have used the first image of the sequence for neutral emotion and the last four images for each emotion.
The second dataset that is mostly used is KDEF [30] . KDEF was created from seventy actors using five different angles and with six basic and neutral facial expressions. Originally, this dataset was proposed for medical and psychological purposes. From this dataset only frontal images were used. In total, 2,893 images with frontal facial expressions are collected from the two datasets. For normalizing training data, the present work proposes preprocessing methods and augmentation operations, which are described in Section 4.

3 The Training Process

The present work proposes preprocessing methods, which improve the learning process. The first one is face detection and alignment. Typically, for FER tasks, it is desirable only to select the image of the face. Nevertheless, the datasets contain of a lot unnecessary information. For example, samples of CK+ dataset which are showed in Figure 3, contain a large amount of unnecessary information such as background, body, hair,

Table 1: List of selected datasets of Facial Expressions (i - images, v - videos, s - sequences of images, b - basic emotions, n - neutal emotion, GS - grayscale, NIR - Near Infra Red).

Name	Year	Samples	Emotions	Source	Type
JAFFE [7]	1998	213 i	6 b + 1 n	Posed	GS
KDEF [30]	1998	4900 i	6 b + 1 n	Posed	RGB
CK [16]	2000	486 s	6 b + 1 n	Posed	RGB + GS
MMI [17]	2005	740 i, 2900 v	6 b + 1 n	Posed	RGB
Multi-PIE [31]	2010	755,370 i	5 b + 1 n	Posed	RGB
TFD [29]	2010	112,234 i	6 b + 1 n	Posed	GS
CK+ [19]	2010	593 s	6 b + 1 n	Posed	RGB + GS
Oulu-CASIA [22]	2011	2,880 s	6 b	Posed	RBG + NIR
SFEW [20]	2011	1,766 i	6 b + 1 n	Movie	RGB
AFEW [20]	2011	1,809 v	6 b + 1 n	Movie	RGB
FER-2013 [23]	2013	35,887 i	6 b + 1 n	Posed + Web	GS
EmotioNet [25]	2016	1,000,000 i	23 b	Web	RGB
BU-3DFE [28]	2016	2,500 i	6 b + 1 n	Posed	3D
RAF-DB [26]	2017	1,608 i	6 b + 1 n	Posed	RGB
AffectNet [27]	2017	450,000 i	6 b + 1 n	Web	RGB

etc. To get only the face, we used Viola-Jones algorithm [15] and cut the face from the original image. In the next step our system normalizes the position of the face, using Face Alignment [33]. This Face Alignment system gives sixty eight landmarks of the face in a short time. Using three reference landmarks, the face is transformed, as shown in Figure 4. After this preprocessing method, all faces are normalized, and have the same position. At the end, the output image is grayscaled and resized into 256×256 pixels. After normalizing, data augmentation functions are called. Real-time data augmentation is used, while training steps are going. However, only few

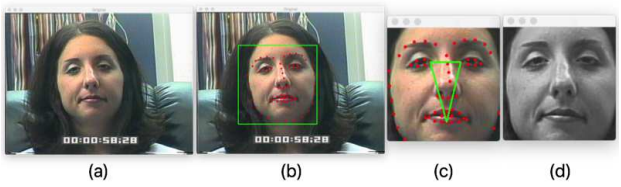

Fig. 4: Preprocessing steps used in this work. (a) - input image, (b) - face detection and 68 landmarks, (c) - transformed face into reference points, (d) - grayscaled output image.

operations are used for data augmentation. They are mirroring, random zoom in the range +/- 10%, random rotation in +/- 5 degrees and random brightness change in the range +/- 20%.

Figure 5 demonstrates some examples after data augmentation procedures. During training, we have called the data augmentation function fifty times per epoch with a batch size of thirty two. More details about training parameters are in Section 5.

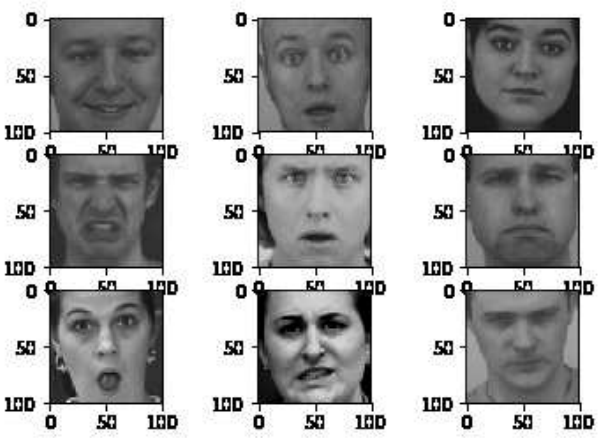

Fig. 5: Images after data augmentation procedures.

4 Using Deep Convolutional Neural Networks and Hyperparameter tuning

The present work is based on the new Deep CNN model, called Xception [32] which was proposed by Francois Chollet. Xception was inspired by Inception V3 [34] and it has the same number of training parameters. However, Xception has a different construction. Inception modules are changed to deep separated convolutions. As result, Xception is more efficient than Inception V3 model. Figure 6 illustrates Xception convolution module. More information about this architecture can be found in [32]. Xception architecture consists of 20,861,480 parameters, where 20,806,952 parameters are trainable. Additionally, this model has thirty six conv-layers. These layers are built-in into fourteen modules. Also, they have residual

connections [35], which are located around them. However, the first and last modules do not have these connections. We have used three different optimizers in the present work. They are Adadelta [36], RMS Prop [37] and Adam Optimizer [38]. Adadelta Optimizer is used with initial learning rate 1.0, decay factor (rho) 0.95, and fuzz factor (epsilon) 10^{-6}. RMS Prop with initial learning rate 0.001, decay factor(rho) 0.94. Epsilon is not used in the present work. Adam Optimizer is used with initial learning rate 0.001, beta one 0.9, beta two 0.999. Training is done on two Nvidia Quadro M2000 GPUs and it is repeated for each optimizer with 50 training epochs.

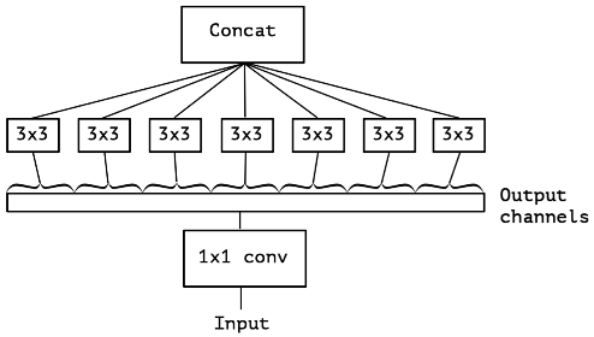

Fig. 6: Xception module with one spatial convolution per output channel of the 1×1 convolution.

5 Experiments

Different experiments are performed in order to determine the best preprocessing methods to use with fine-tuned Xception architecture with the preprocessing procedures. Results of experiments without preprocessing procedures are also shown, as a reference. Details of preprocessing are described in Section 3. For validation of the training model, it uses the most popular ten-fold validation protocol. Firstly, all data is shuffled, and after that it is divided into ten groups. Each experiment was repeated ten times, where eight groups of data are used for training. The remaining two are used for validation and testing. Results, which are presented in Section 4, are the average values of ten iterations. The present work shows results based on several datasets. The main dataset is Extended Cohn-Kanade dataset [19]. Results of this work are compared with other related works using this dataset. However, JAFFEE [7] and KDEF [30] datasets are used for further evaluation of Xception DeepNets.

6 Results

As described above, we have trained Xception DeepNets model using JAFFE, KDEF and Extended Cohn-Kanade datasets with three types of fine-tuned optimizers. The Xception model is also trained without preprocessing on Extended Cohn-Kanade dataset. Table 2 shows the results, where proposed preprocessing procedures improve performance by about seven percent. The best result is achieved using RMS Prop Optimizer. This Optimizer shows best results on JAFFE and KDEF1 datasets, where performances are 94.84 % and 96.72 % respectively. Additionally, three datasets were mixed, trained and tested in the present system with three Optimizers. Figure 7 shows the progress of losses and performance, during one of the ten training episods. All three optimizers have similar results, with small exclusions. Training and validation results have approximately the same values. Only after 30 epochs validation performances are stabilized and training test results are going to small over-fitting (Figure 7). After ten-fold iteration, accuracy of the trained model using Adadelta Optimizer is 86.9 % with loss 0.459 on the test set. The model trained by Adam Optimizer shows the lowest result, 84.8 % accuracy and 0.614 of loss. The leader of this experiment is RMS Prop Optimizer again. The model trained using this Optimizer shows 88.3 % of accuracy and 0.446 of loss. Table 3 illustrates comparison with other latest related works.

Table 2: Averaged results after 10 fold testing on CK+ dataset using Xception model with and without preprocessing procedures.

Model	Optimizer	Error	Accuracy
Xception	Adadelta	0.207	89.17 %
Xception	RMS Prop	0.187	91.26 %
Xception	Adam	0.152	91.88 %
Xception + preprocessing	Adadelta	0.061	96.81 %
Xception + preprocessing	RMS Prop	0.014	98.96 %
Xception + preprocessing	Adam	0.062	97.73 %

Table 3: Comparison with related works.

Model	Data Group	Accuracy
AlexNet [40]	LOSO	94.40 %
DSAE [41]	LOSO	95.79 %
(N+M)-tuplet clusters loss [42]	8 folds	97.10 %
Xception + preprocessing	**10 folds**	**98.96 %**

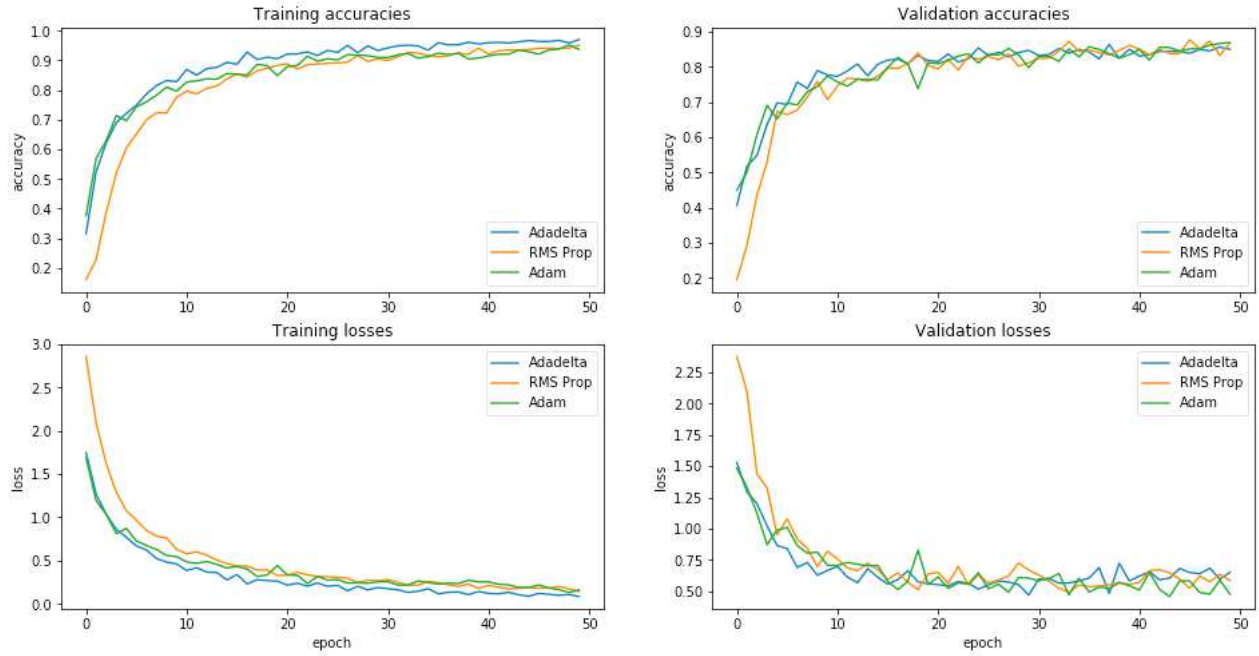

Fig. 7: One of the ten training steps, based on three mixed datasets.

7 Perspective

Xception is one of the most powerful deep network models. We have a best result with accuracy 98.96% on extended Cohn-Kanade Dataset. Xception with preprocessing procedures, proposed in this work, and fine-tuned optimizer shows good results. Xception is a very deep model of convolutional neural networks. Training can take a lot of time. However, as it was described above, a properly configured and trained model can cope with complex tasks such as FER. Also, the present work proves the superiority of Deep CNN on image recognition tasks. We plan to continue this work using spontaneous facial expression dataset AffectNet [27] and FER2013 [23].

Acknowledgement

This research was supported by grant of the program of Ministry of Education of the Republic of Kazakhstan BR05236699 "Development of a digital adaptive educational environment using Big Data analytics". We thank our colleagues from Suleyman Demirel University(Kazakhstan) who provided insight and expertise that greatly assisted the research. We express our hopes that they will agree with the conclusions and findings of this paper.

References

[1] G.-B. D. de Boulogne and R. A. Cuthbertson, *The Mechanism of Human Facial Expression*, Cambridge University Press, (1990).
[2] C. Darwin, *The expression of emotion in man and animals*, Oxford University Press, (1872).
[3] C. E. Izard, *The face of emotion*, (1971).
[4] P. Ekman, Universal and cultural differences in facial expression of emotion, *Nebr. Sym. Motiv.*, **19**, 207-283, (1971).
[5] P. Ekman and W. V. Friesen, Constants across cultures in the face and emotion, *1. Pers. Soc. Psycho!.*, **17**, 124-129, (1971).
[6] Ekman P., Facial expression and emotion, *American psychologist*, **48**, 384, (1993).
[7] Lyons M. et al., *Coding facial expressions with gabor wavelets*, Proceedings Third IEEE international conference on automatic face and gesture recognition, **IEEE**, 200-205, (1998).
[8] Mase K., Recognition of facial expression from optical flow, *IEICE transactions (E)*, **74**, 3474-3483 (1991).
[9] Tian Y. I., Kanade T., Cohn J. F., Recognizing action units for facial expression analysis, *IEEE Transactions on pattern analysis and machine intelligence*, **23**, 97-115, (2001).
[10] Bartlett M. S. et al., *Real Time Face Detection and Facial Expression Recognition: Development and Applications to Human Computer Interaction*, 2003 Conference on computer vision and pattern recognition workshop, 5, (2003).
[11] J. A. Russell, A circumplex model of affect. *Journal of personality and social psychology*, **39**, 161-178, (1980).
[12] P. Ekman, An argument for basic emotions. *Cognition and Emotion*, **6**, 169-200, (1992).

[13] R. Plutchik, Nature of emotions. *Am. Sci.*, **89**, 344-350, (2002).
[14] Zhou, F., Kong, S., Fowlkes, C., Chen, T., Lei, B., Fine-Grained Facial Expression Analysis Using Dimensional Emotion Model. *arXiv preprint arXiv:1805*, **01024**, (2018).
[15] P. Viola and M. J. Jones, Robust real-time face detection, *International journal of computer vision*, **57**, 137-154, (2004).
[16] T. Kanade, J. F. Cohn, and Y. Tian, *Comprehensive database for facial expression analysis*, Proceedings Fourth IEEE International Conference on Automatic Face and Gesture Recognition (Cat. No. PR00580), **IEEE**, 46-53, (2000).
[17] M. Pantic, M. Valstar, R. Rademaker and L. Maat, *Web-based database for facial expression analysis*, IEEE International Conference on Multimedia and Expo, (2005).
[18] M. Pantic and I. Patras, Dynamics of facial expression: recognition of facial actions and their temporal segments from face profile image sequences, *IEEE Transactions on Systems, Man, and Cybernetics*, **36**, 433-449, (2006).
[19] P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, Z., Ambadar, and I. Matthews, *The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression*, 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition-Workshops, 94-101, (2010).
[20] Dhall A. et al., Acted facial expressions in the wild database, *Australian National University, Canberra, Australia, Technical Report TR-CS-11.*, **2**, (2011).
[21] A. Dhall, R. Goecke, S. Lucey, and T. Gedeon, *Static facial expression analysis in tough conditions: Data, evaluation protocol and benchmark*, 2011 IEEE International Conference on Computer Vision Workshops (ICCV Workshops), 2106-2112, (2011).
[22] Zhao, G., Huang, X., Taini, M., Li, S. Z., and Pietikalnen, M., Facial expression recognition from near-infrared videos, *Image and Vision Computing*, **29(9)**, 607-619, (2011).
[23] I. J. Goodfellow, D. Erhan, P. L. Carrier,et al., *Challenges in representation learning: A report on three machine learning contests*, International Conference on Neural Information Processing, 117-124, (2013).
[24] Benitez-Quiroz, C. F., Srinivasan, R., and Martinez, A. M., *EmotioNet: An Accurate, Real-Time Algorithm for the Automatic Annotation of a Million Facial Expressions in the Wild*, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 5562-5570, (2016).
[25] Benitez-Quiroz, C. F., Srinivasan, R., Feng, Q., Wang, Y., and Martinez, A. M., EmotioNet Challenge: Recognition of facial expressions of emotion in the wild, *arXiv preprint arXiv:1703*, **01210**, (2017).
[26] Li, S., Deng, W., and Du, J., *Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild*, Proceedings of the IEEE conference on computer vision and pattern recognition, 2584-2593, (2017).
[27] Mollahosseini, A., Hasani, B., and Mahoor, M. H., Affectnet: A database for facial expression, valence, and arousal computing in the wild. arXiv preprint arXiv:1708.03985. 2017.
[28] Lijun Yin, L. Yin, X. Wei, Y. Sun, J. Wang, and M. J. Rosato, *A 3D Facial Expression Database For Facial Behavior Research*, 7th In- ternational Conference on Automatic Face and Gesture Recognition, (2016).
[29] J. M. Susskind, A. K. Anderson, and G. E. Hinton, *The toronto face database*, Department of Computer Science, University of Toronto, Toronto, ON, Canada, Tech. Rep, vol. 3, (2010).
[30] D. Lundqvist, A. Flykt, and A. Ohman, *The karolinska directed emotional faces (KDEF)*, CD ROM from Department of Clinical Neu- roscience, Psychology section, Karolinska Institutet, no. 1998, (1998).
[31] R. Gross, I. Matthews, J. Cohn, T. Kanade, and S. Baker, Multi-pie, *Image and Vision Computing*, **28(5)**, 807-813, (2010).
[32] Chollet, F., *Xception: Deep learning with depthwise separable convolutions*, Proceedings of the IEEE conference on computer vision and pattern recognition, 1610-02357, (2017).
[33] Kazemi, V., and Sullivan, J. (2014). One millisecond face alignment with an ensemble of regression trees. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 1867-1874).
[34] C.Szegedy,V.Vanhoucke,S.Ioffe,J.Shlens,andZ.Wojna, Rethinking the inception architecture for computer vision, *arXiv preprint arXiv:1512.00567*, (2015).
[35] K. He, X. Zhang, S. Ren, and J. Sun, Deep residual learning for image recognition, *arXiv preprint arXiv:1512.03385*, (2015).
[36] Zeiler, M. D., ADADELTA: an adaptive learning rate method. *arXiv preprint arXiv:1212.5701*, (2012).
[37] T Tieleman and G Hinton, *Rmsprop: Divide the gradient by a running average of its recent magnitude*, COURSERA: Neural Networks for Machine Learning. Technical Report, **31**, (2012).
[38] Kingma, D. P., and Ba, J., Adam: A method for stochastic optimization, *arXiv preprint arXiv:1412.6980*, (2014).
[39] Kanatov M., Atymtayeva L., Deep Convolutional Neural Network based Person Detection and People Counting System, *Advanced Engineering Technology and Application*, **7(3)**, 21-25, (2018).
[40] S. Ouellet, Real-time emotion recognition for gaming using deep convolutional network features, *arXiv preprint arXiv:1408.3750*, (2014).
[41] N. Zeng, H. Zhang, B. Song, W. Liu, Y. Li, and A. M. Dobaie, Facial expression recognition via learning deep sparse autoencoders, *Neurocomputing*, **273**, 643-649, (2018).
[42] X. Liu, B. Kumar, J. You, and P. Jia, *Adaptive deep metric learning for identity-aware facial expression recognition*, Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW), 522-531, (2017).

Maksat Kanatov master of Computer science, researcher at Kazakh National Research Technical University after K. I. Satpayev in Almaty, Kazakhstan. His research interest includes Machine Learning, Computer Vision, Deep Convolutional and Recurrent Neural Networks, Automated Facial Expression Recognition systems using Deep CNNs.

Lyazzat Atymtayeva received the Ph.D and Doctor of Science degree in Mechanics, Mathematics and Computer Science at Suleyman Demirel University, Kazakhstan. Her research interests are in the areas of mechanics, applied mathematics and computer science including the numerical and rigorous mathematical methods and models for mechanical engineering and computer science, intelligent and expert systems in Information Security, Project Management and HCI. She has published research papers in reputed international journals of mathematical and computer sciences. She is reviewer and editor of international journals in mathematics and information sciences.

Mateus Mendes Adjunct Professor of Higher School of Technology and Management of Oliveira do Hospital Polytechnic Institute of Coimbra, Portugal. Since 2005: Researcher at the Institute of Systems and Robotics of the University of Coimbra, member of the Computational Intelligence Lab. Member 103338 of the International Association of Engineers (IAENG). Main research interests: Artificial and Computational Intelligence, Image Processing and Language Processing.