DOI: https://doi.org/10.15276/aait.02.2021.6
UDC 004.93

DEEP LEARNING TECHNOLOGY OF CONVOLUTIONAL NEURAL NETWORKS FOR FACIAL EXPRESSION RECOGNITION

Denys V. Petrosiuk[1)]
ORCID: https://orcid.org/0000-0003-4644-3678; d.petrosyuk1994@gmail.com
Olena O. Arsirii[1)]
ORCID: https://orcid.org/0000-0001-8130-9613; e.arsiriy@gmail.com
Oksana Ju. Babilunha[1)]
ORCID: https://orcid.org/0000-0001-6431-3557; babilunga.onpu@gmail.com
Anatolii O. Nikolenko[1)]
ORCID: https://orcid.org/0000-0002-9849-1797; anatolyn@ukr.net
[1)] Odessa National Polytechnic University. 1, Shevchenko Ave. Odesa, 65044, Ukraine

ABSTRACT

The application of deep learning convolutional neural networks for solving the problem of automated facial expression recognition and determination of emotions of a person is analyzed. It is proposed to use the advantages of the transfer approach to deep learning convolutional neural networks training to solve the problem of insufficient data volume in sets of images with different facial expressions. Most of these datasets are labeled in accordance with a facial coding system based on the units of human facial movement. The developed technology of transfer learning of the public deep learning convolutional neural networks families DenseNet and MobileNet, with the subsequent "fine tuning" of the network parameters, allowed to reduce the training time and computational resources when solving the problem of facial expression recognition without losing the reliability of recognition of motor units. During the development of deep learning technology for convolutional neural networks, the following tasks were solved. Firstly, the choice of publicly available convolutional neural networks of the DenseNet and MobileNet families pre-trained on the ImageNet dataset was substantiated, taking into account the peculiarities of transfer learning for the task of recognizing facial expressions and determining emotions. Secondary, a model of a deep convolutional neural network and a method for its training have been developed for solving problems of recognizing facial expressions and determining human emotions, taking into account the specifics of the selected pretrained convolutional neural networks. Thirdly, the developed deep learning technology was tested, and finally, the resource intensity and reliability of recognition of motor units on the DISFA set were assessed. The proposed technology of deep learning of convolutional neural networks can be used in the development of systems for automatic recognition of facial expressions and determination of human emotions for both stationary and mobile devices. Further modification of the systems for recognizing motor units of human facial activity in order to increase the reliability of recognition is possible using of the augmentation technique.

Keywords: Deep Learning; Transfer Learning; Facial Expression Recognition; Convolutional Neural Networks;

For citation: **Petrosiuk D. V., Arsirii O. O., Babilunha O. Ju., Nikolenko A. O. "Deep Learning Technology of Convolutional Neural Networks for Facial Expression Recognition".** ***Applied Aspects of Information Technology*****. 2021; Vol.4 No.2: 192–201. DOI: https://doi.org/10.15276/aait.02.2021.6**

INTRODUCTION

Automated recognition of facial expressions recognition (FER) and emotion detection (ED) of a person is an urgent task in the development of various intelligent systems and technologies such as [1]:

– systems of human-machine interaction, in which the state of the operator is determined as a reflective agent;

– information technologies of emotional marketing for the promotion of goods, taking into account their perception by a person;

– specialized systems for conducting behavioral and neurobiological studies of the human condition;

– game applications and machine graphics for realistic animation of a human face, etc.

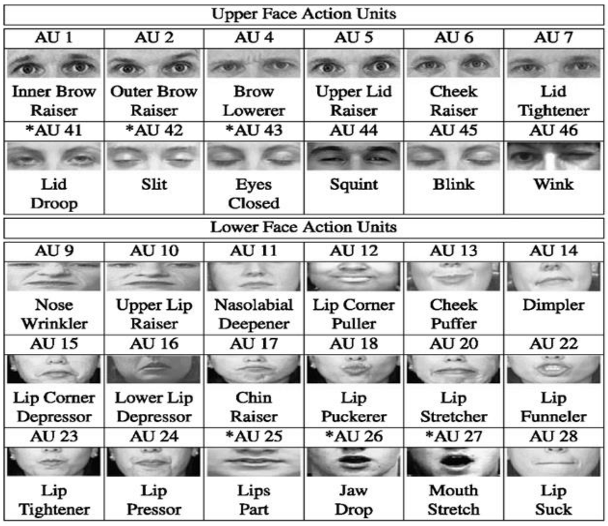
With the growth of the computational capabilities of modern computer systems, automated solutions for FER and ED based on Deep Learning (DL) of Convolutional Neural Networks (CNN) began to appear [2]. However, despite the successful use of CNN for object recognition and classification in computer vision systems [3, 4], solving human FER and ED problems using DL CNN remains a difficult problem.

This is due to the fact that in order to implement DL CNN technological solutions with the required accuracy, in addition to high-performance graphics processors unit (GPU), a sufficiently large set of pre-labeled data, such as the ImageNet set [5], is required, but focused on FER, the receipt of which is still a matter of the future. However, even with the availability of computing power and the specified data set, the time for deep learning of any complex CNN suitable for solving practical problems requiring reliable FER will be quite long. It should be borne in mind that the availability of a solution for automated FER requires intelligent systems that are developed not only for stationary, but also for mobile platforms, which imposes additional restrictions on the resource intensity of the CNN architectures used and the speed of their learning. Therefore, the development of technological solutions for deep learning CNN for solving the FER problem with the required accuracy and resource intensity is an urgent scientific and practical task.

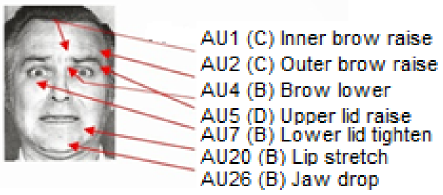
LITERATURE OVERVIEW

The basis for FER is the Facial Action Coding System (FACS), proposed by P. Ekman in collaboration with W.V. Friesen [6, 7], which describes the movement of facial muscles using different action units (AU). In the FACS system, to describe all possible and visually observed changes in a person's face, 12 motor units for the upper part of the face and 18 action units for the lower part (Fig. 1a) are defined, which are associated with the contraction of a certain set of muscles. In this case, action units can occur separately or in combination (Fig. 1b). Two types of state coding are used to describe motor units. The first, more simple, is the coding of the presence or absence of action unit on the face. In the second case, in addition to the first, the intensity or strength of the AU action is also indicated, while 5 levels of intensity coding are possible (neutral <A <B <C <D <E), where A is the least intense action, and E – by the action of maximum force [7]. It is also worth noting that the six universal emotions introduced by P. Ekman such as “anger”, “disgust”, “fear”, “happiness”, “sadness” and “surprise” can be described using a combination of several AUs (Fig. 1c) [8, 9].

In recent years, DL CNNs have become widely used to automate AU-based FERs due to their powerful feature representation and end-to-end effective training scheme, which largely contributed to the first practical successes in the field of human FER and ED [10, 11], [12, 13], [14]. Within DL CNN, an approach is implemented in which features are extracted directly from the input data itself, trying to capture high-level abstractions through hierarchical architectures of multiple nonlinear transformations and representations. Works based on DL CNN models, which currently show the best results on the DISFA set [15], evidence the success of this approach for human FER and ED. Thus, in [16, 17], an EAC-Net approach for AU detection was proposed, based on the addition of two new networks to the previously trained network, trained to recognize AU by features extracted from the entire image and by pre-cut separate areas of face images, representing



a



b

Emotion	AU (Ekman & Friesen)
Happiness	6+12
Sadness	1+4+15
Surprise	1+2+5B+26(or 27)
Disgust	9+15+16
Angry	4+5+7+23
Fear	1+2+4+5+20+26

c

Fig. 1. **Representation of action units [8]:**
a – AU for the upper and lower parts of the face; b – an example of AU distribution with different intensities on a face image; c – combinations of AU in emotions

Source: **[8]**

the area of interest from the point of view of recognition of individual AU species. The authors use the CNN VGG-19 model in their approach [18]. The assumed area of each AU in the image has a fixed size and a fixed location, which is determined by the feature labels on the face image. Based on the structure of the E-Net [16], a method of adversarial learning between AU recognition and face recognition is proposed. In [19], a JAA-Net approach with an adaptive learning module is proposed to improve the initially defined areas of each AU in the image. All of these works demonstrate the effectiveness of modeling the distribution of spatial attention for detecting AU, that is, tracking the area of location of a specific AU (by analogy with a human visual analyzer) when analyzing a scene in FER and ED problems. The results of the AU classification obtained in the studies considered will be used in this work as the baseline for comparing the recognition reliability based on the proposed technological solutions.

However, as noted in the introduction, to implement the DL CNN approach with the required accuracy in the FER and ED problems, a sufficiently large set of pre-labeled data is required. The publicly available DISFA set used in the listed works is limited to videos from 27 subjects – 12 women and 15 men, each of whom recorded a video with 4845 frames [15]. For comparison, we present the characteristics of the ImageNet test dataset [5], the use of which for DL of various CNNs makes it possible to develop intelligent systems for recognizing objects on the scene with the best accuracy and speed [16, 17], [18, 19], [20]. ImageNet is a collectively collected dataset of more than 15 million high-resolution images from 22,000 categories, harnessed by Amazon Mechanical Turk. About 1.2 million training images, 50,000 images for verification and 150,000 images for testing are used to implement DL.

All of the above gives impetus to the development of DL CNN for FER and ED problems using the transfer learning approach, which involves the use of CNNs trained on the data of the ImageNet dataset. This article is devoted to the development of such a technological solution for DL CNN in the FER and ED problems.

THE AIM AND OBJECTIVES OF THE RESEARCH

The aim of the work is to reduce the resource intensity of recognizing human facial expressions without loss of accuracy by reducing the number of trained parameters of neural networks by developing deep learning technology for public convolutional neural networks.

To develop deep learning technology for public CNN, the following tasks were solved:

– taking into account the peculiarities of transfer learning for the FER problem, publicly available pre-trained on the ImageNet CNN set were reasonably selected;

– taking into account the specifics of the selected pre-trained CNNs, a DL CNN model and a method for its training were developed to solve FER problems;

– the developed technology was tested by assessing the resource intensity and reliability of AU recognition on the DISFA set.

MAIN PART

Transfer learning consists in transferring the functions of describing features obtained by the DL CNN model with multiple layers in the process of solving the original recognition problem to the target recognition problem [21, 22], [23, 24]. In general, the process of transfer learning in the context of DL CNN can be represented by the following stages:

Stage 1. Convolutional layers are extracted from the previously trained model (pre-train).

Stage 2. Convolutional layers are frozen to avoid the destruction of any information they contain during future training epochs (train).

Stage 3. Add several new trainable layers on top of the frozen layers. They will learn how to turn old feature maps into predictions for a new dataset.

Stage 4. Train new layers on the target dataset.

Stage 5. The last step is fine-tune, which consists in unblocking the entire model obtained above (or part of it) and retraining on the target dataset with a very low learning rate. This can potentially lead to significant improvements by gradually adapting pre-trained networks to new data.

These stages form the basis of the learning transfer technology aimed at detecting and recognizing AU on a static image of a human face. At the same time, we used public DL CNN pre-trained on the ImageNet set, the results of which research on resource capacity, classification accuracy and fast action on the NVIDIA Titan X Pascal GPU platform are described in [25].

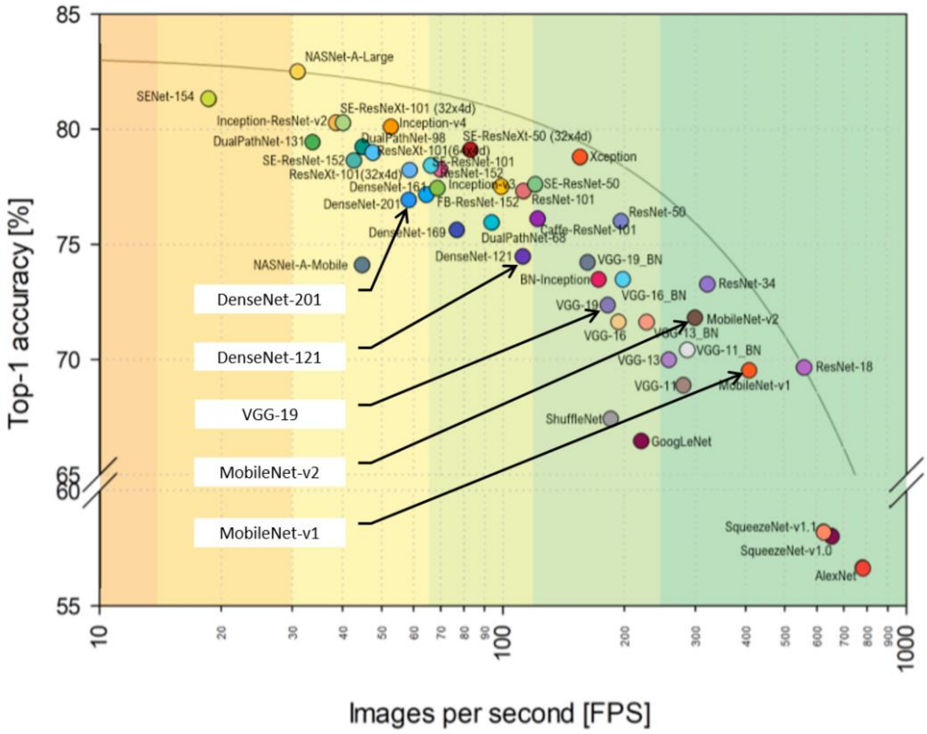


Fig. 2. **A graph of the dependence of the accuracy of CNN classification on the ImageNet set on the frame rate (frames per second, FPS) [25]**

Source: **[25]**

In Fig. 2 shows a graph of the dependence of the classification accuracy on the number of images processed by the network per second (FPS) for a sufficiently wide list of public DL CNNs. The authors evaluated the performance of networks on the NVIDIA Titan X Pascal GPU platform

In the graph below, DL CNNs are highlighted, which were selected by the authors for further research. These are DenseNet-121, DenseNet-201 [26], MobileNet-v1 [27], MobileNet-v2 [28].

Also in Fig. 2 marked the VGG-19 network, which is used by the authors of the EAC-Net approach, which is mentioned in the overview and with which the results are compared further.

The choice was made taking into account the requirements for recognition accuracy and video processing speed. The latter requirement is very important when creating mobile applications for FER and ED human. Let us give the following explanations. As you can see in the graph, the MobileNet family of networks outperforms DenseNet and VGG-19 in video processing speed, but inferior to them in classification accuracy on the ImageNet dataset.

Also, briefly note that VGG-19 is one of the first truly deep networks to achieve high accuracy on the ImageNet dataset. The network has 19 layers and contains more than 140×10^6 parameters. DenseNet (Dense Convolutional Network) family of networks is designed for stationary devices and is implemented through the formation of a sequence of “dense” blocks – each block contains a set of convolutional layers and transition layers that resize the feature map. DenseNet is a network with a much smaller number of trained parameters (about 7×10^6), which is two dozen times less than that of the VGG-19, but the classification accuracy is comparable. The family of MobileNet networks [1], [29], due to its lightness (4×10^6 parameters), has revolutionized computer vision on mobile platforms. The MobileNet model is based on a deep convolution structure that can convert standard convolution to deep convolution and point convolution with a 1 × 1 convolution kernel.

At the same time, it is noted in [25] that the MobileNet family of networks is more than three times faster than the selected networks of the DenseNet family and the VGG-19 network is more than 1.5 times faster.

Within the framework of the chosen approach for the implementation of transfer learning, a DL CNN model was developed to solve the human FER and ED problem (Fig. 3), taking into account the pre-trained publicly available DenseNet with

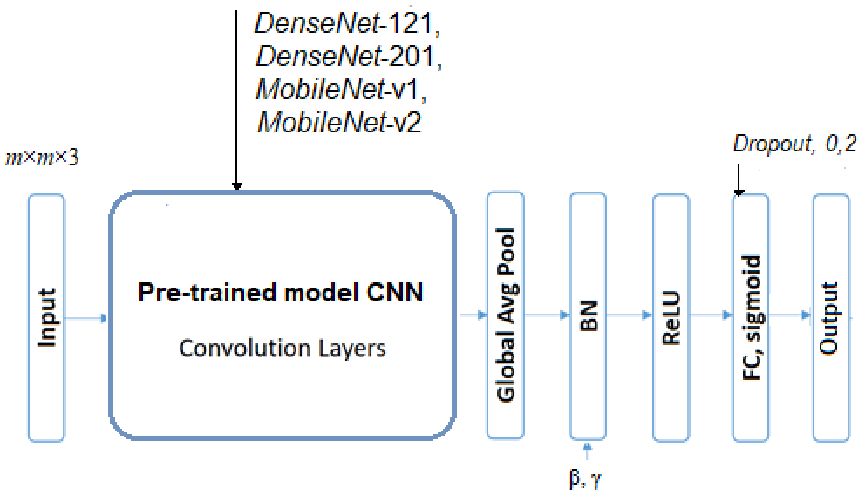


Fig. 3. **DL CNN model for solving the human FER and ED problem based on transfer learning technology**

Source: **compiled by the authors**

numbers 121, 201, MobileNet versions v1, and v2. As shown, CNNs of the selected architectures have both a high learning rate and a significant speed of operation. For each network, the fully connected layers at the output were removed, instead of which, after the sub-sampling layer (Global Average Pooling), an AU Predicted block was added. It consist of from: a layer of batch normalization BN with β and γ – two generalizing variables for each feature [30], a layer with an activation function ReLU, and a new fully connected output layer with a sigmoidal activation function as an AU classifier.The transfer learning method of the proposed DL CNN model consists of two stages: training the classifier on the target dataset and fine-tuning the pre-trained model CNN.

Stage 1. Batch training of the classifier by the backpropagation method on the target dataset consists of the following steps:

1. Initialization of the weight coefficients of the classifier with random values is carried out.

2. The target set of color images is fed to the input of the pre-trained model CNN – a tensor of dimension m×m×3×N, where (m×m is the image size, N is the batch size).

3. (Direct pass): as a result of passing the pre-trained model CNN, feature maps of certain sizes are formed, (for example, 7×7) in the number L, which determines the size of the Global Average Pooling layer (for example, L = 1024). The Global Average Pooling output corresponds to the averaged value of each input feature map and has the form of an L × N matrix. Batch Normalization (BN) is performed for each row of the resulting matrix. At the output of the ReLU layer, only positive values of the coefficients remain, (negative values are zeroed), which are fed to the fully connected classifier layer. Using the target values, the average training error is calculated, which is averaged over the entire batch of size N.

4. (Backward pass the weight coefficients of the classifier are adjusted by the method of back propagation of the error, taking into account the Dropout operation with a decimation factor of 0.2. For the BatchNormalization layer, the β and γ coefficients are adjusted [30].

5. When the retraining of the classifier is achieved (errors of the test and validation samples are tracked), the return pass is completed.

Stage 2. Fine-tuning is performed to improve the quality of the classification. This unfreezes all or part of the pre-trained model CNN coefficients, which are also corrected by the error backpropagation method with a low level of learning rate. Starting from point 2 of the main teaching method, all steps of the forward and backward pass are performed.

The constructed DL CNN model and its transfer learning method form the basis of the deep learning technology of convolutional neural networks. This technology makes it possible to retrain the last DL CNN layer using a set of DISFA images in a reasonable time without changing the weights of other layers, providing the necessary reliability of AU recognition.

EVALUATION AU RECOGNITION RELIABILITY BASED ON THE DL CNN DEVELOPED TECHNOLOGY FOR FER AND ED

Examples of images from the DISFA set are shown in Fig. 4. It contains videos from 27 subjects – 12 women and 15 men, each of whom recorded a video with 4845 frames [9]. Each frame is annotated with AU intensity on a 6-point ordinal scale from 0 to 5, where 0 indicates the absence of AU, while 5 corresponds to the maximum AU intensity. Based on the analysis of previous works [29], an assumption was made about the presence of AU in the image if its intensity is equal to two or more, and about its absence – otherwise. The frequency of occurrence of each AU among 130814 frames of the DISFA dataset is shown in Table. A serious problem of data imbalance should be noted, in which most AUs have a very low frequency of occurrence, while only a few other AUs have a higher frequency of occurrence. Testing was carried out in the form of a subjective-exclusive three-fold cross-check on eight AUs, which determine the following states of motor activity of the muscles of the human face: AU1 – the inner parts of the eyebrows are raised; AU2 – the outer parts of the eyebrows are raised; AU4 – dropped eyebrows; AU6 – cheeks raised; AU9 – wrinkled nose; AU12 – the corners of the lips are raised; AU25 – lips parted; AU26 – drooping jaw (Table).

Convolutional network layers were initialized with pre-trained weights on the ImageNet set, while fully connected layers were initialized with random values. Adam was used as an optimization algorithm with a learning rate of all networks of 0.00001. For the loss function, the Log-Sum-Exp Pairwise (LSEP) function was chosen [31], which gives better results than the weighted binary cross-entropy. Log-Sum-Exp Pairwise Function:

$$l_{lsep} = \log(1 + \sum_{v \notin Y_i} \sum_{u \in Y_i} \left(f_v(x_i) - f_u(x_i)\right),$$

where: $f(x)$ is a label prediction function that maps a vector of an object x into a K-dimensional label space representing the confidence scores of each label; K is equal to the number of unique labels.

One of the main properties of the function (x) is that it must create a vector whose values for true labels Y are greater than for false labels

$$f_u(x) > f_v(x), \forall u \in Y, v \notin Y,$$

where: $f_u(x)$ is the u-th confidence level element for the i-th instance in the dataset, respectively; Y_i is the corresponding label set for the i-th instance in the dataset.

Due to the large imbalance in the data in the DISFA set, the recognition reliability was estimated by the value of the F1-measure (the harmonic mean of Precision and Recall indicators) as an average value – Avg. F1 for all AUs:

$$F_1 = 2\frac{Precision \times Recall}{Precision + Recall},$$

$$Precision = \frac{TP}{TP+FP},$$

$$Recall = \frac{TP}{TP+FN},$$

where: TP – true positive examples, FP – false positive examples; FN – false negative examples.

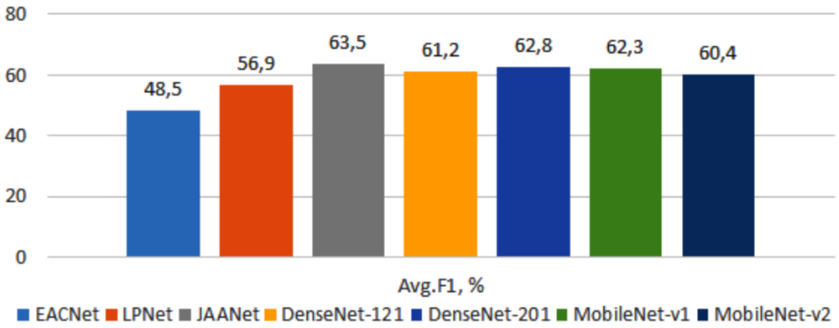
Fig. 4. **Sample images from the DISFA dataset**

Source: **compiled by the authors**

Table. **Number of different types of AU in the DISFA dataset**

AU	1	2	4	6	9	12	25	26
No.	6506	5644	19933	10327	5473	16851	36247	11533

Source: **compiled by the authors**



Number of learning parameters:	150 $\times 10^6$	7×10^6 - 18×10^6	2×10^6 - 3×10^6

Fig. 5. **Diagram of comparative estimation of the AU recognition quality when solving the FER problem using CNN for the DISFA dataset**

***Source:* compiled by the authors**

The models of DenseNet-201 and MobileNet-v1 networks (Fig. 5) showed the highest values of the F1-measure. At the same time, the networks MobileNet-v2 and MobileNet-v1 have the smallest number of trained parameters (2×10^6 and 3×10^6, respectively); the networks DenseNet-201 and DenseNet-121 are more resource-intensive (18×10^6 and 7×10^6, respectively). The heaviest VGG-like convolutional neural networks EACNET, LPNET and JAANET have the largest number of trained parameters – 150×10^6.

CONCLUSIONS

To summarize, we can say that the developed technology of deep learning of convolutional neural networks for facial expression recognition problems is based on the proposed model and the method of transfer learning of pre-trained public DL CNNs with subsequent “fine tuning” of the network parameters. Testing the proposed technology on a small set of DISFA data made it possible to obtain a reduction resource intensity by reducing the number of CNN trained parameters for solving human FER and ED problems using various motor units of human facial activity without reducing the recognition quality indicators.

The proposed technology uses the publicly available deep CNNs of the DenseNet and MobileNet families pre-trained on the ImageNet dataset and pre-trained on the DISFA set for FER and ED problems, thus solving the problem of weakly labeled datasets [32]. As a direction for further research, the following should be noted. One of the problems of the developed CNN deep learning technology for human FER and ED problems, in addition to the need for a large set of pre-labeled data, is the tendency of such networks to retrain. To combat retraining, the authors are trying to use the augmentation technique aimed at increasing the variability of the training sample. The essence of the proposed method of augmentation is to concatenate (unite) halves of two different faces in one image. Such combining can be carried out both vertically and horizontally of the image. As further studies, the authors experimentally tested both options for augmentation. The results obtained indicate the prospects for further research in this direction.

REFERENCES

1. Samadiani, N., Huang, G., Cai, B., Luo, W., Chi, C., Xiang, Y. & He, J. “A Review on Automatic Facial Expression Recognition Systems Assisted by Multimodal Sensor Data”. *Sensors*. 2019; Vol. 19 No 8: p. 1863. DOI: https://doi.org/10.3390/s19081863.

2. Li S. & Deng W. “Deep Facial Expression Recognition: A Survey”. *IEEE Transactions on Affective Computing*. 2018. DOI: https://doi.org/10.1109/TAFFC.2020.2981446.

3. Le, Quoc Tuan , Antoshchuk, S. G., Nguyen, Thi Khanh Tien , Tran, The Vinh & Dang, Nhan Cach . “Automated Student Attendance Monitoring System in Classroom Based on Convolutional Neural Networks”. *Applied Aspects of Information Technology. Publ. Nauka i Tekhnika*. Odessa: Ukraine. 2020; Vol. 3 No.3: 179–190. DOI: https://doi.org/10.15276/aait.03.2020.6.

4. Nguyen, Thi Khanh Tien, Antoshchuk, S. G., Nikolenko, A. O., Tran, Kim Thanh & Babilunha, O. Yu. “Non-Stationary Time Series Prediction Using One-Dimensional Convolutional Neural Network Mod-

els”. *Herald of Advanced Information Technology. Publ. Science i Technical.* 2020; Vol.3 No.1: 362–372. Odesa. Ukraine. DOI: https://doi.org/10.15276/hait.01.2020.3.

5. “ImageNet: ImageNet overview”. – Available from: https://image-net.org/about.php – [Accessed Dec 2020].

6. Ekman, P. & Friesen, W. “Facial Action Coding System: A Technique for the Measurement of Facial Movement”. *Consulting Psychologists Press* .1978.

7. Ekman, P., Friesen, W.V. & Hager, J.C. “Facial Action Coding System (FACS)”. *A Human Face.* 2002. – Available from: https://web.archive.org/web/20080607095042/http://www.face-and-emotion.com/dataface/facs/manual/TitlePage.html.

8. Soysal, Ö. M., Shirzad, S. & K. Sekeroglu, K. “Facial Action Unit Recognition Using Data Mining Integrated Deep Learning”. *2017 International Conference on Computational Science and Computational Intelligence (CSCI)*, 2017. p. 437–443. DOI: https://doi.org/10.1109/CSCI.2017.74.

9. Ekman, P. “Facial Expression and Emotion”. *American Psychologist.*1993; Vol. 48, No. 4: p.p.384-392. https://doi.org/10.1037/0003-066X.48.4.384

10. Tian, Y.-I., Kanade, T. & Cohn, J. F. “Recognizing Action Units for Facial Expression Analysis”. *IEEE Transactions on Pattern Analysis and Machine Intelligence.* 2001; Vol.23 No.2: 97–115. DOI: https://doi.org/10.1109/34.908962.

11. Lin, Q. & He, R. & Jiang, P. “Feature Guided CNN for Baby’s Facial Expression Recognition”. *Complexity*. 2020. p.1–10. DOI: https://doi.org/10.1155/2020/8855885.

12. Liu, M., Li, S., Shan, S., Wang, R. & Chen, X. “Deeply Learning Deformable Facial Action Parts Model for Dynamic Expression Analysis”. *Asian Conference on Computer Vision*; *Publ. Springer.* Berlin: Germany. 2014; Vol. 9006: 143–157. DOI: https://doi.org/10.1007/978-3-319-16817-3_10

13. Jung, H., Lee, S., Yim, J., Park, S. & Kim, J. “Joint Fine-Tuning in Deep Neural Networks for Facial Expression Recognition”. *Proceedings of the IEEE International Conference on Computer Vision (ICCV).* Santiago: Chile. 7–13 December, 2015. p. 2983–2991. DOI: https://doi.org/10.1109/ICCV.2015.341.

14. Mollahosseini, A., Chan, D. & Mahoor, M.H. “Going deeper in facial expression recognition using deep neural networks”. *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*. 2016. p.1–10. DOI: https://doi.org/10.1109/WACV.2016.7477450.

15. Mavadati, S. M., Mahoor, M. H., Bartlett, K., Trinh, P. & Cohn, J. F.”DISFA: A Spontaneous Facial Action Intensity Database”. *IEEE Transactions on Affective Computing.* April-June, 2013; Vol. 4 No. 2: 151–160. DOI: https://doi.org/10.1109/T-AFFC.2013.4.

16. Li, W., Abtahi, F., Zhu, Z., & Yin, L. “EAC-Net: Deep Nets with Enhancing and Cropping for Facial Action Unit Detection”. *In IEEE Transactions on Pattern Analysis and Machine Intelligence.* Nov. 2018; Vol. 40 No. 11: 2583–2596. DOI: https://doi.org/10.1109/tpami.2018.2791608.

17. Zhang, Z., Zhai, S., & Yin, L. “Identity-based Adversarial Training of Deep CNNs for Facial Action Unit Recognition”. *British Machine Vision Conference. BMVA Press.* 2018. 226 p. – Available from: http://bmvc2018.org/contents/papers/0741.pdf.

18. Simonyan, K. & Zisserman. “A.Very Deep Convolutional Networks for Large-Scale Image Recognition”. In: Bengio, Y., LeCun, Y. (eds.). *3rd International Conference on Learning Representations, ICLR* San Diego: CA. USA. *Conference Track Proceedings* (May 7-9, 2015). – Available from: https://arxiv.org/abs/1409.1556 .

19. Shao, Z., Liu, Z., Cai, J. & Ma, L. “JÂA-Net: Joint Facial Action Unit Detection and Face Alignment Via Adaptive Attention”. *Int. J. Comput. Vis*. 2021; Vol.129 No.2: 321–340. DOI: https://doi.org/10.1007/s11263-020-01378-z.

20. Walecki, R., Rudovic, O., Pavlovic, V., Schuller, B. & Pantic, M. “Deep Structured Learning for Facial Action Unit Intensity Estimation”. *In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2017. p. 5709–5718. DOI: https://doi.org/10.1109/CVPR.2017.605.

21. Almaev, T., Martinez, B. & Valstar, M. “Learning to Transfer: Transferring Latent Task Structures and Its Application to Person-Specific Facial Action Unit Detection”. In: *IEEE International Conference on Computer Vision.* 2015. p. 3774–3782. DOI: https://doi.org/ 10.1109/ICCV.2015.430.

22. Lim, Y., Liao, Z., Petridis, S. & Pantic, M. “Transfer Learning for Action Unit Recognition*”. ArXiv.* 2018. – Available from: https://arxiv.org/abs/1807.07556v1 .

23. Ntinou, I., Sanchez, E., Bulat, A., Valstar, M. & Tzimiropoulos, G. “A Transfer Learning Approach to Heatmap Regression for Action Unit Intensity Estimation”. *ArXiv*. 2020. – Available from: https://arxiv.org/abs/2004.06657.

24. Akhand, M. A. H., Roy, S., Siddique, N., Kamal, Md. A. S. & Shimamura, T. “Facial Emotion Recognition Using Transfer Learning in the Deep CNN”. *Electronics.* 2021; Vol. 10 No 9. DOI: https://doi.org/10.3390/ electronics10091036.

25. Bianco, S., Cadene, R., Celona, L. & Napoletano, P. “Benchmark Analysis of Representative Deep Neural Network Architectures”. *IEEE Access*. 2018; Vol. 6: 64270–64277. DOI: https://doi.org/10.1109/ ACCESS.2018.2877890.

26. Huang, G., Liu, Z., Van Der Maaten, L. & Weinberger, K. “Densely Connected Convolutional Networks”. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR).* 2017. p. 2261–2269. DOI: https://doi.org/10.1109/CVPR.2017.243.

27. Howard, A., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M. & Adam, H. “Mobilenets: Efficient Convolutional Neural Networks for Mobile Vision Applications”. *ArXiv*. 2017. – Available from: https://arxiv.org/abs/1704.04861.

28. Sandler, M., Howard, A., Zhu, M., Zhmoginov, A. & Chen, L.C. “Mobilenetv2: Inverted Residuals and Linear Bottlenecks”. *IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2018. p.4510–4520. DOI: https://doi.org/10.1109/CVPR.2018.00474.

29. Shao, Z., Liu, Z., Cai, J., Wu, Y. & Ma, L. “Facial Action Unit Detection Using Attention and Relation Learning”. *IEEE Transactions on Active Computing*. 2019. DOI: https://doi.org/10.1109/ ta_c.2019.2948635

30. Ioffe, S. & Szegedy, C. “Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift”. *In: Bach, F., Blei, D. (eds.) Proceedings of the 32nd International Conference on Machine Learning. Proceedings of Machine Learning Research*. France. 07–09 Jul, 2015; Vol. 37: 448–456. – Available from: https://proceedings.mlr.press/v37/ioffe15.html.

31. Li, Y., Song, Y. & Luo, J. “Improving Pairwise Ranking for Multi-Label Image Classification”. In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2017. p. 1837–1845. DOI: https://doi.org/10.1109/CVPR.2017.199

32. Arsirii O., Antoshchuk S., Babilunha O., Manikaeva O. & Nikolenko A. “Intellectual Information Technology of Analysis of Weakly-Structured Multi-Dimensional Data of Sociological Research Lecture Notes in Computational Intelligence and Decision Making”. *Advances in Intelligent Systems and Computing. Publ. Springer, Cham.* 2020; Vol. 1020: 242–258. DOI: https://doi.org/10.1007/978-3-030-26474-1_18.

Conflicts of Interest: the authors declare no conflict of interest

Received 22.12.2020
Received after revision 25.02.2021
Accepted 15.03.2021

DOI: https://doi.org/10.15276/aait.02.2021.7
УДК 004.93

ТЕХНОЛОГІЯ ГЛИБОКОГО НАВЧАННЯ ЗГОРТКОВИХ НЕЙРОННИХ МЕРЕЖ ДЛЯ РОЗПІЗНАВАННЯ ВИРАЗІВ ОБЛИЧЧЯ

Денис Валерійович Петросюк[1]
ORCID: https://orcid.org/0000-0003-4644-3678; d.petrosyuk1994@gmail.com
Олена Олександрівна Арсірій[1]
ORCID: https://orcid.org/0000-0001-8130-9613; e.arsiriy@gmail.com
Оксана Юріївна Бабілунга[1]
ORCID: https://orcid.org/0000-0001-6431-3557; babilunga.onpu@gmail.com
Анатолій Олександрович Ніколенко[1]
ORCID: https://orcid.org/0000-0002-9849-1797; anatolyn@ukr.net
[1] Одеський національний політехнічний університет, пр. Шевченка, 1. Одеса, 65044, Україна

АНОТАЦІЯ

Проаналізовано застосування глибокого навчання згорткових нейронних мереж для вирішення проблеми автоматизованого розпізнавання виразу обличчя та визначення емоцій людини. Запропоновано використовувати переваги трансферного підходу до глибокого навчання згорткових нейронних мереж для вирішення проблеми недостатнього обсягу даних у наборах зображень з різними виразами обличчя. Більшість із цих наборів даних маркуються відповідно до системи кодування обличчя, заснованої на одиницях руху обличчя людини. Розроблена технологія трансферного навчання загальнодоступних сімейств глибоких згорткових нейронних мереж DenseNet та MobileNet з подальшим «тонким налаштуванням» параметрів мережі дозволила скоротити час навчання та обчислювальні ресурси при вирішенні задачі розпізнавання виразу обличчя без втрати надійності розпізнавання моторних одиниць. Під час розробки технології глибокого навчання для згорткових нейронних мереж були вирішені наступні завдання. По-перше, вибір загальнодоступних згорткових нейронних мереж сімейств DenseNet та MobileNet, попередньо навчених на наборі даних ImageNet, був обґрунтований з урахуванням особливостей трансферного навчання для розпізнавання виразу обличчя та визначення емоцій. По-друге, розроблено модель глибокої згорткової нейронної мережі та метод її навчання для вирішення задач розпізнавання виразу обличчя та визначення людських емоцій з урахуванням особливостей обраних попередньо навчених згорткових нейронних мереж. По-третє, випробувана розроблена технологія глибокого навчання. На останок оцінено ресурсоємність та надійність розпізнавання моторних одиниць на наборі DISFA. Запропонована технологія глибокого навчання згорткових нейронних мереж може бути використана при розробці систем для автоматичного розпізнавання виразу обличчя та визначення людських емоцій як для стаціонарних, так і для мобільних пристроїв. Подальша модифікація систем розпізнавання рухових одиниць обличчя людини з метою підвищення надійності розпізнавання можлива за допомогою методу аугментації.

Ключові слова: Глибоке навчання; трансферне навчання; розпізнавання виразу обличчя; згорткові нейронні мережі

ABOUT THE AUTHORS

Denys Valeriiovych Petrosiuk – PhD Student of the Department of Information Systems. Odessa National Polytechnic University. 1, Shevchenko Ave. Odesa, 65044, Ukraine
ORCID: https://orcid.org/0000-0003-4644-3678; d.petrosyuk1994@gmail.commm
***Research field*:** Information Technology; Artificial Intelligence; Machine Learning; Neural Networks

Денис Валерійович Петросюк – аспірант кафедри Інформаційних систем. Одеський національний політехнічний університет, пр. Шевченка, 1. Одеса, 65044, Україна

Olena Oleksandrivna Arsirii – Dr. Sci. (Eng), Professor, Head of the Department of Information Systems. Odessa National Polytechnic University. 1, Shevchenko Ave. Odesa, 65044, Ukraine
ORCID: https://orcid.org/0000-0001-8130-9613; e.arsiriy@gmail.com
***Research field*:** Information Technology; Artificial Intelligence; Decision Support Systems; Machine Learning; Neural Networks

Олена Олександрівна Арсірій – доктор технічних наук, професор, завідувач кафедри Інформаційних систем. Одеський національний політехнічний університет, пр. Шевченка, 1. Одеса, 65044, Україна

Babilunha Oksana Yurievna – Candidate of Technical Sciences**,** Associate Professor of the Department of Information Systems. Odessa National Polytechnic University. 1, Shevchenko Ave. Odesa, 65044, Ukraine
ORCID: https://orcid.org/0000-0001-6431-3557; babilunga.onpu@gmail.com
***Research field*:** Image Processing, Data analysis; Artificial Intelligence Methods and Systems

Бабілунга Оксана Юрійовна – кандидат технічних наук, доцент кафедри Інформаційних систем. Одеський національний політехнічний університет, пр. Шевченка, 1. Одеса, 65044, Україна

Anatolii Oleksandrovych Nikolenko – Candidate of Technical Sciences, Associate Professor of the Department of Information Systems. Odessa National Polytechnic University. 1, Shevchenko Ave. Odesa, 65044, Ukraine
ORCID: https://orcid.org/0000-0002-9849-1797; anatolyn@ukr.net
***Research field*:** Modelling Systems; Digital Signal and Image Processing; Artificial Intelligence Methods and Systems

Ніколенко Анатолій Олександрович – кандидат технічних наук, доцент кафедри Інформаційних систем. Одеський національний політехнічний університет, пр. Шевченка, 1. Одеса, 65044, Україна