Audio-Visual Continuous Recognition of Emotional State in a Multi-User System Based on Personalized Representation of Facial Expressions and Voice

A. V. Savchenko[a],* and L. V. Savchenko[a],**

[a] *HSE University, Laboratory of Algorithms and Technologies for Network Analysis, Nizhny Novgorod, 603155 Russia*

** e-mail: avsavchenko@hse.ru*

*** e-mail: lsavchenko@hse.ru*

Abstract—This paper is devoted to tracking dynamics of psycho-emotional state based on analysis of the user's facial video and voice. We propose a novel technology with personalized acoustic and visual lightweight neural network models that can be launched in real-time on any laptop or even mobile device. At first, two separate user-independent classifiers (feed-forward neural networks) are trained for speech emotion and facial expression recognition in video. The former extracts acoustic features with OpenL3 or OpenSmile frameworks. The latter is based on preliminary extraction of emotional features from each frame with a pre-trained convolutional neural network. Next, both classifiers are fine-tuned using a small number of short emotional videos that should be available for each user. The face of a user is identified during the real-time tracking of emotional state to choose the concrete neural networks. The final decision about current emotion in a short time frame is predicted by blending the outputs of personalized audio and video classifiers. It is experimentally demonstrated for the Russian Acted Multimodal Affective Set that the proposed approach makes it possible to increase the emotion recognition accuracy by 2–15%.

Keywords: audio-visual emotion recognition, emotional state tracking, personalized facial expression recognition, speaker-dependent speech emotion recognition, fusion of audio and video classifiers

DOI: 10.1134/S1054661822030397

INTRODUCTION

Emotions affect the psychological status of any person and play an important role in human life and work. They usually appear spontaneously, which makes recognizing them accurately and on time a challenging problem. The change of a person's internal affective state or intention is reflected in many human physical signals [3], among which the most useful for practical applications are facial expressions and voice. Automation of facial expression recognition (FER) and speech emotion recognition (SER) methods is one of the crucial points of pattern recognition having decisive importance for increasing the efficiency of emotion analysis. Availability of both audio and visual modalities in a video makes it possible to develop the audio-visual emotion recognition techniques. They can be applied in human computer interfaces, affective computing, lie detection, intelligent environments [3], assessment of several neuropsychiatric disorders [20], etc.

Unfortunately, constructing image models and representations allowable by efficient emotion recognition algorithms is very difficult because the datasets available for FER and SER are small and dirty. In fact, the labeling of an emotional video may be very difficult as perception of emotions varies from person to person, so many labels are ambiguous [12]. Moreover, the labeling of the beginning and end positions of each emotion at frame level [15] is required to track the changes in the emotional state [11]. As a result, the accuracy of even the state-of-the-art models trained on such datasets is still limited to 50–70% if the subjects from the training and testing sets are disjoint. For example, the single ResNet model with multilevel attention mechanism and self-training on unlabeled body language dataset with iterative training [8] reached validation accuracy 55.2% for the AFEW (Acted Facial Expressions in the Wild) database [3]. Representations of faces based on carefully pre-trained EfficientNet-B0 is the best-known single model for AFEW with accuracy greater than 59% [13]. The factorized bilinear pooling in the attention cross-modal feature fusion mechanisms [22] lead to the greatest validation accuracy (65.5%) on the same dataset. The highest accuracy on the testing set (62.78%) is obtained by the bi-modality fusion [9] of audio and video features extracted by four different CNNs. The multimodal dynamic fusion network [6] reached an accuracy 68.2% on IEMOCAP (Interac-

Received June 1, 2022; revised June 1, 2022; accepted June 1, 2022

tive Emotional Dyadic Motion Capture) database for the emotion recognition in conversations problem. The pre-trained deep convolutional neural network (CNN) with a correlation-based feature selection [6] in the speaker-independent mode achieved an SER accuracy of 56.5 and 63% for IEMOCAP and RAVDESS (Ryerson Audio-Visual Database of Emotional Speech and Song), respectively.

It is important to emphasize that many recent papers report much better performance for the speaker-dependent mode, in which the training and testing sets contain data from the same subjects [12, 17]. For instance, the model from the above-mentioned paper [6] trained in this mode achieved accuracy of 83.8% for IEMOCAP and 81.3% for RAVDESS. The multiview facial expression lightweight network [7] had the FER accuracy 90–95% for several datasets with random train-test split. The subject-dependent challenge of the FER task is accomplished in [19] with a novel face recognition-based attention framework. The greatest unweighted average recall (UAR) for the RAMAS (Russian Acted Multimodal Affective Set) [11] has been established by fusing audio and video classifiers [12] with the random train-test split of the sequences of frames with the same emotion.

Thus, in this paper, we propose to develop personalized short-term FER [14] and SER [17] representations that have been adapted to each user of a multiuser system. The audio and video classifiers are fused in a novel technology for automatic audio-visual tracking of changes in the psycho-emotional state of the subject. The concrete audio and video models are chosen using preliminary video-based face recognition. The remaining part of this paper discusses the details of the proposed approach and its experimental study for the RAMAS dataset [11]. The results of the research and the conclusions can be useful for many engaged in the field of pattern recognition and image mining.

TASK FORMULATION

The task of continuous recognition of emotional state is formulated as follows. Let a set of K users (speakers) be available. Given an input video with the face and voice of one of these users, it is required to assign one of $C > 1$ emotional classes for every moment in time. In this paper, the typical assumption is made about the smoothness of psycho-emotional state. Hence, it is possible to split the whole signal into partially overlapped video X_v and audio X_a fragments of short duration (0.5–5 s), for which emotion is considered to be constant. Thus, the task is to predict the class label c of emotions represented by audio signal $X_a = \{x_a(t)\}$, $t = 1, 2, ..., T_a$ and a sequence of $T_v > 1$ video frames (facial images) $X_v = \{X_v(t)\}$, $t = 1, 2, ..., T_v$, where the number of samples T_a in speech signal and number of video frames T_v are relatively small. For simplicity, we assume that only one facial image has been preliminary extracted from each frame by using appropriate face detection technique [21]. In order to solve this task, the training set of $N > 1$ pairs of facial video and audio signals $\{(X_{v;n}, X_{a;n})\}$, $n = 1, 2, ..., N$ of other persons with known emotional category c_n should be available. Here each video signal is represented by a sequence of facial frames $X_{v;n} = \{X_{v;n}(t)\}$.

At first, it is necessary to extract visual and acoustic emotional features. In this paper, we use the MobileNet [2] and EfficientNet [13, 15] models pretrained on the AffectNet dataset of facial photos [10]. The facial images $X_v(t)$ and $X_{v;n}(t)$ are fed into a CNN, and the D-dimensional feature vectors (embeddings) $\mathbf{x}_v(t)$ and $\mathbf{x}_{v;n}(t)$ are extracted at the output of the penultimate layer. There are several techniques to compute descriptor of the whole video X_v, such as attention mechanism [14, 19, 22], but we will use the simple component-wise averaging of feature vectors $\mathbf{x}_v(t)$ and $\mathbf{x}_{v;n}(t)$ to obtain D-dimensional video descriptors $\mathbf{x}_v$ and $\mathbf{x}_{v;n}$, respectively.

The audio feature vectors $\mathbf{x}_a$ and $\mathbf{x}_{a;n}$ are extracted from the speech signal X_a and $X_{a;n}$ using libraries OpenSmile [4] and OpenL3 [1]. The former uses Emobase configuration of traditional acoustic features, such as pitch frequency, Mel-frequency cepstral coefficients, etc. The latter extracts deep audio embeddings based on the L^3-Net (Look, Listen, and Learn) CNN trained through self-supervised learning of audio-visual correspondence in videos as opposed to other embeddings requiring labeled data.

Finally, arbitrary audio and video classifiers are trained on the sets $\{(\mathbf{x}_{v;n}, c_n)\}$ and $\{(\mathbf{x}_{a;n}, c_n)\}$. In this paper, we will use RF (Random Forest) and SVM (Support Vector Machine) from scikit-learn and a feed-forward neural network, such as multiclass logistic regression or MLP (multilayer perceptron) from the TensorFlow 2 framework.

PROPOSED APPROACH

The specificity and complexity of audio-visual emotion recognition problems stem from necessity to achieve some balance between such highly contradictory factors as variability of emotional features for different persons, ambiguous labeling of existing emotional datasets and requirement for near-real-time processing for continuous tracking of emotional state. Hence, the typical approach from the previous section based on lightweight visual and acoustic representations is not very accurate if the training and testing set do not contain data of the same subjects [14]. In this section, we describe the possibility to develop personalized representations with an assumption that a small

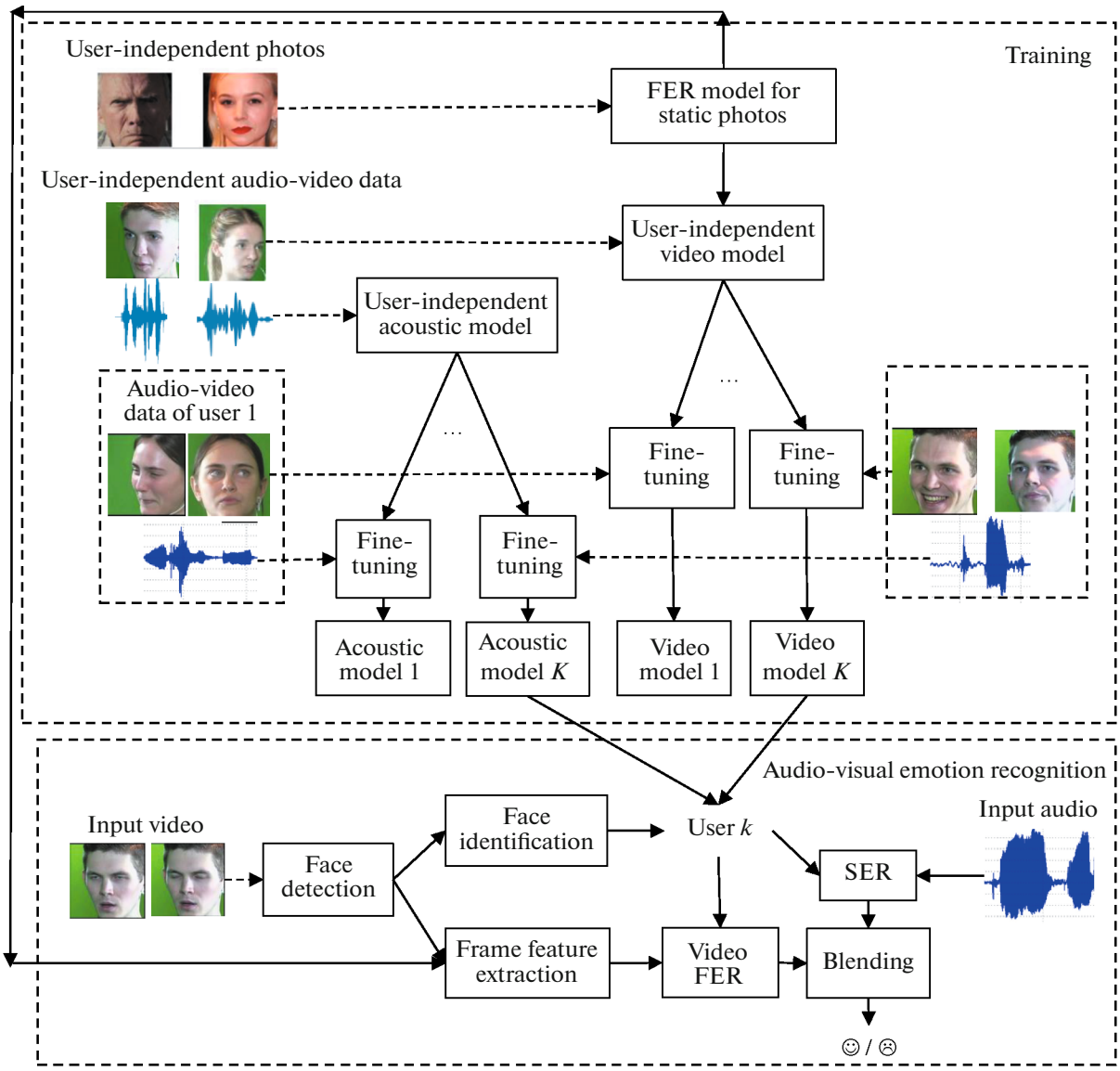

Fig. 1. Proposed technology for audiovisual tracking of user's emotional state in a multi-user system.

set of $N_k > C$ utterances and facial videos is available for every kth user ($k = 1, 2, ..., K$). The proposed technology for continuous recognition of emotional state in a multi-user system is shown in Fig. 1.

The top part of this figure contains the training of user-independent audio and video MLP-based classifiers from the previous section. Next, the personalized acoustic and video models are obtained for every mth user by fine-tuning these MLPs given only the data from this user. The MLP is initialized by the weights of the speaker-independent model, and the training process is repeated over 50 epochs using SGD (stochastic gradient descent) optimizer with learning rate 0.001.

The audio-visual emotion recognition is implemented as follows. At first, facial regions are detected in each video frame using MTCNN (multi-task cascaded CNN) [21]. The face is recognized by the nearest neighbor classifier of average facial embeddings extracted by our lightweight MobileNet or EfficientNet from each video frame [13]. As a result, the user identifier k is obtained. Next, short fragments of the input video and audio signals are stored in X_v and X_a, and their visual and acoustic representations are estimated as described in the previous section. The input features $\mathbf{x}_v$ and $\mathbf{x}_a$ are fed into the kth video and speech models to obtain the C-dimensional scores (estimates of posterior probabilities) $[p_{v;1}, ..., p_{v;C}]$ and $[p_{a;1}, ..., p_{a;C}]$, respectively. The simple blending rule [16] is used for fusion of audio and video modalities to compute the final vector of scores $p_c = wp_{v;c} + (1 - w)p_{a;c}$, $c = 1, 2, ..., C$, where the weight w is estimated using

Table 1. Classification results of speaker-dependent video-based FER

Features	Classifier	Metric, %	Number of agreed annotators				
			1	2	3	4	5
VGGFace [12]	RF	UAR	42.9	51.4	57.1	60.2	56.4
Fine-tuned EfficientNet-B3 [12]		UAR	45.3	53.1	65.3	74.8	70.8
MobileNet v1	SVM	UAR	40.1	65.8	68.4	76.8	78.7
		Accuracy	46.7	69.7	71.4	79.2	81.2
	MLP	UAR	34.4	59.8	67.9	73.4	73.3
		Accuracy	49.8	68.9	73.5	78.1	84.6
EfficientNet-B0	SVM	UAR	39.8	66.4	67.9	75.9	78.9
		Accuracy	46.0	68.5	70.6	78.1	80.8
	MLP	UAR	34.1	67.6	72.6	74.5	75.3
		Accuracy	47.1	69.7	73.5	80.8	83.3

Table 2. Classification results of speaker-dependent audio-based SER

Features	Classifier	Metric, %	Number of agreed annotators				
			1	2	3	4	5
LLD from OpenSmile	SVM [12]	UAR	28.9	31.1	38.3	40.4	46.4
	LSTM [12]	UAR	34.4	42.3	46.0	46.3	46.3
Emobase from OpenSmile	SVM	UAR	28.6	29.9	43.5	48.8	50.4
		Accuracy	35.0	33.6	43.4	49.0	51.2
	MLP	UAR	29.8	37.6	47.2	49.2	53.5
		Accuracy	35.5	42.3	49.7	50.7	55.7
OpenL3	SVM	UAR	26.1	27.3	49.1	48.5	53.4
		Accuracy	32.0	38.9	50.5	51.5	53.6
	MLP	UAR	29.5	36.7	47.1	49.4	54.1
		Accuracy	35.9	43.2	51.4	53.6	59.1

cross-validation. The emotional class with the greatest score p_c is returned as a final solution for the current moment in time. The dynamics of predicted emotional states can be further processed in various practical applications. For example, the standard deviation of emotions computed for all time moments [2] can be useful for stress-level analysis or lie detection. Let us experimentally prove the claim that the proposed personalized models are much more accurate when compared to the speaker-independent classifiers.

EXPERIMENTAL RESULTS

In this section, the RAMAS dataset [11] was used because it is the only one publicly available multimodal emotional dataset with frame-level annotations and known subjects. It contains 564 audio and facial videos from 10 actors. The beginning and the end of each of $C = 6$ emotions (anger, sadness, disgust, happiness, fear, or surprise) and neutral class are labeled by at least 5 annotators for each video. In this paper, we borrowed the testing protocol originally introduced in [12]. The neutral emotion was dropped, and a threshold (level of confidence) n_a was set for a number of agreed annotators to obtain emotional intervals for each threshold. As a result, we obtain different sets of video and audio fragments with corresponding class labels that were chosen by at least n_a annotators.

In the first experiment, the speaker-dependent mode with the random train-test split from the paper [12] was used. As a result, the training/testing sets contain 2277/380, 1539/265, 1425/244, 1468/294 and 1124/234 samples for $n_a = 1, 2, ..., 5$, respectively. The UAR and accuracy of video and audio emotion recognition are shown in Tables 1 and 2, respectively. These results demonstrate that our visual models are much better (up to 10%) than VGGFace and fine-tuned EfficientNet-B3 from [12] for video data from at least $n_a = 2$ agreed annotators. We used the same OpenSmile library as the authors of [12], so the UAR for OpenSmile features are more or less equal. However,

Table 3. Classification results of personalized emotion recognition

Modality	Features	Classifier	Metric	Number of agreed annotators				
				1	2	3	4	5
Audio	Emobase from OpenSmile	Speaker-independent MLP	UAR, %	33.3	41.2	49.7	53.1	46.6
			Accuracy, %	32.6	42.5	47.9	52.3	46.5
		Proposed personalized model	UAR, %	35.0	44.2	54.5	55.7	52.4
			Accuracy, %	34.3	43.8	51.2	52.9	51.2
Audio	OpenL3	Speaker-independent MLP	UAR, %	37.3	47.9	54.1	52.7	53.4
			Accuracy, %	32.8	46.1	51.5	51.3	52.5
		Proposed personalized model	UAR, %	37.4	50.5	58.7	58.7	60.2
			Accuracy, %	38.6	49.7	52.7	54.5	56.6
Video	EfficientNet	Speaker-independent MLP	UAR, %	33.3	47.9	50.4	52.7	47.0
			Accuracy, %	43.1	57.5	58.9	66.7	68.4
		Proposed personalized model	UAR, %	47.7	64.2	72.2	79.1	80.7
			Accuracy, %	49.3	66.4	71.9	80.2	83.6
Video	MobileNet	Speaker-independent MLP	UAR, %	32.4	49.5	49.7	54.8	46.5
			Accuracy, %	43.8	54.6	54.3	61.4	66.3
		Proposed personalized model	UAR, %	46.5	63.2	69.5	77.1	79.1
			Accuracy, %	48.9	65.0	56.4	77.8	82.0
Audio + video	OpenL3 + EfficientNet	Speaker-independent MLP	UAR, %	45.0	58.1	65.0	69.5	71.3
			Accuracy, %	45.0	62.6	64.6	72.0	74.4
		Proposed personalized model	UAR, %	46.4	66.9	76.9	82.1	81.8
			Accuracy, %	49.5	68.2	76.0	82.5	83.6
Audio + video	OpenL3 + MobileNet	Speaker-independent MLP	UAR, %	42.5	54.9	61.8	67.9	69.2
			Accuracy, %	48.8	59.2	62.5	69.0	72.0
		Proposed personalized model	UAR, %	46.5	66.9	73.2	81.9	83.8
			Accuracy, %	48.8	67.8	73.5	81.2	84.8
Audio + video	OpenSmile + EfficientNet	Speaker-independent MLP	UAR, %	46.2	56.2	61.5	65.6	67.5
			Accuracy, %	46.2	60.8	62.6	67.4	70.5
		Proposed personalized model	UAR, %	50.3	64.1	74.7	77.1	82.0
			Accuracy, %	51.1	67.7	73.6	77.9	84.3
Audio + video	OpenSmile + MobileNet	Speaker-independent MLP	UAR, %	42.5	52.0	57.3	61.5	67.1
			Accuracy, %	45.8	57.1	59.7	62.9	69.8
		Proposed personalized model	UAR, %	48.5	61.0	71.9	76.4	81.0
			Accuracy, %	51.1	66.7	71.0	77.0	83.5

embeddings of the L^3-Net are classified slightly more accurately (Table 2).

In the next experiment, the study of the proposed approach (Fig. 1) was carried out. The following implementation of 10-fold cross-validation was used. The audio and video of 9 actors were chosen to train the speaker-independent models. The videos of the remaining actor were randomly split into two equal parts. One of them was used to fine-tune the MLPs trained on the data of other actors. The accuracy and UAR of such a personalized model were estimated on the remaining half of audio-visual data. An experiment with selection of a testing actor was repeated 10 times to verify that all actors were involved in the testing process, and the average metrics were computed. In addition, we estimated the average accuracy of the speaker-independent classifiers trained on the data of 9 actors and tested on the same half of data of the remaining subject. The results of this experiment are presented in Table 3.

The proposed approach has approximately the same performance as the speaker-dependent recognition (Tables 1, 2), but is more flexible, because models for new users can be added in any moment without

affecting the results for previous users. Our method (Fig.1) is much more accurate when compared to conventional speaker-independent mode in all cases except audio modality and only one agreed annotator. For example, the best visual representations (EfficientNet-B0) [13] with a personalized classifier increase the accuracy by 15% if $n_a > 1$. The quality of SER is typically much lower than FER, so that their fusion leads to only 1–2% higher accuracy and UAR when compared to processing of visual modality only. Conventional OpenSmile features [4] are significantly worse than OpenL3 deep audio embeddings [1]. Their fusion with visual representations let us achieve the greatest accuracy (up to 84.8%), and the difference between MobileNet and EfficientNet in this ensemble is not significant. Because the MobileNet is faster and has lower size, it is more preferable in real-time practical applications.

CONCLUSIONS

In this paper, the novel technology was proposed for continuous emotion recognition with personalized audio and visual neural models. It brings a real opportunity to efficiently solve various practical problems via extracting from videos the information necessary for analysis of dynamics of psycho-emotional state. It was experimentally shown that our adaptation of the user-independent classifiers significantly increased the recognition accuracy when compared to universal speaker-independent models (Table 3). We demonstrated that the lightweight MobileNet-based visual representations [2] are more suitable for practical applications due to their high accuracy and known excellent running time and model size. It was also shown that emotional speech is better represented by the L^3-Net from OpenL3 [1] rather than with the traditional OpenSmile features [4].

The main direction for future research is a development of more sophisticated fusion algorithms for acoustic and facial representations instead of our simple blending of audio and video predictions. For example, it is possible to detect moments with pronunciation of vowels and aggregate the features of video frames in these moments only [18]. In addition, it is important to study the semi-supervised methods for development of personalized models without need for labeled audio-visual data for a concrete user.

FUNDING

The work is supported by Russian Science Foundation, grant no. 20-71-10010.

COMPLIANCE WITH ETHICAL STANDARDS

This article is a completely original work of its authors; it has not been published before and will not be sent to other publications until the *PRIA* Editorial Board decides not to accept it for publication.

Conflict of Interest

The authors declare that they have no conflicts of interest.

REFERENCES

1. J. Cramer, H. H. Wu, J. Salamon, and J. P. Bello, "Look, listen, and learn more: Design choices for deep audio embeddings," in *ICASSP 2019–2019 IEEE Int. Conf. on Acoustics, Speech and Signal Processing, Brighton, UK, 2019* (IEEE, 2019), pp. 3852–3856. https://doi.org/10.1109/ICASSP.2019.8682475
2. P. Demochkina and A. V. Savchenko, "MobileEmotiFace: Efficient facial image representations in video-based emotion recognition on mobile devices," in *Pattern Recognition. ICPR International Workshops and Challenges. ICPR 2021*, Ed. A. Del Bimbo, Lecture Notes in Computer Science, Vol. 12665 (Springer, Cham, 2021), pp. 266–274. https://doi.org/10.1007/978-3-030-68821-9_25
3. A. Dhall, R. Goecke, S. Lucey and T. Gedeon, "Collecting large, richly annotated facial-expression databases from movies", IEEE Multimedia **19**, 34–41 (2012). https://doi.org/10.1109/MMUL.2012.26
4. F. Eyben, M. Wöllmer, and B. Schuller, "OpenSmile: the Munich versatile and fast open-source audio feature extractor," in *Proc. 18th ACM Int. Conf. on Multimedia, Firenze, 2010* (Association for Computing Machinery, New York, 2010), pp. 1459–1462. https://doi.org/10.1145/1873951.1874246
5. M. Farooq, F. Hussain, N. K. Baloch, F. R. Raja, H. Yu, and Y. Bin Zikria, "Impact of feature selection algorithm on speech emotion recognition using deep convolutional neural network," Sensors **20**, 6008 (2020). https://doi.org/10.3390/s20216008
6. D. Hu, X. Hou, L. Wei, L. Jiang, and Y. Mo, "MM-DFN: Multimodal dynamic fusion network for emotion recognition in conversations," in *ICASSP 2022–2022 IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP), Singapore, 2022* (IEEE, 2022), pp. 7037–7041. https://doi.org/10.1109/ICASSP43922.2022.9747397
7. S. Jie, and Q. Yongsheng, "Multi-view facial expression recognition with multi-view facial expression light weight network," Pattern Recognit. Image Anal. **30**, 805–814 (2020). https://doi.org/10.1134/S1054661820040197
8. V. Kumar, S. Rao, and L. Yu, "Noisy student training using body language dataset improves facial expression recognition," in *Computer Vision–ECCV 2020 Workshops*, Ed. by A. Bartoli, Lecture Notes in Computer Science, Vol. 12535 (Springer, Cham, 2020), pp. 756–773. https://doi.org/10.1007/978-3-030-66415-2_53
9. S. Li, W. Zheng, Y. Zong, C. Lu, C. Tang, X. Jiang, J. Liu, and W. Xia, "Bi-modality fusion for emotion recognition in the wild," in *ICMI'19: Int. Conf. on Mul-*

timodal Interaction, Suzhou, China, 2019 (Association for Computing Machinery, New York, 2019), pp. 589–594.
https://doi.org/10.1145/3340555.3355719

10. A. Mollahosseini, B. Hasani, and M. H. Mahoor, "AffectNet: A database for facial expression, valence, and arousal computing in the wild," IEEE Trans. Affective Comput. **10**, 18–31 (2017).
https://doi.org/10.1109/TAFFC.2017.2740923
11. O. Perepelkina, E. Kazimirova, and M. Konstantinova, "RAMAS: Russian multimodal corpus of dyadic interaction for affective computing," in *Speech and Computer. SPECOM 2018,* Ed. by A. Karpov, O. Jokisch, and R. Potapova, Lecture Notes in Computer Science, Vol. 11096 (Springer, Cham, 2018), pp. 501–510.
https://doi.org/10.1007/978-3-319-99579-3_52
12. E. Ryumina, O. Verkholyak, and A. Karpov, "Annotation confidence vs. training sample size: trade-off solution for partially-continuous categorical emotion recognition", in *Interspeech 2021* (IEEE, 2021), pp. 3690–3694.
https://doi.org/10.21437/Interspeech.2021-1636
13. A. V. Savchenko, "Facial expression and attributes recognition based on multi-task learning of lightweight neural networks," in *IEEE 19th Int. Symp. Intelligent Systems and Informatics (SISY), Subotica, Serbia, 2021,* Ed. by L. Kovács (IEEE, 2021), pp. 119–124.
https://doi.org/10.1109/SISY52375.2021.9582508
14. A. V. Savchenko, "Personalized frame-level facial expression recognition in video," in *Pattern Recognition and Artificial Intelligence. ICPRAI 2022,* Ed. by M. El Yacoubi, E. Granger, P. C. Yuen, U. Pal, and N. Vincent, Lecture Notes in Computer Science, Vol. 13363 (Springer, Cham, 2022), pp 447–458.
https://doi.org/10.1007/978-3-031-09037-0_37
15. A. V. Savchenko, "Video-based frame-level facial analysis of affective behavior on mobile devices using EfficientNets," in *IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR), 2022*, Ed. by D. Kollias (IEEE, 2022), pp. 2359–2366.
16. A. Savchenko, A. Alekseev, S. Kwon, E. Tutubalina, E. Myasnikov, and S. Nikolenko. "Ad lingua: Text classification improves symbolism prediction in image advertisements," in *Proc. 28th Int. Conf. on Computational Linguistics, Barcelona, 2020*, Ed. by D. Scott, N. Bel, and Ch. Zong (Association for Computational Linguistics, 2020), pp. 1886–1892.
https://doi.org/10.18653/v1/2020.coling-main.171
17. A. V. Savchenko and L. Savchenko, "Speaker-aware training of speech emotion classifier with speaker recognition," in *Speech and Computer. SPECOM 2021,* Ed. by A. Karpov and R. Potapova, Lecture Notes in Computer Science, Vol. 12997 (Springer, Cham, 2021), pp. 614–625.
https://doi.org/10.1007/978-3-030-87802-3_55
18. L. V. Savchenko and A. V. Savchenko, "A method of real-time dynamic measurement of a speaker's emotional state from a speech waveform," Meas. Tech. **64**, 319–327 (2021).
https://doi.org/10.1007/s11018-021-01935-z
19. M. Shahabinejad, Y. Wang, Y. Yu, J. Tang, and J. Li, "Toward personalized emotion recognition: A face recognition based attention method for facial emotion recognition," in *16th IEEE Int. Conf. on Automatic Face and Gesture Recognition (FG 2021), Jodhpur, India, 2021* (IEEE, 2021), pp. 1–5.
https://doi.org/10.1109/FG52635.2021.9666982
20. B. Sonawane, and P. Sharma, "Deep learning based approach of emotion detection and grading system," Pattern Recognit. Image Anal. **30**, 726–740 (2020).
https://doi.org/10.1134/S1054661820040239
21. K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, "Joint face detection and alignment using multitask cascaded convolutional networks," IEEE Signal Process. Lett. **23**, 1499–1503 (2016).
https://doi.org/10.1109/LSP.2016.2603342
22. H. Zhou, D. Meng, Yu. Zhang, X. Peng, J. Du, K. Wang, and Yu Qiao, "Exploring emotion features and fusion strategies for audio-video emotion recognition," in *Int. Conf. on Multimodal Interaction, Suzhou, China, 2019*, Ed. by W. Gao, H. M. Ling Meng, M. Turk, S. R. Fussell, B. Schuller, Ya. Song, and K. Yu (Association for Computing Machinery, New York, 2019), pp. 562–566.
https://doi.org/10.1145/3340555.3355713

Andrey V. Savchenko received the BSc degree in applied mathematics and informatics from Nizhny Novgorod State Technical University, Nizhny Novgorod, Russia, in 2006, the Cand. Sci. degree in mathematical modeling and computer science from the State University Higher School of Economics, Moscow, Russia, in 2010, and the Dr. Sci. degree in system analysis and information processing from Nizhny Novgorod State Technical University in 2016. Since 2008, he has been with the HSE University, Nizhny Novgorod, where he is currently a Full Professor with the Department of Information Systems and Technologies. He is also a Leading Research Fellow with the Laboratory of Algorithms and Technologies for Network Analysis and academic supervisor of the Master of Computer Vision programme at HSE University. He has authored or co-authored one monograph and more than 50 articles. His current research interests include statistical pattern recognition, image classification, and biometrics.

Lyudmila V. Savchenko received the Specialist degree in applied mathematics and informatics from Nizhny Novgorod State Technical University, Nizhny Novgorod, Russia, in 2008, the Cand. Sci. degree in system analysis and information processing from Voronezh State Technical University in 2017. Since 2018, she has been with the HSE University, Nizhny Novgorod, where she is currently an Associate Professor with the Department of Information Systems and Technologies. Her current research interests include speech processing and e-learning systems.