ISSN 1060-992X, Optical Memory and Neural Networks, 2022, Vol. 31, No. 3, pp. 237–244. © Allerton Press, Inc., 2022.

Neural Network Model for Video-Based Analysis of Student's Emotions in E-Learning

A. V. Savchenko[a], * and I. A. Makarov[b], **

[a] *Laboratory of Algorithms and Technologies for Network Analysis, Higher School of Economics (HSE) University, Nizhny Novgorod, 603093 Russia*

[b] *Artificial Intelligence Research Institute (AIRI), Moscow, 117246 Russia*

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

***e-mail: makarov@airi.net*

Received March 10, 2022; revised May 30, 2022; accepted May 31, 2022

Abstract—In this paper, we consider a problem of an automatic analysis of the emotional state of students during online classes based on video surveillance data. This problem is actual in the field of e-learning. We propose a novel neural network model for recognition of students' emotions based on video images of their faces and use it to construct an algorithm for classifying the individual and group emotions of students by video clips. At the first step, it performs detection of the faces and extracts their features followed by grouping the face of each student. To increase the accuracy, we propose to match students' names selected with the aid of the algorithms of the text recognition. At the second step, specially learned efficient neural networks perform the extraction of emotional features of each selected person, their aggregation with the aid of statistical functions, and the subsequent classification. At the final step, it is possible to visualize fragments of the video lesson with the most pronounced emotions of the student. Our experiments with some datasets from EmotiW (Emotion Recognition in the Wild) show that the accuracy of the developed algorithms is comparable with their known analogous. However, when classifying emotions, the computational performance of these algorithms is higher.

Key words: image processing, online learning, emotion classification on video, face clustering, text recognition on images

DOI: 10.3103/S1060992X22030055

1. INTRODUCTION

Due to influence of the COVID-19 pandemics, these days we see an explosive growth of education technologies and e-learning. A lot of massive open online courses (MOOC) appear and all around the world the majority of universities and educational institutions moved many classes to online format. Unfortunately, it is very difficult for teachers to control student engagement in an on-line lecture in the same way as well as in the case of an offline learning course [1]. Indeed, during the lecture all microphones except the teacher's one must be muted. This is necessary to exclude an interactive feedback when the majority of students lose their attention with the learning material the teacher presents. Even small video images of each student face on the screen cannot help the teacher during his lecture, especially when the number of students is relatively large [2]. Despite the possibilities to estimate the engagement of students based on their interaction with mobile devices, contemporary analysis shows that automatic algorithms provide the most suitable solution of this problem [1].

Although student emotions and their engagement play a decisive role in the on-line learning process [3], frequently one cannot implement the known very accurate emotion recognition methods for real-time video analysis especially when teachers use typical (low-performance) equipment (laptop or mobile device) [4, 5]. This is a reason why in this paper we propose a novel efficient neural network model for analysis of the dynamics of the emotional state of each student as well as for prediction of emotions of the whole group of students. This can be helpful when it is necessary to define difficult parts of the lecture. The goal of the developed algorithm is to detect the moments of the video lesson, which caused the most pronounced student emotions based on the analysis of video images of their faces. We hope that our results and conclusions will be helpful for a wide range of specialists in the field of the computer vision and pattern recognition.

2. LITERATURE REVIEW

Algorithms for analyzing the students' behavior and detecting their engagement in the online environment are the subject of an intensive study in the field of the data mining for education. In the survey [2], the authors discuss the dependencies of the present methods on the degree of the students' participation and propose to classify these methods into three categories: automatic, semi-automated and performed by the lecturer/teacher. Due to the widest range of applications, they focus on the automatic methods based on the computer vision.

Recognition of the student emotions can influence significantly the quality of the large number of the e-learning systems. The authors of [3] claim that the multimodal emotion recognition based on the analysis of the facial expressions, body gestures and students' messages, provides better performance than the single-modal. In [6], the same approach was used not only for the online learning, but also for processing video recordings of the situation in the classroom.

To estimate different methods of the emotion classification on static images people frequently use AffectNet dataset [7] that contains 287651 face images with 8 emotions (anger, contempt, disgust, fear, happiness, sadness, surprise, and neutral state), as well as 500 test images for each class. In [8], for this high accuracy dataset we proposed several neural network architectures acceptable for implementation on mobile devices. As a rule, the progress in the audiovisual emotion recognition is estimated using the database AFEW (Acted Facial Expression in the Wild) from EmotiW challenge [9]. The baseline of the authors of this set is the calculation of the LBP-TOP (Local Binary Pattern—Three Orthogonal Planes) features for each video frame separately and applying the resultant vectors for training the support vector machine (SVM) to classify emotions. The accuracy of this approach reaches 38.9% on the validation part of AFEW 8.0. A convolutional neural network (CNN) DenseNet-161 pretrained on large additional databases for the face recognition extracts the feature vectors for each video frame and then calculates a descriptor of the entire video using the concatenation of the vectors of the statistical functions (mean, maximum, minimum, and standard deviation) [10]. The noisy student technique for the CNN training with simultaneous using of an additional body language dataset provided one of the best single models for the dataset AFEW [11]. The best accuracy on the validation set is reached when generating an ensemble of attention-based models for the factorized bilinear pooling of the features in the multi-modal emotion recognition [12]. However, this model has a slightly lower accuracy on the test set comparing with the winner of the EmotiW 2019 challenge, which performs the bimodal analysis of the audio and video features extracted by four different CNNs [13].

The predicted emotions are useful not only when we have to understand the behavior of each student but also for a summarization of videos of all the students in the class [14] or for recognition of group emotions of all the students. The appearance of the VGAF (Video-level Group AFfect) dataset [15] made it possible to analyze the last problem in detail. For this set, one can achieve a relatively high accuracy through action recognition and application of the K-injection neural networks [17]. The authors of [18] discussed the generation of an ensemble of the space-time and static characteristics (face, body, all the frame, and so on). The winners of the EmotiW 2020 challenge proposed an ensemble of hybrid neural networks for audio, facial emotions, video stream, environmental object statistics, and "competing" streams of detectors [19].

Thus, presently a growing number of researchers propose to apply different methods of the emotion recognition on video to analyze the behavior of students of online courses and online classes. In this case, the majority of arising problems can be tested with the aid of the datasets from EmotiW 2018–2020 challenges. Unfortunately, typically the winners of these challenges propose high-frequency methods based on large ensembles of deep CNNs and multimodal features of voices, faces, body postures, and so on [12, 19]. As a result, they are not suitable for many practical applications when a real-time processing with the aid of a low-performance equipment is necessary. This means that new computationally effective and high accurate algorithms for the emotion recognition of students and their groups need to be developed.

3. PROPOSED APPROACH

In the present paper, to extract features of the selected tracks and clusters of faces we apply a multi-task neural network [20] adapted to the problems of emotion recognition. At first, we use a traditional approach: a basic CNN is pretrained to recognize emotions with the aid of a very large VGGFace2 dataset [21]. Although traditionally they use a center crop of the area 224×224 as pre-processing of each image, here we train the CNN to recognize faces cropped with the aid of the MTCNN detector (multi-task cascaded convolutional neural network) [22] without any additional margins.

Since the aim of this paper is to develop computationally efficient algorithms for video processing, we decided to use such architectures as MobileNet v1, EfficientNet-B0 and EfficientNet-B2. The first generated neural network (CNN-1) extracts the facial identity features suitable to distinguish one person from another in a photograph. These features allow us to predict attributes stable for a given person (for example, the gender and ethnicity) using a simple classifier such as one fully connected layer. It is important to highlight that other face attributes change quickly so that the facial features obtained in the face recognition have to remain identical with respect to such changes. The problem of the emotion recognition is an example: an interclass distance between facial features of the same person with different emotions should remain much lower than an intraclass distance between photographs of the faces of different persons even with the same facial expression. Consequently, to recognize emotions we cannot use directly the facial features extracted by the CNN-1 trained to solve the identification problems. In the same time, the first layers of such CNN can extract low-level features more suitable for face processing than the neural network pretrained on a dataset such as, for example, ImageNet-1000, which is not related to faces.

This is the reason why in the present paper our CNN generated to identify faces was retrained to recognize emotions using the AffectNet dataset [7]. Since it is not balanced, in the course of training we used a weighted categorical cross entropy loss function

$$L(X,y) = -\text{softmax}(z_y)\max_{c\in\{1,\ldots,C_e\}}\frac{N_c}{N_y},$$

where X is an image from the training set, $y \in \{1,\ldots,C_e\}$ is it the emotion class label, N_y is the total number of training examples of the yth class, z is the output of the yth neuron of the next-to-last layer, and softmax activation function is used in the last layer. In the course of the training, we replace the last layer with a new fully connected layer with C_e outputs whose weights we trained with the aid of the optimizer Adam (learning rate 0.001) during 3 epochs. Finally, the weights of the entire neural network were trained with the aid of SAM (Sharpness-aware minimization) and Adam (learning rate 0.001) during 6 epochs.

When we feed the face images to the input of the thus obtained CNN-2, at the output of the penultimate layer we obtain a D-dimensional vector $\mathbf{x}$ of the emotion characteristic features, which we can use when processing video data. In the Figure, we present the proposed model for emotion recognition on video; in Table 1, we describe the algorithm of the group emotion recognition on video.

We receive the entire video lesson from an online video conferencing tool (for example, Zoom, MS Teams, Google Meet, etc.) to the input of the MTCNN face detector. In addition, on each video frame the text detection and recognition take place using, for example, the Tesseract library. Then the algorithm feeds all the face images to the input of the two CNN (CNN-1 and CNN-2) trained to recognize faces [20] and to classify emotions on static images [8], respectively. The algorithm uses these features to track and group face areas of the same students.

The algorithm combines the emotional features of one student on a video 5–10 s long to recognize individual emotions of each student. Furthermore, it classifies the emotions of the whole group. We use a statistical module (STAT encoding) to aggregate the features of all the videos. This module allows us to combine the results of several statistical functions (such as minimum, maximum, mean value, and standard deviation) calculated component by component when using the features of all the video frames. As a result, we obtain a descriptor, which contains information about all the input frames [23]. We apply the L_2-norm to the descriptors and use the result to train the SVM with linear kernel that predicts one of the classes of emotions.

Emotions of individual students can be visualized in the form of short fragments of a video lesson. For example, we can consider time intervals when according to our prediction a strong emotion has to take place. In the file video_summarizer.ipynb of our repository [24], you can find an example of such visualization for a real lesson. To help a teacher to organize the lesson materials better, the algorithm can draw a dependence of the predicted emotions as function of time and choose the most interesting or the most difficult for understanding parts of the lesson.

4. EXPERIMENTAL RESULTS

In our first experiment, to predict emotions of a person from a video image of his face we used the AFEW dataset [9] that consisted of short video clips selected from different films/serials. The training and validation sets contained 773 video clips and 383 clips, respectively. There were seven categories of emotions in the dataset. Each clip belonged to one of the six emotions (angry, disgust, fear, happiness, sadness, surprise) or to a neutral category. To obtain the descriptor of the whole video we combined the mean,

Table 1. Proposed algorithm for prediction of individual and group emotions of students by short fragments of facial videos

1. Initialize two-dimensional lists of identity $X^{(\mathrm{id})} = [[]]$ and emotional features $X^{(\mathrm{emo})} = [[]]$ as well as list of found names $S^{(\mathrm{txt})} = [[]]$.
2. Repeat for each frame $t = 1, 2, ..., T$:
   2.1. Use face detector to find $M(t)$ facial regions (draw rectangles around faces).
   2.2. Use text detector to find combinations of words (potential names/surnames of participants).
   2.3. Repeat for each found face area $m = 1, 2, ..., M(t)$:
      2.3.1. Find words or phrases S closest to the mth area.
      2.3.2. Feed face image to input of CNN-1 trained for face identification and extract identity features $\mathbf{x}_\mathbf{m}^{(\mathbf{id})}(t)$.
      2.3.3. Feed face image to input of CNN-2 trained for emotion recognition and extract emotional features $\mathbf{x}_\mathbf{m}^{(\mathbf{emo})}(t)$.
      2.3.4. For each element of list $S^{(\mathrm{txt})}$ calculate edit distance $\boldsymbol{\rho}_\mathbf{m}^{(\mathbf{txt})}(t)$ between S and phrases from each element of list.
      2.3.5. Compute index i of list $S^{(\mathrm{txt})}$ corresponding to minimal distance from $\boldsymbol{\rho}_\mathbf{m}^{(\mathbf{txt})}(t)$ to S.
      2.3.6. Find minimal Euclidean distance $\boldsymbol{\rho}_\mathbf{m}^{(\mathbf{id})}(t)$ between vector $\mathbf{x}_\mathbf{m}^{(\mathbf{id})}(t)$ and feature vectors of element from list $X^{(\mathrm{id})}[i]$.
      2.3.7. If minimal distances $\rho_\mathbf{m}^{(\mathbf{txt})}(t)$ and $\rho_\mathbf{m}^{(\mathbf{id})}(t)$ are larger than predefined thresholds then
         2.3.7.1. Create a new cluster by appending lists $\left[\mathbf{x}_\mathbf{m}^{(\mathbf{id})}(t)\right]$, $\left[\mathbf{x}_\mathbf{m}^{(\mathbf{emo})}(t)\right]$ and $[S]$ to ends of lists $X^{(\mathrm{id})}$ $X^{(\mathrm{emo})}$ and $S^{(\mathrm{txt})}$, respectively.
      otherwise
         2.3.7.2. For ith cluster append $\mathbf{x}_\mathbf{m}^{(\mathbf{id})}(t)$, $\mathbf{x}_\mathbf{m}^{(\mathbf{emo})}(t)$ and S to lists $X^{(\mathrm{id})}[i]$, $X^{(\mathrm{emo})}[i]$ and $S^{(\mathrm{txt})}[i]$, respectively.
   2.4. To obtain descriptor $\mathbf{x}^{(\mathrm{emo})}(t)$ apply component-wise STAT functions (mean and standard deviation) to each of D elements of vectors $\mathbf{x}_\mathbf{m}^{(\mathbf{emo})}(t)$.
3. Repeat for each selected cluster i in list $X^{(\mathrm{emo})}$

   3.1. To get the final descriptor of facial emotions $\mathbf{x}_\mathbf{i}^{(\mathbf{emo})}$ apply component-wise STAT functions (mean, minimum, maximum, and standard deviation) to each of D elements of vectors $X^{(\mathrm{emo})}[i]$.

   3.2. Feed descriptor $\mathbf{x}_\mathbf{i}^{(\mathbf{emo})}$ to input of pretrained classifier, display classes of individual emotions.
4. To get the final descriptor of group emotions $\mathbf{x}^{(\mathrm{emo})}$ apply component-wise STAT functions (mean and standard deviation) to each descriptor of frame $\mathbf{x}^{(\mathrm{emo})}(t)$.
5. Feed descriptor $\mathbf{x}^{(\mathrm{emo})}$ to input of pretrained classifier, display classes of emotions of entire group.

maximum, minimum, and the standard deviation calculated for each feature of the frame extracted using the CNN. Consequently, the dimension of the descriptor of the video was 4 times greater than the dimension D of the emotional features of the frames. If the face was not found on the video from the training set the algorithm ignored it, but it related the video with missed faces from the test set with a zero descriptor of the dimension $4D$. *L2*-normalized descriptors were classified using LinearSVC from scikit-learn with the regularization parameter found by cross-validation on the training set. In Table 2, for different CNNs we present the comparison of the accuracy of our algorithm with the known results.

Here the developed algorithm showed itself as 5% more accurate than other single neural network models. Even the MobileNet improved the accuracy of the best-known technology noisy student for the CNN ResNet [11] by 0.1%. It is worth noting that although the accuracy of the EfficientNet-B2 is much higher when recognizing emotions on static images from the AffectNet database, the EfficientNet-B0 neural network with fewer parameters performed slightly better. Even if the best ensembles [12] are still much more accurate, our algorithm is much faster and it can process the students' emotions in real time even on a mobile device.

In the final experiment, we tested the proposed algorithm of the group-level emotion recognition on videos from the VGAF dataset [15], which contained videos downloaded from YouTube with a creative commons license. These videos varies significantly in terms of their context, the number of people, their quality, and so on. The training set provided by the organizers of the EmotiW 2020 challenge contained 2661 videos and 766 videos were accessible for validation. The problem was to classify each video into three

Table 2. Accuracy (%) of recognition of individual emotions on video for AFEW dataset

Method		Accuracy, %
Committee of classifiers, audio + video	Bimodal committee of 4 CNNs [13]	54.3
	5 FBP models [12]	65.5
Single neural network	LBP-TOP (basic method) [9]	38.90
	Noisy student [11]	55.17
	Proposed algorithm, MobileNet-v1	55.35
	Proposed algorithm, EfficientNet-B0	59.27
	Proposed algorithm, EfficientNet-B2	59.00

Table 3. Accuracy (%) of recognition of group emotions for VGAF dataset

Method		Accuracy, %
Committee of classifiers, audio + video	VGAFNet (face + holistic + audio) [15]	61.61
	K-injection networks [17]	66.19
	Fusion of 14 models [18]	71.93
	Hybrid Networks [19]	74.28
Single neural network	VGAFNet (faces) [15]	60.18
	DenseNet-121 (Hybrid Networks) [19]	64.75
	Self-attention K-injection network [17]	65.01
	Slowfast [18]	68.57
	Proposed algorithm, MobileNet-v1	68.92
	Proposed algorithm, EfficientNet-B0	66.80
	Proposed algorithm, EfficientNet-B2	70.23

Table 4. Efficiency of models for emotional feature extracting

CNN	Recognition time of one image, ms	Number of parameters, millions
VGG-16	224.7	134.3
ResNet-18	58.7	11.7
Inception-v3	160.4	19.7
SENet-50	128.4	25.5
MobileNet-v1	40.6	3.2
EfficientNet-B0	54.8	4.3
EfficientNet-B2	97.0	7.8

classes that are positive, negative, and neutral. Since each frame in this dataset contains several areas of the face, we firstly combined the mean and the standard deviation of the emotional features of the faces that the CNN selected in each frame. The algorithm calculated the final descriptor of the whole video as a combination of the mean and standard deviation of the characteristics of the frame. In all the videos, the absent faces were processing similarly to previous experiments with the AFEW dataset. This means that the algorithm removed the empty videos from the training set, but to compere the accuracy with the accuracies of the existing models it associated empty videos with zero descriptors. The obtained videos were classified with the aid of SVM with the RBF (Radial Basis Functions) kernel. In Table 3, we present the main obtained results.

In this case, our algorithm leads to better results comparing with the known non-ensemble methods. The best architecture, namely, EfficientNet-B2, increased the accuracy of the best Slowfast network [18] by 1.7%. Even the fastest architecture MobileNet turned out to be 0.35% more accurate than the Slowfast. Although the winners of the EmotiW 2020 challenge are unsurprisingly accurate, the results of the proposed approach remain competitive even when a combination of 14 deep CNNs is applied to different features of audio and video modalities [18].

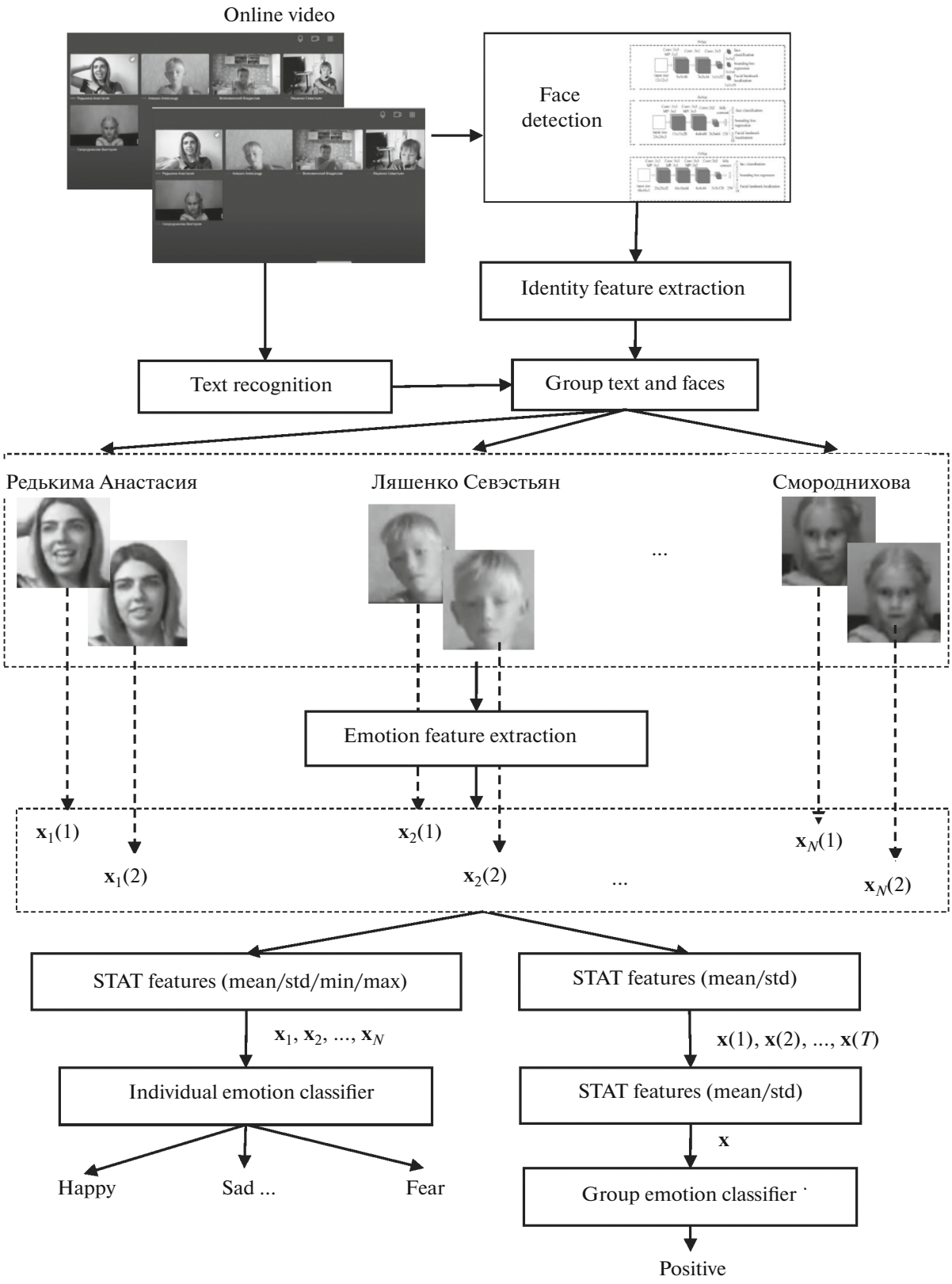

Fig. 1. Proposed video-based emotional recognition model.

In the concluding experiment, we examined the efficiencies of the neural network architectures used for emotion prediction. We measured an average time of a face recognition on the MSI GP63 8RE laptop (Intel Core i7-8750H 2.2 GHz processor, 16 GB RAM); the number of the parameters of each architecture we presented in Table 4. As expected, the number of parameters and the running time of the trained CNNs were comparatively small, although ResNet-18 had a comparable speed.

5. CONCLUSIONS

In the present paper, we propose the neural network model of students' emotions recognition during an online course (Fig. 1) based on the analysis of images of their faces. Implementation of the corresponding algorithms, the model training method with the use of the library PyTorch and a prototype software for visualizing emotions of students are available in the repository [24]. Within a comparable accuracy, the developed algorithms achieve qualitatively higher computational efficiency compared with the best analogous results for computational efficiency of the classification of facial emotions. In the same time, they are suitable for implementation on a low-performance equipment, including mobile devices of lecturers and/or students.

Our experiments showed that the developed algorithm for recognition of individual and group emotions (Table 1) has the highest accuracy compared to the best-known neural network models not united in an ensemble of classifiers (Tables 2, 3). Frequently, the computational efficiency of an ensemble makes it difficult or impossible to realize it in real time, while the efficiency of our neural network models is acceptable even without using graphics accelerators (Table 3). The obtained results are an important step towards improving the quality of the online materials and the analysis of the listeners' involvement in the ongoing online event.

The main direction of the future development of our model is to make it able to use the obtained emotion features for prediction of the students' engagement from video [9]. In addition, a full-scale implementation of the proposed approach for analysis of video conferences in real time is of interest. Finally, as preliminary experiments show, it is not sufficiently accurate to group images of each student basing on the recognition of his name with the aid of the face recognition and clustering technologies [25]. Despite the errors when recognizing surnames, especially Russian (see Fig. 1), the proposed procedure of recognition of the students' names using the Tesseract library allow us to reduce significantly the number of the selected students. This means that in the future it is necessary to investigate the ways of improving the quality of the face grouping. You can do that, for example, by using the tracking technology DeepSORT [26] and its integration with processing of the detected text.

FUNDING

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

CONFLICT OF INTEREST

The authors declare that they have no conflicts of interest.

REFERENCES

1. Bhardwaj, P., Gupta, P., Panwar, H., Siddiqui, M.K., Morales-Menendez, R., and Bhaik, A., Application of deep learning on student engagement in e-learning environments, *Comput. Electr. Eng.,* 2021, vol. 93, p. 107277.
2. Dewan, M.A.A., Murshed, M., and Lin, F., Engagement detection in online learning: a review, *Smart Learn. Environ.*, 2019, vol. 6, no. 1, pp. 1–20.
3. Imani, M. and Montazer, G.A., A survey of emotion recognition methods with emphasis on E-Learning environments, *J. Network Comput. Appl.*, 2019, vol. 147, p. 102423.
4. Savchenko, A.V., Deep neural networks and maximum likelihood search for approximate nearest neighbor in video-based image recognition, *Opt. Mem. Neural Networks,* 2017, vol. 26, no. 2, pp. 129–136
5. Savchenko, A.V., Probabilistic Neural Network with complex exponential activation functions in image recognition, *IEEE Trans. Neural Networks Learn. Syst.*, 2020, vol. 31, Issue 2, pp. 651–660
6. Ashwin, T.S. and Guddeti, R.M.R., Affective database for e-learning and classroom environments using Indian students' faces, hand gestures and body postures, *Future Generation Comput. Syst.*, 2020, vol. 108, pp. 334–348.
7. Mollahosseini, A., Hasani, B., and Mahoor, M.H., AffectNet: A database for facial expression, valence, and arousal computing in the wild, *IEEE Trans. Affective Comput.*, 2017, vol. 10, no. 1, pp. 18–31.

8. Savchenko, A.V., Facial expression and attributes recognition based on multi-task learning of lightweight neural networks, *Proceedings of 19th IEEE International Symposium on Intelligent Systems and Informatics (SISY)*, 2021, pp. 119–124
9. Dhall, A., EmotiW 2019: Automatic emotion, engagement and cohesion prediction tasks, *Proceedings of the International Conference on Multimodal Interaction (ICMI)*, 2019, pp. 546–550.
10. Liu, C. et al., Multi-feature based emotion recognition for video clips, *Proceedings of the ACM International Conference on Multimodal Interaction (ICMI)*, 2018, pp. 630–634.
11. Kumar, V., Rao, S., and Yu, L., Noisy student training using body language dataset improves facial expression recognition, *Proceedings of the European Conference on Computer Vision (ECCV)*, Cham: Springer, 2020, pp. 756–773.
12. Zhou, H. et al., Exploring emotion features and fusion strategies for audio-video emotion recognition, *Proceedings of the ACM International Conference on Multimodal Interaction (ICMI)*, 2019, pp. 562–566.
13. Li, S. et al., Bi-modality fusion for emotion recognition in the wild, *Proceedings of the ACM International Conference on Multimodal Interaction (ICMI)*, 2019, pp. 589–594.
14. Zeng, H. et al., EmotionCues: Emotion-oriented visual summarization of classroom videos, *IEEE Trans. Visual. Comput. Graphics*, 2020, vol. 27, no. 7, pp. 3168–3181.
15. Sharma, G., Dhall, A., and Cai, J., Audio-visual automatic group affect analysis, *IEEE Trans. Affective Comput.*, 2021.
16. Pinto, J.R. et al., Audiovisual classification of group emotion valence using activity recognition Networks, *Proceedings of the 4th IEEE International Conference on Image Processing, Applications and Systems (IPAS),* 2020, pp. 114–119.
17. Wang, Y. et al., Implicit knowledge injectable cross attention audiovisual model for group emotion recognition, *Proceedings of the ACM International Conference on Multimodal Interaction (ICMI)*, 2020, pp. 827–834.
18. Sun, M. et al., Multi-modal fusion using spatio-temporal and static features for group emotion recognition, *Proceedings of the ACM International Conference on Multimodal Interaction (ICMI)*, 2020, pp. 835–840.
19. Liu, C. et al., Group level audio-video emotion recognition using hybrid Networks, *Proceedings of the ACM International Conference on Multimodal Interaction (ICMI)*, 2020, pp. 807–812.
20. Savchenko, A.V., Efficient facial representations for age, gender and identity recognition in organizing photo albums using multi-output ConvNet, *Peer J. Comput. Sci.*, 2019, vol. 5, e197.
21. Cao, Q., Shen, L., Xie, W., Parkhi, O.M., and Zisserman, A., VGGface2: A dataset for recognising faces across pose and age, *Proceedings of the 13th IEEE International Conference on Automatic Face and Gesture Recognition (FG*), 2018, pp. 67–74.
22. Zhang, K. et al., Joint face detection and alignment using multitask cascaded convolutional networks, *IEEE Signal Process. Lett.*, 2016, vol. 23, no. 10, pp. 1499–1503.
23. Savchenko, A. V., Savchenko, L. V., and Makarov I. A., Classifying emotions and engagement in online learning based on a single facial expression recognition neural network, *IEEE Trans. Affective Comput.*, 2022, pp. 1–12.
24. Facial emotion recognition repository. https://github.com/HSE-asavchenko/face-emotion-recognition/.
25. Sokolova, A.D., Kharchevnikova, A.S., and Savchenko, A.V., Organizing multimedia data in video surveillance systems based on face verification with convolutional neural networks, *Proceedings of International Conference on Analysis of Images, Social Networks and Texts (AIST)*, Cham: Springer, 2017, pp. 223–230
26. Veeramani, B., Raymond, J.W., and Chanda, P., DeepSort: deep convolutional networks for sorting haploid maize seeds, *BMC Bioinform.*, 2018, vol. 19, no. 9, pp. 1–9.