Classifying Emotions and Engagement in Online Learning Based on a Single Facial Expression Recognition Neural Network

Andrey V. Savchenko, Lyudmila V. Savchenko, and Ilya Makarov

Abstract—In this article, behaviour of students in the e-learning environment is analyzed. The novel pipeline is proposed based on video facial processing. At first, face detection, tracking and clustering techniques are applied to extract the sequences of faces of each student. Next, a single efficient neural network is used to extract emotional features in each frame. This network is pre-trained on face identification and fine-tuned for facial expression recognition on static images from AffectNet using a specially developed robust optimization technique. It is shown that the resulting facial features can be used for fast simultaneous prediction of students' engagement levels (from disengaged to highly engaged), individual emotions (happy, sad, etc.,) and group-level affect (positive, neutral or negative). This model can be used for real-time video processing even on a mobile device of each student without the need for sending their facial video to the remote server or teacher's PC. In addition, the possibility to prepare a summary of a lesson is demonstrated by saving short clips of different emotions and engagement of all students. The experimental study on the datasets from EmotiW (Emotion Recognition in the Wild) challenges showed that the proposed network significantly outperforms existing single models.

Index Terms—Online learning, e-learning, video-based facial expression recognition, engagement prediction, group-level emotion recognition, mobile devices

1 Introduction

There is an explosive growth of education and electronic learning (e-learning) technologies due to the impact of the COVID-19 pandemic [1]. A lot of new Massive Open Online Courses (MOOCs) have been recently appeared [2]. Moreover, many universities and educational institutions all over the world have moved many classes to an online format. There exist many essential factors for an effective e-learning environment, such as different elements related to assessing the value of e-learning and the need for sufficient tools available at the university for remote teaching [3].

One of the key challenges in online learning is a difficulty for a teacher to control the engagement of students in an online lecture similarly to traditional offline education [4]. Indeed, all the microphones except the educator's one should be muted during a lecture so that it is impossible to provide interactive feedback when most students become disentangled and/or noisy. While delivering an *offline* lecture, a teacher can use student's individual of group-level emotions to slow down or modify the presentation [5], but the presence of small facial videos of each student cannot help a teacher during an *online* lecture when the number of listeners is relatively high [6]. Although there exist some studies of the mobile-based formative assessment of the student's engagement [7], it seems that only automatic detection of students engagement is the most appropriate solution [8].

It is important to emphasize that engagement is not the only factor that is essential for an e-learning environment. For example, the dynamics of the emotional state of each student play an important role in the learning process [9]. Analysis of the group-level affect may be important to find the difficult or weird parts of the lecture [10]. It is known that all these tasks are connected to each other, because engagement and affect are linked to increased learning gain and productivity [6]. Moreover, the authors of the article [5] demonstrated that facial expressions of the students are significantly correlated to their comprehension towards the lecture.

Most of the above-mentioned tasks were considered in several sub-challenges of Emotion Recognition in the Wild (EmotiW) contests. Their winners proposed rather accurate techniques that are typically based on large ensembles of deep convolutional neural networks (CNNs) [11], [12] and multi-modal features of audio, faces and body pose [13], [14], [15]. As a result, they cannot be used in many practical applications with requirements to real-time processing in a low-resource environment. Moreover, to guarantee the privacy of a student, it is preferable to process the facial videos directly on his or her personal (often, mobile) device [16].

- *Andrey V. Savchenko and Lyudmila V. Savchenko are with the Laboratory of Algorithms and Technologies for Network Analysis, HSE University, 603155 Nizhny Novgorod, Russia. E-mail: {avsavchenko, lsavchenko}@hse.ru.*
- *Ilya Makarov is with the Artificial Intelligence Research Institute (AIRI), 105064 Moscow, Russia, and also with the Big Data Research Center, National University of Science and Technology MISIS, 119991 Moscow, Russia. E-mail: makarov@airi.net.*

Manuscript received 6 December 2021; revised 28 April 2022; accepted 29 June 2022. Date of publication 4 July 2022; date of current version 15 November 2022.
The publication was supported by the grant for research centers in the field of AI provided by the Analytical Center for the Government of the Russian Federation (ACRF) in accordance with the agreement on the provision of subsidies (identifier of the agreement 0D730321P5Q0002) and the agreement with HSE University No. 70-2021-00139.
(Corresponding author: Andrey V. Savchenko.)
Recommended for acceptance by M. Mahoor.
Digital Object Identifier no. 10.1109/TAFFC.2022.3188390

Thus, the aim of this article is a development of fast and accurate technique for classifying emotions and engagement that can be implemented in online learning software at laptops without powerful GPUs (graphics processing units) and/or mobile devices of students and teachers. The main contribution consists of the following:

- Lightweight FER (facial expression recognition) models based on EfficientNet and MobileNet architectures for emotional features extraction from facial images. It is proposed to borrow the idea of robust data mining [17] to modify the softmax loss function for the training of this model to predict emotions on static images.
- Efficient neural network model for simultaneous engagement detection and recognition of individual- and group-level emotions in facial videos. Lightweight CNN from the previous item extracts unified emotional features of each frame on the learner's device. The features of several frames are aggregated into a video descriptor using statistical (STAT) functions (mean, standard deviation, etc.) [18]. The resulting models let us reach the state-of-the-art results in several emotion recognition and engagement detection tasks.
- A novel technological framework for real-time video-based classification of emotions and engagement in online learning by using only facial modality. The engagement and individual emotions of each student are predicted on the device of each student. Obtained emotional feature vectors may be sent to the teacher's device to classify the emotions of the whole group of students. If the faces of some students in the online video conferencing tool (Zoom, MS Teams, Google Meet, etc.) are turned on, it is proposed to additionally cluster these faces and summarize their emotions and engagement during the whole lesson into short video clips [19]. This helps the lecturers to understand their own weakness and to change it [5]. The sources of training and testing code using Tensorflow 2 and Pytorch frameworks together with demo Android application and several models are made publicly-available[1].

The remaining part of this article is structured as follows. Section 2 contains a brief survey of related articles. The details of the proposed framework are given in Section 3. Section 4 provides experimental results of our models on EngageWild [8], AFEW (Acted Facial Expression In The Wild) [20] and VGAF (Video-level Group AFfect) [10] datasets from EmotiW challenges. Finally, concluding comments and future work are discussed in Section 5.

2 Literature Survey

2.1 Video-Based Emotion Recognition

Recognition of students' emotions may have a great impact on the quality of many e-learning systems. The authors of the review [9] claimed that the multi-modal emotion recognition based on a fusion of facial expressions, body gestures and user's messages provide better efficiency than the single-modal ones. Similar features have been used in [21] for offline learning and videos of classroom environments. It is known that facial emotions, which are a form of non-verbal communication, can be used to estimate the learning affect of a student and enhance the current e-learning platforms [22]. Hence, in this article, it was decided to deal only with an analysis of facial modality.

The FER models are typically pre-trained on single images from a rather large dataset, such as AffectNet [23]. Excellent results have been recently obtained by using supervised learning (SL) and self-supervised learning (SSL) [24] of EfficientNets [25], visual transformers and attentional selective fusion [26], relation-aware transformers (TransFER) [27] and the lightweight models with careful pre-training on the face recognition datasets [28]. Very accurate EmotionGCN explores emotional dependencies between facial expressions and valence-arousal by training the graph convolutional networks in the multi-task learning framework [29].

The progress in the video-based FER is mainly measured on various versions of the AFEW dataset from EmotiW 2013-2019 challenges [20]. One of the best single models is obtained via the noisy student training using body language [30], while the old method with the STAT aggregation of features extracted by three CNNs (VGG13, VGG16 and ResNet) [18] is still one of the best ensemble-based techniques. The best validation accuracy is achieved by the attention cross-modal feature fusion mechanisms that highlight important emotion feature by exploring feature concatenation and factorized bi-linear pooling (FBP) [31]. However, the latter model has slightly lower accuracy on the testing set when compared to bi-modality fusion [32] of audio and video features extracted by four different CNNs.

Predicted emotions can be used not only for understanding the behaviour of each learner but also for visual summarization of classroom videos [19] or classification of the group-level emotions on videos. The latter task has become studied since the appearance of the VGAF dataset [10]. Rather high accuracy is achieved by activity recognition and K-injection networks [33], [34]. The winner of the EmotiW 2020 Audio-video group emotion recognition sub-challenge developed an ensemble of hybrid networks for audio, facial emotion, video, environmental object statistics and fighting detector streams [14].

2.2 Automatic Engagement Detection in E-Learning Systems

Parental involvement, interaction, and students' engagement are the key factors that may influence online learning effects [1]. Though most e-learning techniques are focused on improving the learners' interaction, the algorithms of behavioural analysis and engagement detection have become recently studied in educational data mining [35]. Researchers do not have a consistent understanding of the definition of learning engagement and regard it as a multidimensional concept [36]. In this article, a special type of the student's persistent effort to accomplish the learning task [8] is considered, namely, the emotional engagement. It focuses on the extent of positive and negative reactions, feeling of interest towards a particular theme and enjoying learning about it [36].

[1] https://github.com/HSE-asavchenko/face-emotion-recognition

A survey [6] considered the dependencies of existing methods on learners' participation and classified them into automatic, semi-automatic (engagement tracing) and manual categories. The most popular one is still the latter category. It includes self-reports, observational checklists and rating scales, and typically requires a great deal of time and effort from observers [36]. As a result, the recent research focus has shifted to the automatic engagement detection that infers the social cues of engagement/disengagement from facial expressions, body movements and gaze pattern [8]. Particular attention is paid to the FER-based methods due to simplicity of their usage [6]. Indeed, the FER and engagement prediction tasks are strongly correlated [5]. For example, a lecturer use students' facial expressions as valuable sources of feedback. Moreover, the emotions of the lecturers kept the students motivated and interested during the lectures [37].

One of the first techniques that applied machine learning and FER to predict student's engagement, was proposed in [38]. Their experiments with support vector machines (SVM) with Gabor features and regression for expression outputs from the Computer Expression Recognition Toolbox proved that the automated engagement detectors perform with comparable accuracy to humans. Traditional computer vision for FER was used in [36], where the adaptive weighted histograms of eight-bit gray codes calculated by Local Gray Code Patterns (LGCP) were classified by SVM. The authors of the latter article introduced two datasets for learning engagement detection based on facial and mouse movement data, but they are not publicly available.

Nowadays, the progress of deep learning caused the widespread use of CNNs. For example, the Mean Engagement Score was proposed in [4] by analyzing the results from facial landmark detection, emotional recognition and the weights from a special survey. The non-contact engagement prediction in unconstrained environments is applied not only in e-learning but in other interactive tasks, such as gaming [39]. The framework of learning engagement assessment [2] timely acquired the emotional changes of the learners using a special CNN trained based on domain adaptation, which is suitable for the MOOC scenario.

The rapid growth of studies in engagement prediction has begun from the introduction of the EngageWild dataset [8] in EmotiW 2018-2020 challenges. This dataset contains facial videos with corresponding engagement labels of the user's, while they are watching educational videos such as the ones in MOOCs. The gaze, head pose and action unit intensities features from the OpenFace library [40] were concatenated into the Gaze-AU-Pose (GAP) descriptor [41]. Its classification using the GRU (gated recurrent unit) networks leads to MSE (mean square error) on the validation set, which is 0.03 lower when compared to the baseline solution for the OpenFace features [8]. The usage of the dilated Temporal Convolutional Network (TCN) classifier [42] for similar OpenFace features led to slightly lower MSE 0.0655. The best results on the testing set in the 2018 challenge were obtained by additional LBP-TOP facial descriptor and C3D action features [43].

The authors of the latter approach improved it for the EmotiW 2019 challenge by using the classical bootstrap aggregation and designing a rank loss as a regularization which enforces a distance margin between the features of distant category pairs and adjacent category pairs [44]. The anti-overfitting strategy with training on the overlapped segments of inputs videos was proposed in [45]. Solution of the winners [46] used the facial behaviour features extracted by OpenFace and ResNet-50 model pre-trained on face identification on the large VGGFace2 dataset [47]. These results were improved in the 2020 challenge by using an attention-based GRU and multi-rate video processing [13].

3 Materials and Methods

3.1 Proposed Approach

Most of the techniques mentioned in the previous section used complex ensemble models and various sets of features to boost their performance. Unfortunately, every single model for one feature set reported in these articles cannot compete with the final solutions. Thus, in this article, the novel technological framework (Fig. 1) is proposed to analyse the behaviour of students at online lectures.

Here each student may launch an application on his or her device to provide the results of behaviour analysis without the need to share the facial video. As a result, the high level of data privacy may be achieved because the video of a face is not required to be sent to the remote server or teacher's PC. In this case, the largest facial region is located in each t-th video frame in the unit "Face detection 1" by using any fast technique, such as MTCNN (multi-task CNN). Next, the emotional features $\mathbf{x}(t)$ of an extracted face are obtained in the unit "Emotional feature extraction 2" [48] using the proposed lightweight CNN trained to classify emotions on static images. The details about the training of this neural network will be provided in the next subsections. Finally, the features of several sequential frames with duration 5-10 seconds are aggregated using STAT functions to classify engagement level and individual emotions in the units "Engagement prediction 3" and "Emotion recognition 4", respectively. Predicted engagement level (from disengaged to highly engaged) and individual emotions (happy, angry, sad, neutral, etc.) for each time frame at the output of units 3 and 4 together with the emotional features at the output of unit 2 are sent to the teacher's PC. As the models in first four units are very efficient, the inference can be launched even on any low-resource environment, such as a mobile device of the learners [16], [28].

The most difficult processing is implemented on the teacher's device in units 5-13. These steps may run in offline mode after obtaining the recording of the whole lecture in the online video conferencing tool. This video is fed into the "Face detection 5", which works similarly to the first unit on the student's device, but can return several ($K \geq 1$) faces of learners who agreed to transfer their videos. It is still possible that several extracted facial images have very low resolution for accurate emotion recognition [49]. In this article, the simplest solution is implemented, so that the faces with size lower than a predefined threshold (64x64 pixels) are ignored. Next, the emotional feature vector $\mathbf{x}_k(t)$ of every k-th face are obtained in "Emotional feature extraction 6", which may use either the same CNN as in unit 2 or more complex architecture if the processing is implemented on a

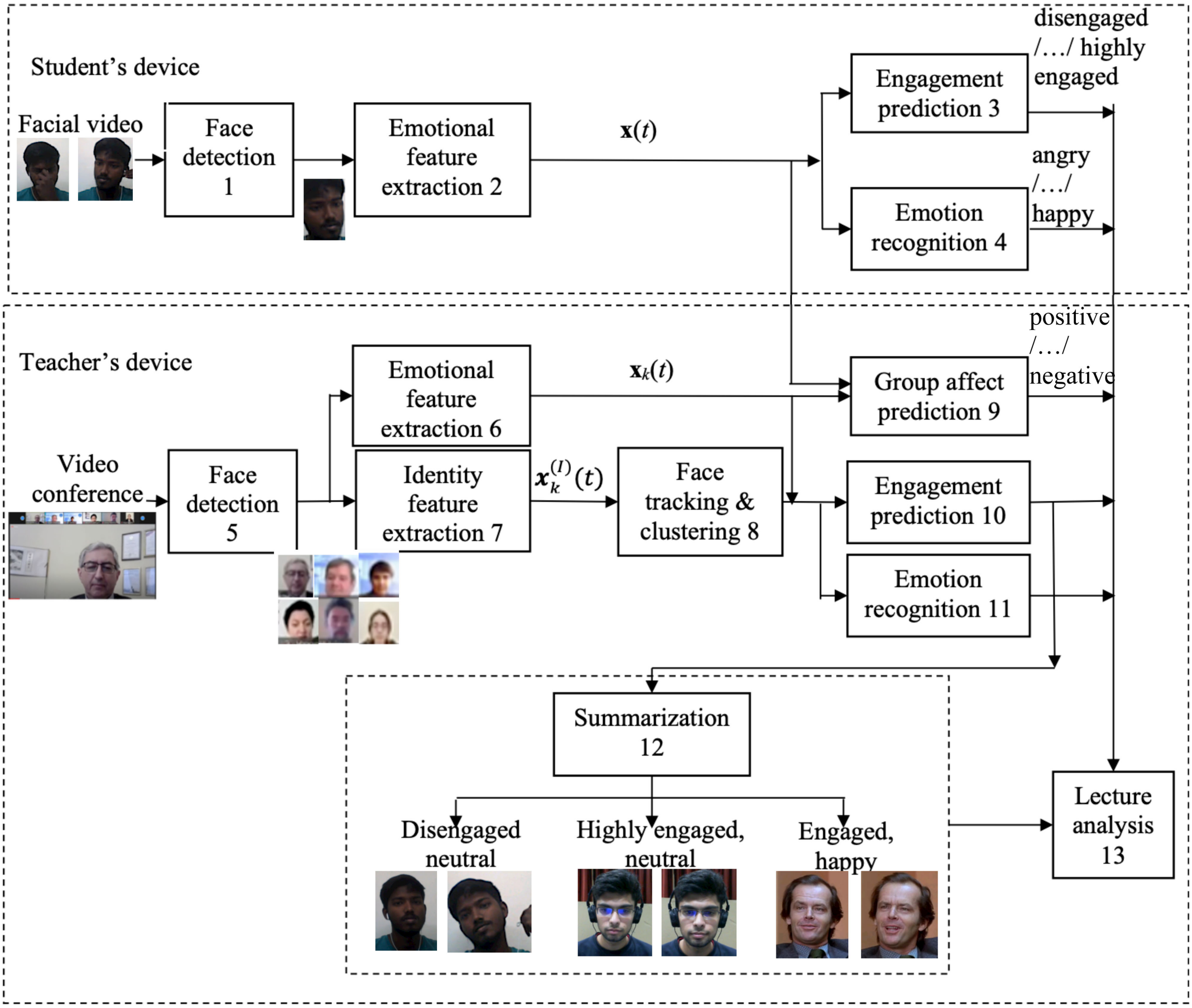


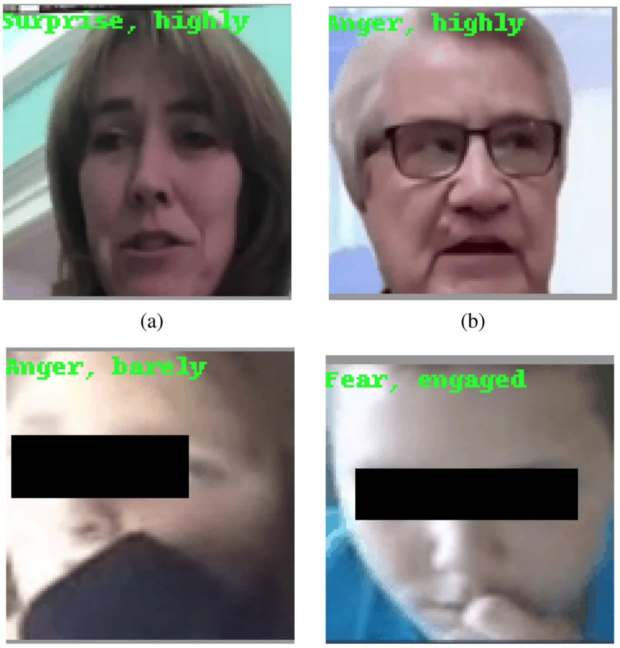
Fig. 1. Proposed pipeline.

rather powerful PC. The identity features $\mathbf{x}_k^{(I)}(t)$ are extracted in unit 7 by using appropriate face recognition CNN [28], [47], [50]. The latter features are used to track and group the facial regions of the same students in the unit "Face tracking & clustering 8". The emotional features $\mathbf{x}_k(t)$ at the output of unit 8 of the same track are combined to solve the down-stream tasks in units 9, 10 and 11. The unit "Group affect prediction 9" first aggregates emotional features of faces from the same frame into single frame features of the whole group of students. Next, all frame features during 5-10 seconds of a video are combined into a single descriptor which can be fed into an appropriate classifier. The 'Engagement prediction 10" and "Emotion recognition 11" work identically to units 3 and 4, but repeat the processing for every k-th face and each group of frames.

Finally, the emotions and engagement of individual students can be summarized into short videos clips and visualized in the unit "Summarization 12". For example, it is possible to take the time points where the strong emotion is predicted. The typical results for several real conferences or lessons are presented in Fig. 2. Another opportunity is the grouping of different emotions based on Russel's 2D space of affect [51] that can give the teacher an initial impression about how students are concentrated, affected and inspired during the e-lesson. Finally, a short GIF with different emotions and engagement during a lesson can be sent to a learner of his or her relatives to increase the parental involvement [1]. Moreover, these clips together with the charts of predicted students' emotions, group affect and engagement depending on time are stored in the unit "Lecture analysis 13" which can help the online educators to detect their online learners' engagement status precisely [6] and to better organize his or her materials. It is also possible to highlight the time points with either high or very low engagement to find the weird or difficult parts of the lecture. Such data let track the efficiency of lessons and increase conversion of online courses.

3.2 Robust Optimization of the FER Network

In this subsection, let us describe the procedure to train the FER network that extracts robust emotional features from either static photos or video frames. At first, a lightweight CNN is trained for face identification on a very large dataset [47]. Next, this network is fine-tuned on any emotional dataset with static facial photos [23]. As existing emotional datasets are typically highly imbalanced, the weighted



[redacted]
[redacted]

(a) (b)

(c) (d)

Fig. 2. Sample frames from the summarized conference presentation (a),(b) and lesson (c),(d). The faces of children are partially hidden.

categorical cross-entropy (softmax) loss is optimized [23], [28]:

$$\mathcal{L}_\theta(X,y) = -\log\left(w_y \cdot softmax(z_y(X;\theta))\right), \tag{1}$$

where θ is the vector of weights of a neural network, X is the training image, $y \in \{1,\ldots,C\}$ is its emotional label, $z_y(X;\theta)$ is the y-th output of the penultimate (logits) layer of the CNN with input X, and $softmax$ is the softmax activation function.

Here the class weights are defined inverse proportional to the total numberN_y of training examples of the y-th class

$$w_y = \max_{c\in\{1,\ldots,C\}} \frac{N_c}{N_y}. \tag{2}$$

To improve the quality of trained emotional models, it is proposed to use robust data mining [17] and formulate the optimization task as follows:

$$\min_\theta \max_{||\boldsymbol{\Delta}\theta||<\epsilon} \mathcal{L}_{\theta+\boldsymbol{\Delta}\theta}(X,y). \tag{3}$$

It is known that the gradient vector gives the direction of maximum increase of a function. Hence, robust optimization task (3) can be simplified as follows:

$$\min_\theta \mathcal{L}_{\theta+\epsilon\tilde{\nabla}\mathcal{L}_\theta}(X,y), \tag{4}$$

where $\tilde{\nabla}\mathcal{L}_\theta$ is defined as the L_2 normalized gradient of the loss function:

$$\tilde{\nabla}\mathcal{L}_\theta = \frac{1}{||\nabla\mathcal{L}_\theta||}\nabla\mathcal{L}_\theta. \tag{5}$$

The optimization task (4) is solved by using a modification of any stochastic gradient descent method. The complete optimization procedure is shown in Algorithm 1. It modifies the Adam optimizer [52] in the notation from the classical book [11]. In addition to parameters of Adam [52], namely, learning rate η and exponential decay rates for moment estimates ρ_1 (default value 0.9) and ρ_2 (0.999 by default), the parameter ϵ is added to control the level of uncertainty.

Algorithm 1. Adam-Based Robust Optimization of FER Neural Network

```
Require: Weights $\theta$ of a CNN pre-trained on face recognition task, training set $\{(X_m, y_m)\}$ of facial images with emotional labels
Ensure: Weights $\theta$ that optimize the robust loss (3)
Initialize accumulators $\boldsymbol{s} := 0, \boldsymbol{r} := 0$ and time step $t = 1$
for $epoch \in \{1,\ldots,NumEpochs\}$ do
  for $batch \in \{1,\ldots,NumBatches\}$ do
    Sample mini-batch of $M$ examples $\{(X_m, y_m)\}$
    for $m \in \{1,\ldots,M\}$ do
      Feed image $X_m$ into a CNN with weights $\theta$ and obtain the $y_m$-th output $p_m^{(1)} := softmax(z_{y_m}(X_m;\theta))$
    end for
    Compute gradient using backprop $\boldsymbol{g}_1 := -\frac{1}{m}\sum_{m=1}^{M}(w_{y_m} \cdot \nabla \log p_m^{(1)})$
    Compute weights $\theta_\epsilon := \theta + \epsilon\boldsymbol{g}_1/||\boldsymbol{g}_1||$
    for $m \in \{1,\ldots,M\}$ do
      Feed image $X_m$ into a CNN with weights $\theta_\epsilon$ and obtain the $y_m$-th output $p_m^{(2)} := softmax(z_{y_m}(X_m;\theta_\epsilon))$
    end for
    Compute gradient using backprop $\boldsymbol{g}_2 := -\frac{1}{m}\sum_{m=1}^{M}(w_{y_m} \cdot \nabla \log p_m^{(2)})$
    Assign $t := t + 1$
    Assign $\boldsymbol{s} := \rho_1 \cdot \boldsymbol{s} + (1-\rho_1) \cdot \boldsymbol{g}_2$
    Assign $\boldsymbol{r} := \rho_2 \cdot \boldsymbol{r} + (1-\rho_2) \cdot \boldsymbol{g}_2 \odot \boldsymbol{g}_2$
    Update weights $\theta := \theta - \frac{\eta\boldsymbol{s}/(1+\rho_1^t)}{\delta+\sqrt{\boldsymbol{r}/(1+\rho_2^t)}}$
  end for
end for
return learned weights $\theta$
```

3.3 Training Details of the Facial Processing Model

Let us summarize the whole training procedure of the FER models in this article. Several lightweight architectures were trained, namely, MobileNet v1, EfficientNet-B0 and EfficientNet-B2 [25]. They were trained in two stages with (1) pre-training on face recognition; and (2) fine-tuning on emotion classification. The details of the training procedure for the first stage were described in the previous articles [28], [50]. It is important to emphasize that the CNNs were trained on the cropped faces without any margins, so that most parts of the background, hairs, etc. is not presented. As a result, the face recognition accuracy may slightly degrade, but the learned facial features are more suitable for emotional analytic.

The second training stage was implemented as follows. A highly imbalanced training set of 287,651 images was utilized from the AffectNet dataset [23] annotated with $C = 8$ basic expressions (Anger, Contempt, Disgust, Fear, Happiness, Neutral, Sadness, and Surprise). The official validation set of 4,000 images (500 per class) was used for testing purposes. The faces were cropped with the bounding boxes

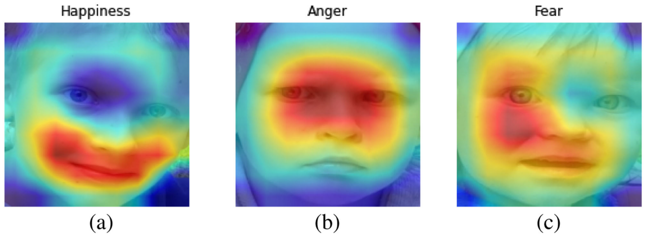


Fig. 3. Visualization of children emotions predicted by EfficientNet-B2.

provided by the authors of the AffectNet. The impact of additional pre-processing was examined, in which the cropped facial images were rotated to align them based on the position of the eyes similarly to [29] but without data augmentation.

The last layer of the network pre-trained on VGGFace2 was replaced by the new head (fully-connected layer with C outputs and softmax activation), so that the penultimate layer with D neurons can be considered as an extractor of facial features. The model was trained totally of 8 epochs by the Adam-based robust optimizer (Algorithm 1). In particular, the learning rate was set to 0.001 at the first three epochs and fit only the weights of the last layer of pre-trained CNN. Finally, the whole network was trained with a learning rate of 0.0001 at the last five epochs. The uncertainty parameter ϵ is equal to 0.05 in all experiments.

Though the AffectNet contains mainly photos of adults, the resulting FER models can classify emotions even for young children. For example, Fig. 3 contains the predicted emotional classes and easily interpreted GradCAM visualization of the CNN's decision.

It is known that many existing articles [29], [53] report the performance of their methods only for 7 basic emotions (without Contempt), for each there exist 283,901 training and 3500 validation images. Hence, two options to compare the proposed models with existing results for 7 emotional states (without Contempt) were studied. The first option is to train the CNN on the complete training set with 8 classes but to use only 7 predictions from the last Softmax layer for 3500 validation images so that the output that corresponds to the Contempt class is simply ignored. The second option is to re-train the model with 7 outputs on the reduced training set. This approach leads to better accuracy than the former one, though the usage of the universal 8-class model is desirable if the contempt emotion may be used in future.

The obtained CNNs were applied for the extraction of facial emotion features for video processing. In particular, the last Softmax layer was removed, and the outputs of D neurons at the penultimate layer were used as a D-dimensional feature vector in further experiments. The proposed complete multi-task model for recognizing the emotions and engagement of students in a video is shown in Fig. 4. It explores the known link between a facial expression and the level of comprehension which helps the teachers to improve their style accordingly and keeps the students interested and enthusiastic during the virtual lectures [5].

Here, two EfficientNets [25] are applied to extract facial identity and emotional features, respectively. The second CNN is obtained by fine-tuning the first CNN on a large FER dataset. Two statistical modules (STAT encoding) are used to aggregate the features of all faces in a video. The first unit combines the result of several statistical functions (minimum, maximum, mean and standard deviation) calculated for each feature of all video frames. The L_2-normalized descriptor is classified to predict the emotion of each student. The standard deviation of frame-wise emotional features of each face is used to predict the engagement level. Finally, the mean and standard deviation of emotional features of all faces at each frame is concatenated and aggregated in a single video descriptor, and the overall affect of the whole group is recognized.

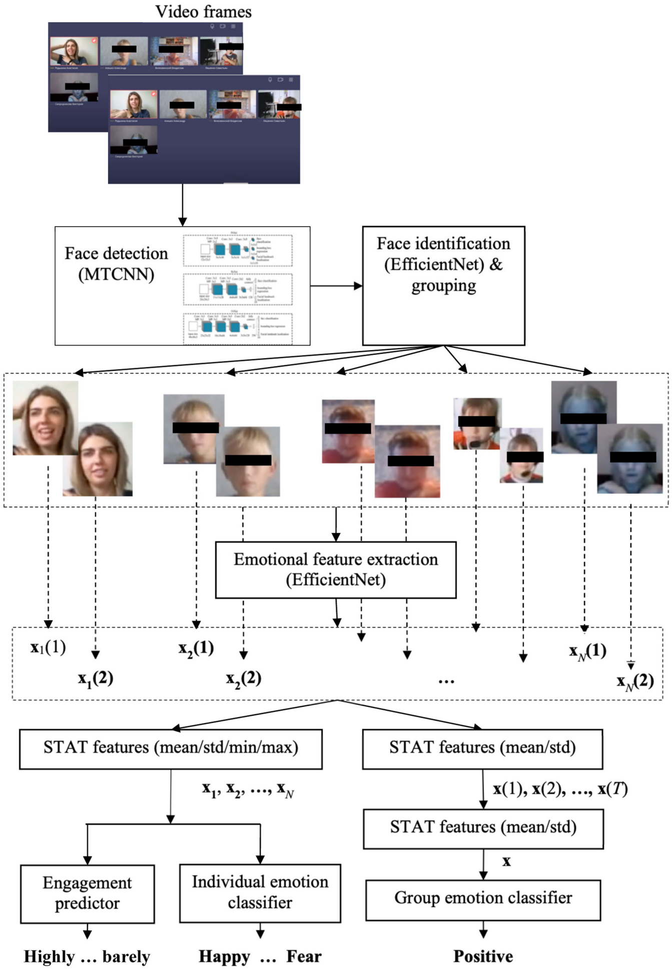


Fig. 4. The proposed model based on EfficientNet features.

4 Experimental Results

4.1 Emotional Feature Extraction

In this subsection, emotion recognition is studied for on static images for the AffectNet dataset. The results on the validation sets for the proposed FER models compared to the state-of-the-art results are shown in Table 1. Bold font indicates the best score for experiments with 8 emotions (Anger, Contempt, Disgust, Fear, Happiness, Neutral, Sadness and Surprise) and 7 basic emotional categories (the same without Contempt).

As one can notice, the best EfficientNet-B2 model improves the previously-known state-of-the-art accuracy [55] for complete validation set from 62.09% to 63.025%. It is slightly (0.1%) less accurate than EmotionGCN [29] for 7 classes, though the network architecture was not modified, and the

TABLE 1
Accuracy for the AffectNet Validation Set

	Accuracy, %	
Model	8 classes	7 classes
Baseline (AlexNet) [23]	58.0	-
Deep Attentive Center Loss [53]	-	65.20
Distilled student [54]	61.60	65.40
EfficientNet-B2 (SL + SSL in-panting-pl) [24]	61.32	-
EfficientNet-B0 (SL + SSL in-panting-pl) [24]	61.72	-
DAN [55]	62.09	65.69
TransFER [27]	-	66.23
EmotionGCN [29]	-	**66.46**
Our MobileNet-v1	60.20	64.71
Our EfficientNet-B0	61.32	65.74
Our EfficientNet-B2	**63.03**	66.34

TABLE 2
Class-Level Accuracy of Emotion Recognition on Statis Photos, AffectNet Dataset

Emotion	MobileNet-v1	EfficientNet-B0	EfficientNet-B2
Anger	**62.8**	61.4	54.2
Contempt	48.0	60.4	**66.0**
Disgust	51.8	50.0	**65.4**
Fear	**66.8**	66.2	63.8
Happiness	**81.8**	78.0	74.6
Neutral	**58.6**	53.4	54.6
Sadness	61.8	59.4	**65.4**
Surprise	56.0	**61.8**	60.2

TABLE 3
Ablation Study of the Proposed Models, AffectNet Dataset

		Accuracy, %	
Model	Pre-train set	8 classes	7 classes
MobileNet-v1	ImageNet	56.88	60.4
EfficientNet-B0	ImageNet	57.55	60.8
EfficientNet-B2	ImageNet	60.28	64.3
SENet-50	VGGFace2	58.70	62.31
Our MobileNet-v1, 8 classes	VGGFace2	60.25	63.21
Our MobileNet-v1, 7 classes	VGGFace2	-	64.71
Our EfficientNet-B0, 8 classes	VGGFace2	61.32	64.57
Our EfficientNet-B0, 7 classes	VGGFace2	-	65.74
Our EfficientNet-B2, 8 classes	VGGFace2	**63.03**	66.29
Our EfficientNet-B2, 7 classes	VGGFace2	-	**66.34**

TABLE 4
Ablation Study of Optimizers, AffectNet Dataset

	8-class accuracy, %		7-class accuracy, %	
CNN	Adam	Robust (3)-(5)	Adam	Robust (3)-(5)
MobileNet-v1	59.87	**60.25**	64.54	**64.71**
EfficientNet-B0	60.94	**61.32**	65.46	**65.74**
EfficientNet-B2	62.11	**63.03**	65.89	**66.34**

training process was straightforward. The accuracy of MobileNet and EfficientNet-B0 is lower, but still comparable with the best-known results reported for this dataset. It is important to emphasize that though the average accuracy of the deepest EfficientNet-B2 model is greater, the class accuracy for every type of emotions is sometimes lower when compared to EfficientNet-B0 and MobileNet (Table 2), so that all our models may be useful in different down-stream tasks.

The detailed ablation study of experiments for AffectNet is presented in Tables 3 and 4. In the latter table, the greatest 8-class and 7-class accuracy for each row (model) are marked by bold.

Here two datasets for pre-training were examined, namely, (1) conventional ImageNet; and (2) VGGFace2 [47] to learn the facial embeddings suitable for face recognition. The official models pre-trained on ImageNet were taken from Tensorflow 2 and PyTorch Image Models (timm) for the former approach. The latter technique was implemented as described in Section 3.3. As one can notice, such a pre-training leads to much better FER's accuracy, even though facial identity features should not depend on the emotional state [28]. Moreover, Table 4 demonstrates that the robust optimization (Algorithm 1) makes it possible to increase the accuracy. It is especially noticeable for the best EfficientNet-B2 model which established a new state-of-the-art result for complete validation set with 8 classes.

4.2 Engagement Prediction

In this subsection, the results on the EngageWild [8] are reported. This dataset contains 147 training and 48 validation videos with an average duration of 5 minutes. Each video is associated with one of 4 engagement levels 0, 0.33, 0.66 and 1 representing engagement mapped to disengaged, barely engaged, engaged and highly engaged.

The frame images were extracted from the video using the FFmpeg tool, and the facial regions were found in each frame using the MTCNN detector. If no faces were detected, the frame was ignored. Next, the developed emotional models were used to extract features of the largest facial region. The final descriptor of the whole video was computed as a standard deviation of frame-wise facial features similar to the baseline [20]. We tried to use other STAT features (mean, max, min) but did not observe the improvements in the MSE (mean squared error) measured on the validation set. The obtained video descriptor was fed into ridge regression from the MORD package because the initial engagement prediction task may be formulated as an ordinal regression. The results of the best attempts compared to the results of the participants of the EmotiW challenge on the official training and validation set are shown in Table 5.

It is important to emphasize that the best results are typically achieved by ensemble models that use several different audio and video features. Hence, the best results of single models are presented here to frankly compare the methods that use only one CNN. Nevertheless, the MSE 0.0563 for EfficientNet-B0 features is the best one when compared to any existing method. The confusion matrix of the best ordinal regression is shown in Fig. 5. Its MSE is lower than the best single model [13] up to 0.01 (15% relative improvement).

However, this point should be clarified. The participants of the Emotion Engagement in the Wild challenge verified that achieving better results on the validation set does not lead to excellent quality on the testing set. For example, the winner of the first challenge (EmotiW 2018) has high

TABLE 5
MSE for the EngageWild Validation Set

Model		MSE
Ensemble	Attention+multi-rate+hybrid [13]	0.0609
	4 models for 2 behavior features [45]	0.0572
	GAP+LBP-TOP [41]	0.0569
Single model	Baseline (OpenFace+LSTM) [20]	0.10
	LGCP [36]	0.0884
	Bootstrap (OpenPose+LSTM) [44]	0.0717
	GAP [41]	0.0671
	Dilated-TCN for action units [42]	0.0655
	VGG [13]	0.0653
	Our MobileNet, ridge regression	0.0722
	Our EfficientNet, ridge regression	**0.0563**

validation MSE of 0.0717 [43]. Thus, many researchers [44], [45] tuned the hyper-parameters of their models on different splits of videos from the united training and validation sets provided by the authors of EngageWild dataset [8]. Hence, in the ablation study two splits of the dataset were used: (1) the official split; and (2) the new random split with balanced training/validation sets, which have the same size as the sets from the original split. In addition to ridge ordinal regression, regression models from scikit-learn were tested, namely, random forest (RF), support vector regression (SVR) with linear and RBF (radial basis function) kernels. Moreover, several sequence models were implemented: (1) one GRU with 128 units and one fully connected output layer, and (2) single frame-level attention [56] with one dense output layer. These sequence models were applied to two inputs, namely, initial features of each frame extracted by developed emotional CNN, and their concatenation with a single descriptor (component-wise standard deviation of frame-wise features) used in the regression models. The results are presented in Table 6. Bold font indicates the lowest MSE for each split and the architecture of the CNN.

Here the new split causes more explainable results. For example, a simple ridge regression has 0.01-0.02 lower MSE than more widely used RF/SVR for the original split. However, our stratified split leads to approximately equal MSEs. Moreover, sequence models have significantly better results. For example, the lowest MSE for the EfficientNet-B0 features was obtained by using frame-wise attention. As a result, we decided to train the models for the demo application using the new split which seems to be more consistent with existing studies of regression techniques.

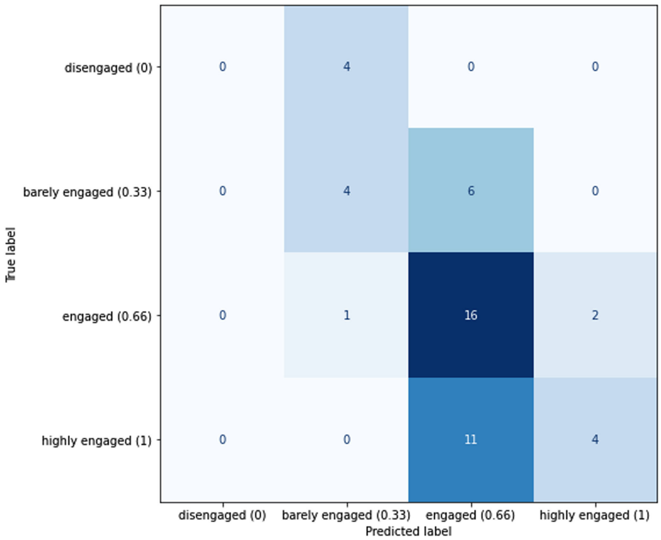


Fig. 5. Confusion matrix for engagement prediction, ridge ordinal regression, EfficientNet-B0 features.

TABLE 6
Ablation Study of the Proposed Models, EngageWild Dataset

		Validation MSE	
Our CNN	Classifier	Original split	Our split
MobileNet v1	RF	0.0844	**0.0511**
	LinearSVR	0.0895	0.0588
	SVR RBF	0.0759	0.0526
	Ridge regression	**0.0722**	0.0547
	GRU (frame only)	0.0981	0.0680
	GRU (frame+std)	0.0970	0.0585
	Attention (frame only)	0.0977	0.0618
	Attention (frame+std)	0.0882	0.0530
EfficientNet-B0	RF	0.0738	0.0540
	LinearSVR	0.0758	0.0560
	SVR RBF	0.0778	0.0543
	Ridge regression	**0.0563**	0.0593
	GRU (frame only)	0.0970	0.0668
	GRU (frame+std)	0.0761	0.0445
	Attention (frame only)	0.0882	0.0626
	Attention (frame+std)	0.0682	**0.0494**
EfficientNet-B2	RF	0.0882	0.0635
	LinearSVR	0.0897	0.0599
	SVR RBF	0.0868	**0.0592**
	Ridge regression	**0.0702**	0.0642
	GRU (frame only)	0.1065	0.0777
	GRU (frame+std)	0.0850	0.0672
	Attention (frame only)	0.0997	0.0715
	Attention (frame+std)	0.0914	0.0652

Remarkably, EfficientNet-B0 with less number of parameters has slightly better performance when compared to EfficientNet-B2, though the latter has much higher emotion recognition accuracy for static images (Table 3). It is important to mention that concatenation of the frame features with the standard deviation of features from the whole video works much better in most cases except the most simple MobileNet and GRU network.

4.3 Video-Based Emotion Recognition

The video-based emotion recognition is studied on two datasets from the EmotiW challenge [20]. First, the AFEW dataset with 773 train and 383 validation samples was examined. It contains audio-video short clips collected from movies and TV serials with spontaneous expressions, various poses, and illuminations. They are labelled using a semi-automatic approach. The task is to assign a single emotion label to the video clip from the six universal emotions (Anger, Disgust, Fear, Happiness, Sad, and Surprise) and Neutral.

The pre-processing of all video clips from the previous subsection was used. However, the point-wise mean, max, min and standard deviation of their frame descriptors were concatenated [57]. Thus, the dimensionality of the video descriptor is 4-times greater than the dimensionality D of the face emotional embeddings. If the face was not found in

TABLE 7
Accuracy for the AFEW Validation Set

Model		Accuracy, %
Ensemble, audio+ video	Bi-modality fusion of 4 CNNs [32]	54.3
	VGG13+VGG16+ResNet [18]	59.42
	5 FBP models [31]	**65.5**
Single model	LBP-TOP (baseline) [20]	38.90
	Frame attention network (FAN) [56]	51.18
	Noisy student with iterative training [30]	55.17
	Our MobileNet-v1	55.35
	Our EfficientNet-B0	**59.27**
	Our EfficientNet-B2	59.00

the training video, it was ignored, but the validation videos with missed faces were associated with zero descriptors [28]. The L_2-normalized descriptors were classified using LinearSVC from scikit-learn with regularization parameters found using cross-validation on the training set. The classification accuracy is presented in Table 7. The greatest accuracy for ensemble and single models are marked by bold.

Here the proposed models are up to 5% more accurate when compared to other single models. Even the MobileNet is 0.1% more accurate than the best-known technique with the ResNet that was iteratively trained as a noisy student [30]. The confusion matrix (Fig. 6) of the best EfficientNet model demonstrates that though many emotions are predicted accurately, the accuracy for at least Disgust and Fear categories should be improved. Though the best ensembles [31] are still much more accurate, our approach is much faster and can be implemented for real-time processing of a student's emotions even on his or her mobile device.

The video group emotion recognition task was studied by using the VGAF dataset [10] that contains group videos downloaded from YouTube with creative commons license. The data has a lot of variations in terms of context, number of people, video quality, etc. The training set provided by the challenge's organizers contains 2661 clips, while 766 videos are available for validation. The task is to classify each video into 3 classes - positive, neutral and negative.

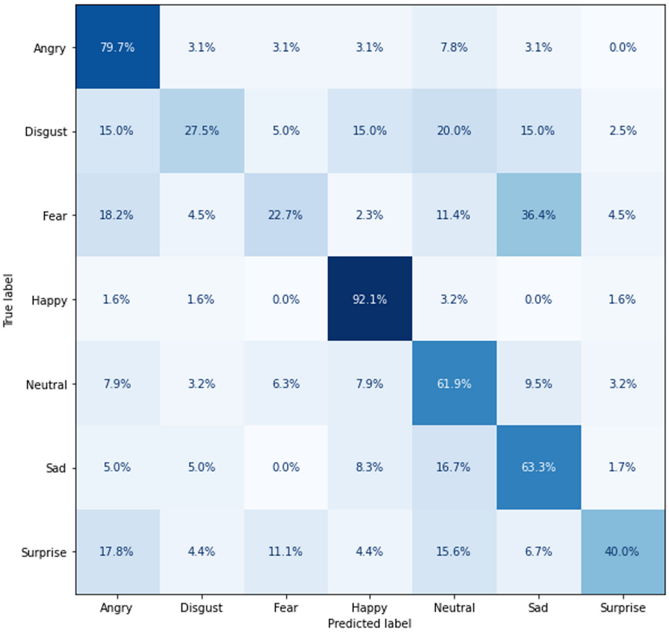


Fig. 6. Confusion matrix for video-based individual emotion recognition, EfficientNet-B0 features.

TABLE 8
Accuracy for the VGAF Validation Set

Model		Accuracy, %
Ensemble, audio+ video	VGAFNet (face + holistic + audio) [10]	61.61
	K-injection networks [34]	66.19
	Fusion of 14 models [58]	71.93
Single model	Hybrid Networks [14]	**74.28**
	VGAFNet (faces) [10]	60.18
	DenseNet-121 (Hybrid Networks) [14]	64.75
	Self-attention K-injection network [34]	65.01
	Slowfast [58]	68.57
	Our MobileNet-v1	68.92
	Our EfficientNet-B0	66.80
	Our EfficientNet-B2	**70.23**

As the frame in this dataset usually contains several facial regions, the mean and standard deviation of the facial emotion embeddings in each frame were concatenated to obtain its features. The final descriptor of the whole video was computed as a concatenation of the mean and standard deviation of frame features. The missing faces in all videos were processed similarly to the previous experiments with the AFEW dataset: the empty videos were removed from the training set, but associated with zero descriptors to compare the validation accuracy with existing models [28]. The video descriptors obtained by the proposed CNNs are classified by SVC with RBF kernel. The main results are presented in Table 8.

The proposed approach leads again to the best results against single models. The best EfficientNet-B2 improves the best-known Slowfast video analysis [58] up to 1.7%. Its confusion matrix is presented in Fig. 7. One can notice the natural behaviour of this classifier, namely, the worst accuracy for the Neutral category, which is typically misclassified as either positive or negative affect. Even MobileNet is 0.35% more accurate than the usage of Slowfast. Though the large ensembles of the winners of this challenge [14] are more accurate, the proposed models are still competitive with even the fusion of 14 deep CNNs applied to various audio and video features [58].

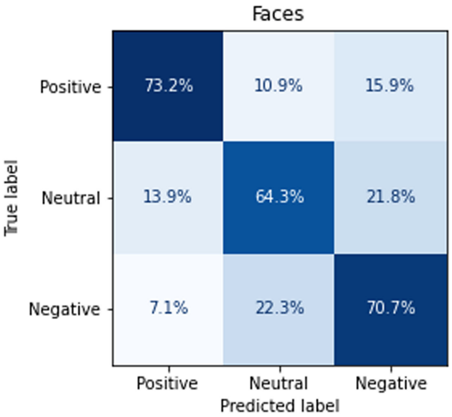


Fig. 7. Confusion matrix for video-based group-level emotion recognition, EfficientNet-B2 features.

5 Conclusion

Though there exist many accurate engagement detection and emotion classification techniques [31], [41], their usage in online learning is still very limited because most students do not want to share their facial videos due to privacy concerns. It is especially difficult to get permission for facial analysis of children. In order to deal with this issue, the novel framework (Fig. 1) was proposed in this article for video-based analysis of the learners' emotional engagement. This framework can be integrated into existing e-learning tools for fast and accurate assessment of students' emotions and comprehension. We believe, the main users of this framework will be the managers and technical specialists of online platforms or online specialities/degrees who have to analyze the success of their courses and find the key factors to improve the quality of online courses, reduce the customer churn, etc. The students need only to turn on their cameras and teachers have to record the video of their webinars. To save the learner's privacy, the facial videos may be processed even on his or her mobile device. The teacher receives only the averaged predicted engagement and emotional scores of the whole group. The developed neural models require facial images to have resolution 224x224 or 300x300, so that the quality of any frontal camera is appropriate. Moreover, our preliminary experiments demonstrate that it is possible to extract emotional features from 1 frame per second without significant accuracy degradation. As a result, even very cheap smartphones or laptops may be used by a learner with any economical background.

The key component of the proposed pipeline is the lightweight neural models (Fig. 4) learned using robust modification of the Adam optimizer (Algorithm 1). The best model outperforms the known state-of-the-art results (Table 6) for prediction of 4 engagement levels (disengaged, barely engaged, engaged and highly engaged) on the EngageWild dataset that contains facial videos of the student's watching educational videos such as the ones in MOOCs [8]. Unfortunately, as far as we know, there are no publicly available datasets for student's emotion recognition. Hence, some of the experiments were done on the datasets that do not belong to e-learning. It was shown that the facial representations obtained by the developed models can be used for fast simultaneous recognition of individual- and group-level emotions. Our best CNN outperforms the known state-of-the-art results of single models for recognition of 8 emotions on static photos from the AffectNet (Table 1), recognition of 7 basic emotions in the AFEW dataset (Table 7) and classification of 3 affects (positive, neutral and negative) in the VGAF dataset (Table 8). However, we claim that the proposed emotional features can be used for accurate emotion recognition on other datasets including e-learning domain. For example, the same models let the first author of this article take the third place in the multi-task learning competition on Affective Behavior Analysis in-the-wild (ABAW) [48].

It is necessary to mention that the proposed approach is less accurate when compared to the best-known multimodal ensembles on the AFEW and VGAF datasets. Nevertheless, as the group videos in e-learning systems typically do not contain the voice of the students and even pose is unclear, many parts of these ensembles are useless if it is required to estimate the overall emotion of the whole group. In this case, analysis of facial regions is the most preferable for emotion recognition in online lessons because it leads to a much higher emotion classification accuracy when compared to the performance of existing single models.

In future, it is necessary to use extra video data to improve the quality of engagement prediction engine, which is limited now due to the usage of a small training set. Second, as the resolution of detected faces can still be low, we are going to study the known low-resolution FER techniques [49] to improve the quality of engagement detection and emotion recognition by modifying the proposed models that were trained on 224x224 and 300x300 facial images. Third, it may be important to predict arousal and valence [51], [59] in addition to facial expressions to give the teacher additional information about the attitude of each student. Another research direction is the improvement of face clustering by applying text detection and recognition techniques to obtain the name of each participant of the video conference and group facial regions that are located close to the detected name. Finally, it is necessary to examine the potential of the proposed technique and emotional features in other tasks of e-learning, such as online proctoring.

References

[1] X. Wang, T. Liu, J. Wang, and J. Tian, "Understanding learner continuance intention: A comparison of live video learning, pre-recorded video learning and hybrid video learning in COVID-19 pandemic," *Int. J. Hum.–Comput. Interact.*, vol. 38, no. 3, pp. 263–281, 2022.

[2] J. Shen, H. Yang, J. Li, and Z. Cheng, "Assessing learning engagement based on facial expression recognition in MOOC's scenario," *Multimedia Syst.*, vol. 28, pp. 469–478, 2022.

[3] Ł. Tomczyk, K. Potyrała, N. Demeshkant, and K. Czerwiec, "University teachers and crisis e-learning: Results of a polish pilot study on: Attitudes towards e-learning, experiences with e-learning and anticipation of using e-learning solutions after the pandemic," in *Proc. IEEE 16th Iberian Conf. Inf. Syst. Technol.*, 2021, pp. 1–6.

[4] P. Bhardwaj, P. Gupta, H. Panwar, M. K. Siddiqui, R. Morales-Menendez, and A. Bhaik, "Application of deep learning on student engagement in e-learning environments," *Comput. Elect. Eng.*, vol. 93, 2021, Art. no. 107277.

[5] M. Sathik and S. G. Jonathan, "Effect of facial expressions on student's comprehension recognition in virtual educational environments," *SpringerPlus*, vol. 2, no. 1, pp. 1–9, 2013.

[6] M. A. A. Dewan, M. Murshed, and F. Lin, "Engagement detection in online learning: A review," *Smart Learn. Environ.*, vol. 6, no. 1, pp. 1–20, 2019.

[7] J. Bacca-Acosta and C. Avila-Garzon, "Student engagement with mobile-based assessment systems: A survival analysis," *J. Comput. Assist. Learn.*, vol. 37, no. 1, pp. 158–171, 2021.

[8] A. Kaur, A. Mustafa, L. Mehta, and A. Dhall, "Prediction and localization of student engagement in the wild," in *Proc. IEEE Digit. Image Comput.: Techn. Appl.*, 2018, pp. 1–8.

[9] M. Imani and G. A. Montazer, "A survey of emotion recognition methods with emphasis on E-learning environments," *J. Netw. Comput. Appl.*, vol. 147, 2019, Art. no. 102423.

[10] G. Sharma, A. Dhall, and J. Cai, "Audio-visual automatic group affect analysis," *IEEE Trans. Affective Comput.*, to be published, doi: 10.1109/TAFFC.2021.3104170.

[11] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. Cambridge, MA, USA: MIT Press, 2016.

[12] T. Liu, J. Wang, B. Yang, and X. Wang, "NGDNet: Nonuniform Gaussian-label distribution learning for infrared head pose estimation and on-task behavior understanding in the classroom," *Neurocomputing*, vol. 436, pp. 210–220, 2021.

[13] B. Zhu, X. Lan, X. Guo, K. E. Barner, and C. Boncelet, "Multi-rate attention based GRU model for engagement prediction," in *Proc. Int. Conf. Multimodal Interact.*, 2020, pp. 841–848.

[14] C. Liu, W. Jiang, M. Wang, and T. Tang, "Group level audio-video emotion recognition using hybrid networks," in *Proc. Int. Conf. Multimodal Interact.*, 2020, pp. 807–812.

[15] T. Liu, J. Wang, B. Yang, and X. Wang, "Facial expression recognition method with multi-label distribution learning for non-verbal behavior understanding in the classroom," *Infrared Phys. Technol.*, vol. 112, 2021, Art. no. 103594.

[16] A. V. Savchenko, K. V. Demochkin, and I. S. Grechikhin, "Preference prediction based on a photo gallery analysis with scene recognition and object detection," *Pattern Recognit.*, vol. 121, 2022, Art. no. 108248.

[17] P. Xanthopoulos, P. M. Pardalos, and T. B. Trafalis, *Robust Data Mining*. Berlin, Germany: Springer, 2012.

[18] S. A. Bargal, E. Barsoum, C. C. Ferrer, and C. Zhang, "Emotion recognition in the wild from videos using images," in *Proc. Int. Conf. Multimodal Interact.*, 2016, pp. 433–436.

[19] H. Zeng et al., "EmotionCues: Emotion-oriented visual summarization of classroom videos," *IEEE Trans. Vis. Comput. Graphics*, vol. 27, no. 7, pp. 3168–3181, Jul. 2021.

[20] A. Dhall, "EmotiW 2019: Automatic emotion, engagement and cohesion prediction tasks," in *Proc. Int. Conf. Multimodal Interact.*, 2019, pp. 546–550.

[21] T. Ashwin and R. M. R. Guddeti, "Affective database for e-learning and classroom environments using indian students' faces, hand gestures and body postures," *Future Gener. Comput. Syst.*, vol. 108, pp. 334–348, 2020.

[22] B. E. Zakka and H. Vadapalli, "Estimating student learning affect using facial emotions," in *Proc. IEEE 2nd Int. Multidisciplinary Inf. Technol. Eng. Conf.*, 2020, pp. 1–6.

[23] 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.*, vol. 10, no. 1, pp. 18–31, Jan.–Mar. 2019.

[24] M. Pourmirzaei, G. A. Montazer, and F. Esmaili, "Using self-supervised auxiliary tasks to improve fine-grained facial representation," 2021, *arXiv:2105.06421*.

[25] M. Tan and Q. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in *Proc. Int. Conf. Mach. Learn.*, 2019, pp. 6105–6114.

[26] F. Ma, B. Sun, and S. Li, "Facial expression recognition with visual transformers and attentional selective fusion," *IEEE Trans. Affective Comput.*, to be published, doi: 10.1109/TAFFC.2021.3122146.

[27] F. Xue, Q. Wang, and G. Guo, "TransFER: Learning relation-aware facial expression representations with transformers," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 3601–3610.

[28] A. V. Savchenko, "Facial expression and attributes recognition based on multi-task learning of lightweight neural networks," in *Proc. IEEE 19th Int. Symp. Intell. Syst. Inform.*, 2021, pp. 119–124.

[29] P. Antoniadis, P. P. Filntisis, and P. Maragos, "Exploiting emotional dependencies with graph convolutional networks for facial expression recognition," in *Proc. 16th IEEE Int. Conf. Autom. Face Gesture Recognit.*, 2021, pp. 1–8.

[30] V. Kumar, S. Rao, and L. Yu, "Noisy student training using body language dataset improves facial expression recognition," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 756–773.

[31] H. Zhou et al., "Exploring emotion features and fusion strategies for audio-video emotion recognition," in *Proc. Int. Conf. Multimodal Interact.*, 2019, pp. 562–566.

[32] S. Li et al., "Bi-modality fusion for emotion recognition in the wild," in *Proc. Int. Conf. Multimodal Interact.*, 2019, pp. 589–594.

[33] J. R. Pinto et al., "Audiovisual classification of group emotion valence using activity recognition networks," in *Proc. IEEE 4th Int. Conf. Image Process. Appl. Syst.*, 2020, pp. 114–119.

[34] Y. Wang, J. Wu, P. Heracleous, S. Wada, R. Kimura, and S. Kurihara, "Implicit knowledge injectable cross attention audiovisual model for group emotion recognition," in *Proc. Int. Conf. Multimodal Interact.*, 2020, pp. 827–834.

[35] I. P. Ratnapala, R. G. Ragel, and S. Deegalla, "Students behavioural analysis in an online learning environment using data mining," in *Proc. IEEE 7th Int. Conf. Inf. Autom. Sustainability*, 2014, pp. 1–7.

[36] Z. Zhang, Z. Li, H. Liu, T. Cao, and S. Liu, "Data-driven online learning engagement detection via facial expression and mouse behavior recognition technology," *J. Educ. Comput. Res.*, vol. 58, no. 1, pp. 63–86, 2020.

[37] T. Dragon, I. Arroyo, B. P. Woolf, W. Burleson, R. E. Kaliouby, and H. Eydgahi, "Viewing student affect and learning through classroom observation and physical sensors," in *Proc. Int. Conf. Intell. Tutoring Syst.*, 2008, pp. 29–39.

[38] J. Whitehill, Z. Serpell, Y.-C. Lin, A. Foster, and J. R. Movellan, "The faces of engagement: Automatic recognition of student engagement from facial expressions," *IEEE Trans. Affective Comput.*, vol. 5, no. 1, pp. 86–98, Jan.–Mar. 2014.

[39] X. Chen, L. Niu, A. Veeraraghavan, and A. Sabharwal, "FaceEngage: Robust estimation of gameplay engagement from user-contributed (YouTube) videos," *IEEE Trans. Affective Comput.*, vol. 13, no. 2, pp. 651–665, Apr.–Jun. 2022.

[40] T. Baltrusaitis, A. Zadeh, Y. C. Lim, and L.-P. Morency, "OpenFace 2.0: Facial behavior analysis toolkit," in *Proc. IEEE 13th Int. Conf. Autom. Face Gesture Recognit.*, 2018, pp. 59–66.

[41] X. Niu et al., "Automatic engagement prediction with GAP feature," in *Proc. Int. Conf. Multimodal Interact.*, 2018, pp. 599–603.

[42] C. Thomas, N. Nair, and D. B. Jayagopi, "Predicting engagement intensity in the wild using temporal convolutional network," in *Proc. Int. Conf. Multimodal Interact.*, 2018, pp. 604–610.

[43] J. Yang, K. Wang, X. Peng, and Y. Qiao, "Deep recurrent multi-instance learning with spatio-temporal features for engagement intensity prediction," in *Proc. Int. Conf. Multimodal Interact.*, 2018, pp. 594–598.

[44] K. Wang, J. Yang, D. Guo, K. Zhang, X. Peng, and Y. Qiao, "Bootstrap model ensemble and rank loss for engagement intensity regression," in *Proc. Int. Conf. Multimodal Interact.*, 2019, pp. 551–556.

[45] J. Wu, Z. Zhou, Y. Wang, Y. Li, X. Xu, and Y. Uchida, "Multi-feature and multi-instance learning with anti-overfitting strategy for engagement intensity prediction," in *Proc. Int. Conf. Multimodal Interact.*, 2019, pp. 582–588.

[46] V. T. Huynh, S.-H. Kim, G.-S. Lee, and H.-J. Yang, "Engagement intensity prediction with facial behavior features," in *Proc. Int. Conf. Multimodal Interact.*, 2019, pp. 567–571.

[47] Q. Cao, L. Shen, W. Xie, O. M. Parkhi, and A. Zisserman, "VGGFace2: A dataset for recognising faces across pose and age," in *Proc. IEEE 13th Int. Conf. Autom. Face Gesture Recognit.*, 2018, pp. 67–74.

[48] A. V. Savchenko, "Video-based frame-level facial analysis of affective behavior on mobile devices using EfficientNets," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops*, 2022, pp. 2359–2366.

[49] Y. Yan, Z. Zhang, S. Chen, and H. Wang, "Low-resolution facial expression recognition: A filter learning perspective," *Signal Process.*, vol. 169, 2020, Art. no. 107370.

[50] A. V. Savchenko, "Efficient facial representations for age, gender and identity recognition in organizing photo albums using multi-output ConvNet," *PeerJ Comput. Sci.*, vol. 5, 2019, Art. no. e197.

[51] J. A. Russell, L. M. Ward, and G. Pratt, "Affective quality attributed to environments: A factor analytic study," *Environ. Behav.*, vol. 13, no. 3, pp. 259–288, 1981.

[52] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*.

[53] A. H. Farzaneh and X. Qi, "Facial expression recognition in the wild via deep attentive center loss," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis.*, 2021, pp. 2402–2411.

[54] L. Schoneveld, A. Othmani, and H. Abdelkawy, "Leveraging recent advances in deep learning for audio-visual emotion recognition," *Pattern Recognit. Lett.*, vol. 146, pp. 1–7, 2021.

[55] Z. Wen, W. Lin, T. Wang, and G. Xu, "Distract your attention: Multi-head cross attention network for facial expression recognition," 2021, *arXiv:2109.07270*.

[56] D. Meng, X. Peng, K. Wang, and Y. Qiao, "Frame attention networks for facial expression recognition in videos," in *Proc. IEEE Int. Conf. Image Process.*, 2019, pp. 3866–3870.

[57] P. Demochkina and A. V. Savchenko, "MobileEmotiFace: Efficient facial image representations in video-based emotion recognition on mobile devices," in *Proc. Pattern Recognit. Int. Workshops Challenge*, 2021, pp. 266–274.

[58] M. Sun et al., "Multi-modal fusion using spatio-temporal and static features for group emotion recognition," in *Proc. Int. Conf. Multimodal Interact.*, 2020, pp. 835–840.

[59] V. Skaramagkas et al., "A machine learning approach to predict emotional arousal and valence from gaze extracted features," in *Proc. IEEE 21st Int. Conf. Bioinf. Bioeng.*, 2021, pp. 1–5.

Andrey V. Savchenko received the BS degree in applied mathematics and informatics from Nizhny Novgorod State Technical University, Nizhny Novgorod, Russia, in 2006, the PhD degree in mathematical modelling and computer science from the State University Higher School of Economics, Moscow, Russia, in 2010, and the DrSc 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, 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 PhD degree in system analysis and information processing from Voronezh State Technical University, in 2017, and the specialist degree in applied mathematics and informatics from Nizhny Novgorod State Technical University, Nizhny Novgorod, Russia, in 2008. 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. She is also a senior research fellow with the Laboratory of Algorithms and Technologies for Network Analysis, HSE University. Her current research interests include speech processing and e-learning systems.

Ilya Makarov received the PhD degree in computer science from the University of Ljubljana, Ljubljana, Slovenia. Since 2011 up to 2022, he was a full-time lecturer with HSE University, School of Data Analysis and Artificial Intelligence. He is senior research fellow with AIRI, HSE University – Nizhniy Novgorod, and researcher with Samsung-PDMI Joint AI Center, St. Petersburg Department of Steklov Institute of Mathematics, Russian Academy of Sciences, St. Petersburg, Russia. His educational career in data science covers positions of program director of BigData Academy MADE from VK, senior lecturer with the Moscow Institute of Physics and Technology, and machine learning engineer and head of Data Science Tech Master program in NLP, National University of Science and Technology MISIS.