Deep learning for real-time robust facial expression analysis

Vladimir Khryashchev
P.G Demidov Yaroslavl State University, Artificial Intelligence Center
Yaroslavl, Russian Federation
vhr@yandex.ru

Leonid Ivanovsky
P.G Demidov Yaroslavl State University, Computer Vision Laboratory
Yaroslavl, Russian Federation
leon.ivanovsky@yahoo.com

Andrey Priorov
P.G Demidov Yaroslavl State University, Computer Vision Laboratory
Yaroslavl, Russian Federation
andcat@yandex.ru

ABSTRACT
The aim of this investigation is to classify real-life facial images into one of the six types of emotions, using deep machine learning and convolutional neural network (CNN). CNN is a special architecture, which allows to recognize various objects rapidly, as well as to make an effective classification. For acceleration of learning of developed CNNs there was used supercomputer NVIDIA DGX-1. This process was implemented parallel, on a large number of independent streams on GPU. Numerical experiments for algorithms were performed on the images of Multi-Pie database with various lighting of scene and rotation of head. For developed models there were given metrics of quality. It is planned, that this algorithm will be used in real-time applications on laptops for real-time video processing, human-computer interaction and statistics gathering, in the field of retail, security etc.

Keywords
Deep machine learning; facial expression analysis; convolutional neural network; real-time videoanalysis.

1. INTRODUCTION
Humans have a wide range of different facial expressions which play an important role in everyday communication. In order to recognize the type of facial expression using the machine learning algorithm, it is necessary to identify certain classes of facial expression. Ekman et al. singled out a set of universal facial expressions that are the same for people across all cultures: anger, disgust, fear, happiness, sadness, surprise [1, 2]. Nowadays most researches aim to identify this set of universal facial expressions with the addition of a neutral emotion using computer vision. Emotion Recognition in the Wild Challenge and Kaggle's Facial Expression Recognition Challenge also present these emotions for classification [3, 4].

Automatic facial expression analysis is used in various applications. Facial expression recognition is successfully applied in robotics to create intelligent visual interface between the man and the machine. More than that, expression recognition is a topic of interest to Human-Computer Interaction research community which aims to build human-computer interfaces that are able to response to one's emotions and mood. Moreover, expression recognition find uses in such fields like Video Games, Animations, Psychiatry, Automobile Safety, Educational Software, etc [5-7].

In practical applications, a fully robust face expression recognition system must possess [2]:

- It must be fully automatic;

- It must have the capability to work with video stream as well as images;

- It must work in real-time;

- It must be able to recognize spontaneous expressions;

- It must be robust against different lighting conditions;

- It must be person independent;

- It must work on people from different cultures and different skin colors. It must also be robust against age (in particular, recognize expressions of both infants, adults and the elderly);

- It must be invariant to facial hair, glasses, makeup etc.;

- It must be able to work with videos and images of different resolutions;

- It must be able to recognize expressions from frontal, profile and other intermediate angles;

As expression recognition systems become more real-time and robust, we will be seeing many other innovative applications and uses.

The task of expression recognition is challenging [2]. Nowadays the recognition of spontaneous expressions is in the focus of research community. The major challenge is the non-availability of spontaneous facial expression data. More than that, capturing spontaneous expressions on images and video and labeling the data is a very time consuming process. The use of semisupervised learning that allows to analyze both labeled and unlabeled data can be used to solve partly this problem.

In addition to the six universal facial expressions there are a lot of other expressions that can be recognized. But capturing and recognizing spontaneous non-basic expressions is even more challenging than capturing and recognizing spontaneous basic expressions.

Another challenge is the fact that different expressions are not being recognized with the same accuracy. More than that, the results of facial expressions recognition usually differs for people of different age groups and different cultures and races. Ideal facial expression recognition systems should be robust against these changes.

The final classification can be performed it two ways: in each videoframe or for the fragment of video. One of the modern approaches to solving this problem is the use of a deep convolutional neural network [7]. The features in such network are formed automatically in the process of training.

The authors of [8] presents the different types of the architecture of the deep convolutional neural network which can be used for solving the problems of the person's gender and age prediction based on the face image analyses. In [9] the solution of the problem of classifying objects of different types in an image by the use of the convolutional neural network is described. In paper [10] there are presented two types of CNNs: five-layer and deeper. For these models there is made a comparative analysis of architectures and numerical experiments. The best performance was 48%. The papers [11, 12] describe a hybrid two-stage classifiers based on the use of deep convolutional neural network. The accuracy of the classifiers was 86.06% and 71%, respectively. The authors of [13] developed convolutional neural network for solving the task of facial expression analysis. A combination of raw pixel data and Histogram of Oriented Gradients features was used for training. The accuracy was 80.5%.

The authors of [14] achieved state-of-the-art results using convolutional neural networks to perform facial expression analysis in Emotion Recognition in the Wild Challenge. They used an ensemble of neural networks with five convolutional layers. Kim et al. [15] also achieved good results by using an ensembles of convolutional neural networks with varying architectures and parameters.

The authors of [16] proposed the network consisted of two convolutional layers, max-pooling, and four Inception layers that was tested on many public data sets. The accuracy was 47% on the EmotiW data set and state-of the-art accuracies on other data sets (93% on CK+). The results of work of this neural network outperformed the results for AlexNet architecture by 1-3 percent on most data sets.

This article consists of five parts including introduction. The second part is devoted to the image databases, which can be applied to the problem of facial expression analysis. The third section is devoted to developed CNNs and the comparison of their architectures. Also in this part there were described tools for building of classifiers, as well as the peculiarities of training process. The forth section shows the results of numerical experiments for developed models. For both networks accuracy values, confusion matrix and some metrics quality were given and compared to each other. Then, in conclusion you can find summarizing and suggestions about the application of proposed classifiers.

2. FACIAL EXPRESSION DATABASES

Databases are one the most important part for tasks of machine learning including facial expression analysis. Nowadays, there are some appropriate available databases of images.

The Binghampton University 3D Facial Expression (BU-3DFE) Database contains 2500 facial shape models of 100 subjects (56% female, 44% male), ranging age from 18 to 70 years old, with a variety of ethnic ancestries. These colour models are captured at 3 views (-45°, 0°, 45°) and divided on 7 types of emotions (neutral, happiness, disguist, fear, angry, surprise and sadness) with 4 levels of expressions, from low to high. [17]. The examples of images from the BU-3DFE Database are shown on Fig. 1.

The Multimedia Understanding Group (MUG) Facial Expression Database consists of image sequences of 86 different subjects (35 – female, 51 – male) between 20 and 35 years of age. Each sequence contains from 50 to 160 frontal colour pictures. In this database there are pictures with 6 different types of emotions: anger, disgust, fear, happiness, sadness, surprise [18]. The examples of images from the MUG Database are shown on Fig. 2.

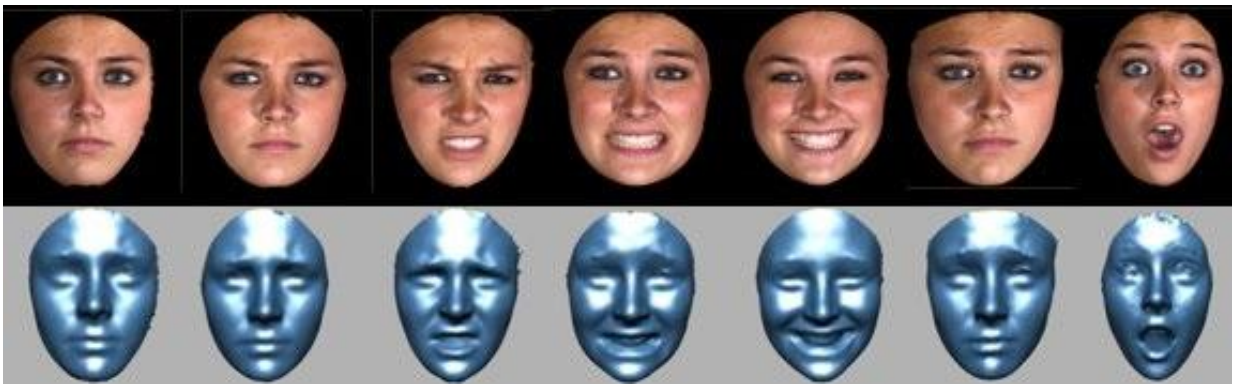

Fig. 1. Examples of images from the BU-3DFE Database

Fig. 2. Examples of images from the MUG Facial Expression Database

Fig. 3. Examples of images from the CMU Multi-Pie Face Database

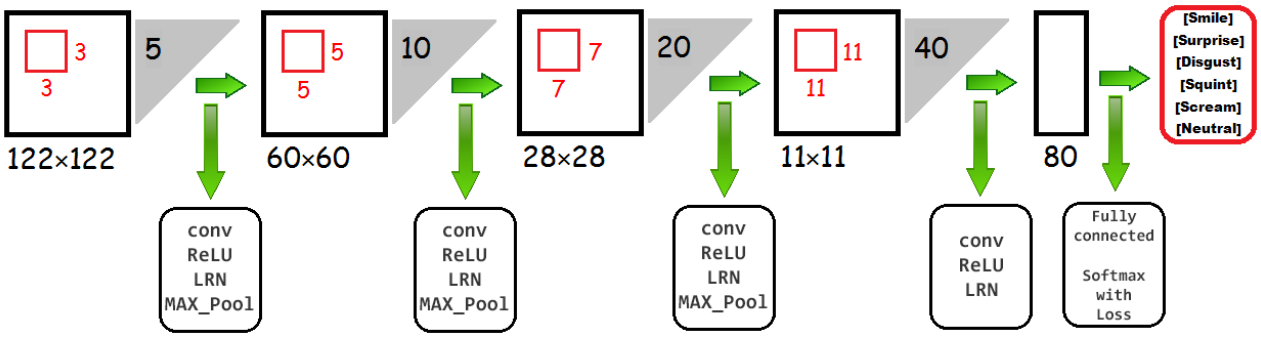

Fig. 4. Architecture of simple CNN

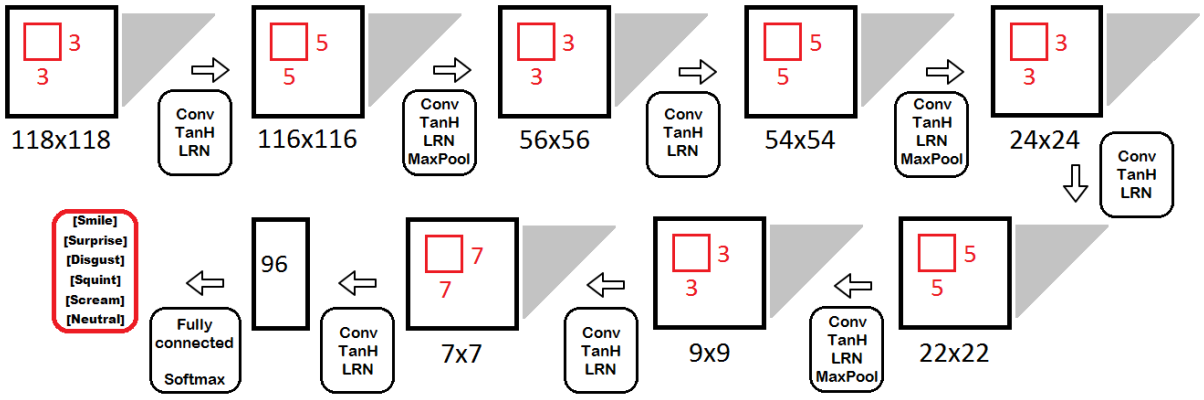

Fig. 5. Architecture of complicated CNN

The CMU Multi-PIE Face Database contains more than 750,000 color images of 6 facial expression (neutral, smile, surprise, squint, disgust, scream) from 337 subjects. The pictures were made on different angles (less than 90°) with various scene illumination [19]. The examples of images from the Multi-Pie Database are shown on Fig. 3.

3. MODELING OF CONVOLUTIONAL NEURAL NETWORK

The paper presents a developed algorithms, based on CNN, which are related to the field of deep machine learning. Conforming to results of modern research, in most problems of classification, this method shows the best results [20].

The network's architecture was written based on the deep CNN proposed in [21]. The new developed algorithm became more complicated than previous model. The architecture of simple CNN is shown on Fig. 4. As shown on Fig. 5, the new developed model consists of 8 convolutional layers (Conv), 8 layers with an activation function $y=\tanh(x)$ (TanH), 8 layers realizing the local normalization process (LRN), 3 sampling layers with the maxpooling operation (MaxPool), one fully connected layer and one softmax layer with loss. The size of every layer, their types and the order in this network were fitted hence the size of input image and the deep convolution neural network proposed in [8, 21].

Compared with algorithm in [21] in developed modernized, complicated CNN there are added more convolutional layers. It was made after the empirical study of ways to improve the quantity of model in [21]. This research shows, that sampling layers with the maxpooling operation and convolutional layers might improve the accuracy of classifier. Also there was changed the type of activation function from linear rectification ReLU $y(x)=\max(0, x)$ to hyperbolic tangent $y=\tanh(x)$.

The network's architecture was written using Caffe framework. Caffe framework allows to describe the CNN and its parameters for the launch in 2 files of the prototxt format. Also this library uses ready algorithms of machine learning [22]. Nowadays, this

framework is used to solve problems of face recognition or gender and age prediction by facial images [8].

In addition to the file describing the structure of developed algorithms, there is also needed another file describing some parameters to train the appropriate model. CNNs were launched on GPU of video card. The learning rate was set equal to 0,01. Stochastic gradient descent (SGD) was chosen as a numerical optimization algorithm. For the regularization the weight updating rule was applied in the learning process with a weight decay equals to 0,0005. Each classifier ended training after completing 70,000 iterations.

The approach based on deep neural network has high resource consumption. To accelerate the neural network operations, the training and testing processes were performed on a large number of independent streams on GPU using parallel computing technology NVIDIA CUDA. This technology is cross-platform and is supported by all modern NVIDIA graphics cards [23].

4. RESULTS

Numerical experiments for developed algorithms were performed on the images of Multi-Pie database [19]. For an experiment there was prepared a sample of 210000 randomly selected and mirrored images (35000 images for each of 6 types of emotions presented in Multi-Pie database) with various lighting of scene and angle of view less than 45°. These images were labeled according to one of 6 facial expression. These labels were saved in text file. On each picture from the sampling there was cropped the facial image of 128 × 128 size and transformed into black-and-white mode. Such conversion was implemented by means of computationally effective PICO algorithm [24].

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. Moreover, the images with the same person were not in the training and test datasets simultaneously.

The launch of the convolution neural networks was carried out on the supercomputer NVIDIA DGX-1 and lasted around 30-40 minutes. As a result of numerical experiments, 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 [25]. The results of numerical experiments cite in Table I.

Table 1. Testing results of convolutional neural networks

Model	Accuracy (A)
Simple	92,29%
Complicated	94,48%

According to Fig. 6 the complicated model converges, and the value of loss function is insignificant, which decreases with the increase of completed training iterations.

In addition, there was built a confusion-matrix (Table II), which allows to evaluate the quality of developed algorithms [26]. Each cell of this matrix conforms to the quantity of recognized pictures to appropriate class. Results in a cell are given in a form simple CNN results/ complicates CNN results.

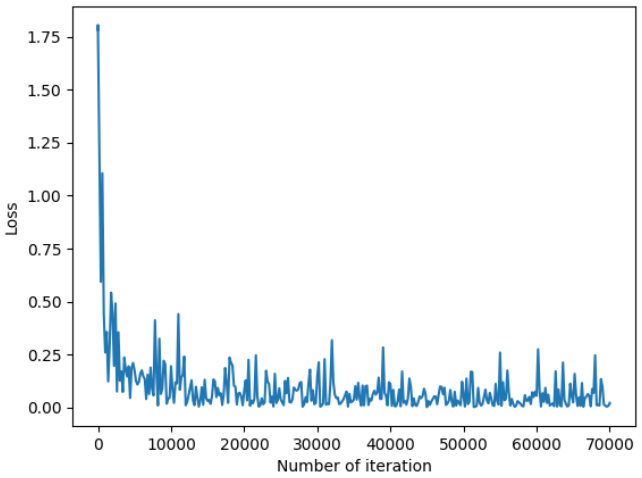

Fig. 6. Graph of loss function for complicated CNN

As shown in Table II, on the test sample both CNNs coped well the task of facial expression analysis. This assumption was proved by values of metrics, such as precision (P), recall (R) and F-score (F). These values were calculated by following formulas

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

where $A_{i,j}$ 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 [27].

Table 2. Confusion-matrix

Classes		Real class: Smile	Surprise	Disgust	Squint	Scream	Neutral
Predicted class	**Smile**	6471/ *6526*	48/ *37*	68/ *53*	8/ *6*	4/ *1*	136/ *145*
	Surprise	87/ *140*	6715/ *6767*	3/ *2*	1/ *0*	47/ *16*	30/ *8*
	Disgust	102/ *72*	28/ *2*	6006/ *6322*	556/ *334*	34/ *33*	61/ *43*
	Squint	143/ *148*	20/ *2*	749/ *541*	6112/ *6479*	12/ *9*	184/ *166*
	Scream	53/ *21*	98/ *156*	48/ *18*	22/ *7*	6903/ *6940*	26/ *1*
	Neutral	144/ *93*	91/ *36*	126/ *64*	301/ *174*	0/ *1*	6563/ *6637*

As shown in Table III, for both developed algorithms, the worst classified facial expressions are “Disguist” and “Squint”. Among

pictures from the Multi-Pie database, these types of emotions are the most difficult to discern [21].

Table 3. Error analysis

Error analysis		Metrics		
		Precision (P)	**Recall (R)**	**F-score (F)**
Classes	**Smile**	0,96/ 0,96	0,82/ 0,93	0,94/ 0,95
	Surprise	0,98/ 0,98	0,96/ 0,97	0,97/ 0,97
	Disgust	0,89/ 0,93	0,86 0,9	0,87/ 0,92
	Squint	0,85/ 0,88	0,87/ 0,93	0,86/ 0,9
	Scream	0,97/ 0,97	0,99/ 0,99	0,98/ 0,98
	Neutral	0,91/ 0,95	0,92/ 0,95	0,92/ 0,95

5. CONCLUSION

Derived results show that the use of complicated CNN allows increasing an accuracy from 92,29% to 94,48% for images from Multi-Pie database with various lighting of scene and angle of view. The confusion-matrix and some metrics of quality show that for facial expression analysis the most difficult types of emotions are "Disguist" and "Squint". But this developed model allows to recognize facial expression. The value of F-score is greater than 0,9 for each class. Moreover, 4 from 6 types of emotions (smile, surprise, scream and neutral) classify well enough. This introduced algorithm is quite simple for its implementation. For learning of CNNs there was used supercomputer NVIDIA DGX-1 in Artificial Intelligence Center of P.G Demidov Yaroslavl State University. This process lasted 30–40 minutes. The use of developed CNN is possible in real-time applications on laptops or in special solutions for embedded systems, such as NVIDIA Jetson.

6. ACKNOWLEDGEMENT

This work was supported by Russian Foundation for Basic Research (http://www.rfbr.ru) grant №15-07-08674 and Fund for the Promotion of Innovation (www.fasie.ru) grant UMNIK-NTI №0033562 "Development of algorithms for predicting individual behavior based on visual recognition of emotions".

7. REFERENCES

[1] Ekman P. and Friesen W.V. 1977. *Manual for the Facial Action Coding System*, Consulting Psychologists Press.

[2] Bettadapura V. 2012. *Face Expression Recognition and Analysis: The State of the Art*, Tech Report, arXiv: 1203.6722.

[3] *Emotion Recognition in the Wild Challenge.* https://sites.google.com/site/emotiwchallenge/

[4] *Challenges in Representation Learning: Facial Expression Recognition Challenge.https*://www.kaggle.com/c/challenges-in-representation-learning-facial-expression-recognition-challenge

[5] Berns K., Hirth J. 2006. Control of Facial expressions of the humanoid robot head roman. In: *IEEE International Conference on Intelligent Robots and Systems*, pp. 3119–3124.

[6] Bartlett M.S., Littlewort G., Fasel I., Movellan J.R. 2003. Real time face detection and facial expression recognition: development and applications to human computer interaction. In: *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 53.

[7] Ucar A. 2017. Deep Convolutional Neural Networks for facial expression recognition. *INnovations in Intelligent SysTems and Applications (INISTA)*, IEEE International Conference on, pp. 371–375.

[8] Niu Z., Zhou M., Wang L., Gao X., Hua G. 2016. Ordinal Regression with Multiple Output CNN for Age Estimation. *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4920–4928.

[9] Paisitkriangkrai S., Sherrah J., Janney P., Van-Den Hengel A. 2015. Effective Semantic Pixel labeling with Convolutional Networks and Conditional Random Fields. *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 36–43.

[10] Raghuvanshi A., Choksi V. 2017. *Facial Expression Recognition with Convolutional Neural Networks*, Tech Report.

[11] Khalajzadeh H., Mansouri M., Teshnehlab M. 2013. *Face Recognition using Convolutional Neural Network and Simple Logistic Classifier*, In: Soft Computing in Industrial Applications. Advances in Intelligent Systems and Computing, vol. 223. Springer, Cham, pp. 197–207.

[12] Wen Z., Huang T. 2003. Capturing subtle facial motions in 3d face tracking. *Proceedings of the 9th IEEE International Conference on Computer Vision*, pp. 1343-1350.

[13] Alizadeh S., Fazel A. 2017. *Convolutional Neural Networks for Facial Expression Recognition*. arXiv: 1704.06756.

[14] Yu Z. and Zhang C. 2015. Image based static facial expression recognition with multiple deep network learning. Proceedings of the *2015 ACM International Conference Multimodal Interaction*, pp. 435-442.

[15] Bo-Kyeong Kim S.-Y.D., Roh J. and Lee S.-Y. 2015. Hierarchical committee of deep convolutional neural networks for robust facial expression recognition. *Journal on Multimodal User Interfaces*, vol. 10, pp. 173–189.

[16] Ali Mollahosseini D.C. and Mahoor M.H. 2016. Going deeper in facial expression recognition using deep neural networks. *IEEE Winter Conference on Applications of Computer Vision*, pp. 1-10.

[17] *The Binghamton University 3D Facial Expression (BU-3DFE) Database*. http://www.cs.binghamton.edu/~lijun/Research/3DFE/3DFE_Analysis.html.

[18] *The Multimedia Understanding Group (MUG) Facial Expression Database*. https://mug.ee.auth.gr/fed.

[19] *The CMU Multi-PIE Face Database*. http://www.cs.cmu.edu/afs/cs/project/PIE/MultiPie/Multi-Pie/Home.html.

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

[21] Ivanovsky L., Khryashchev V., Lebedev A., Kosterin I. 2017. Facial Expression Recognition Algorithm Based on Deep Convolution Neural Network. In *Proceedings of the 21th Conference of Open Innovations Association FRUCT'21*. Helsinki, Finland.

[22] *Caffe Framework*. http://caffe.berkleyvision.org.

[23] Bajpai S., 2015. Implementing Convolutional Neural Network with Parallel Computing Using CUDA. *International Journal of Innovative Science, Engineering & Technology*, Vol. 2, Issue 11, pp. 517 – 520.

[24] De Marco M., Fenu G., Medvet E., Pellegrino F.A. 2016. *Computer Vision for the Blind: A comparison of Face Detectors in a Relevant Scenario*. In: Smart Objects and Technologies for Social Good: Second International Conference, GOODTECHS 2016. Proceedings, Springer International Publishing, Switzerland, 2017, pp. 145–154.

[25] Raschka S. 2015. *Python Machine Learning*, Packt Publishing Ltd., 454 p.

[26] VanderPlas J. 2016. *Python Data Science Handbook: Essential Tools for Working with Data. First Edition.* O'Reilly Media, 541 p.

[27] *Scikit-learn*. http://scikit-learn.org/stable/.

Authors' background

Your Name	Title*	Research Field	Personal website
Vladimir Khryashchev	associate professor	Computer vision, machine learning, big data	www.linkedin.com/in/khryashchev
Leonid Ivanovsky	PhD student	Convolutional neural network, face expression analysis	
Andrey Priorov	associate professor	Image processing, mobile robots	