Neurocomputing 514 (2022) 435–450

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

In search of a robust facial expressions recognition model: A large-scale visual cross-corpus study

Elena Ryumina [a,*,1], Denis Dresvyanskiy [b,c,1], Alexey Karpov [a]

[a] St. Petersburg Institute for Informatics and Automation of the Russian Academy of Sciences, St. Petersburg Federal Research Center of the Russian Academy of Sciences (SPC RAS), St. Petersburg 199178, Russia
[b] Ulm University, Ulm 89081, Germany
[c] ITMO University, St. Petersburg 191002, Russia

ARTICLE INFO

Article history:
Received 3 November 2021
Revised 29 August 2022
Accepted 1 October 2022
Available online 7 October 2022
Communicated by Zidong Wang

Keywords:
Visual emotion recognition
Affective computing
Paralinguistic analysis
Cross-corpus analysis
Deep learning
End-to-end model

ABSTRACT

Many researchers have been seeking robust emotion recognition system for already last two decades. It would advance computer systems to a new level of interaction, providing much more natural feedback during human–computer interaction due to analysis of user affect state. However, one of the key problems in this domain is a lack of generalization ability: we observe dramatic degradation of model performance when it was trained on one corpus and evaluated on another one. Although some studies were done in this direction, visual modality still remains under-investigated. Therefore, we introduce the visual cross-corpus study conducted with the utilization of eight corpora, which differ in recording conditions, participants' appearance characteristics, and complexity of data processing. We propose a visual-based end-to-end emotion recognition framework, which consists of the robust pre-trained backbone model and temporal sub-system in order to model temporal dependencies across many video frames. In addition, a detailed analysis of mistakes and advantages of the backbone model is provided, demonstrating its high ability of generalization. Our results show that the backbone model has achieved the accuracy of 66.4% on the AffectNet dataset, outperforming all the state-of-the-art results. Moreover, the CNN-LSTM model has demonstrated a decent efficacy on dynamic visual datasets during cross-corpus experiments, achieving comparable with state-of-the-art results. In addition, we provide backbone and CNN-LSTM models for future researchers: they can be accessed via GitHub.

1. Introduction

In the recent decades, affective computing has become a new fast growing and perspective domain due to its high importance in human–computer interaction (HCI) systems. Requirements to current HCI systems have increased and now consist of not only recognition of user speech or face, but also analysis of user state, including emotional component. Such information serves then for adjusting system response to fit it in the best way for user. Possible applications are almost limitless: it is being utilized in robotics [1,2], marketing [3], entertainment [4], medicine [5,6], education [7] and many others. Reliable affect recognition in those areas is becoming one of the key features capable to advance the quality of the HCI systems on a new level.

Currently there are two most common models for emotion attribution – the categorical model [8] (also well-known as Ekman's model) and the time-continuous model [9] (also well-known as Russell's circumplex model). The circumplex model has 3 dimensions, which are arousal, valence, and dominance, although the dominance axis is often omitted. Ekman's model divides the emotional space into 7 categorical states (6 salient emotions + *Neutral*) and it is exploited by researchers much more often, becoming a fundamental approach in emotion description. Because of the simplicity of the annotation, the categorical datasets prevail over time-continuous ones, leading to more data available for training machine learning (ML), including deep learning (DL) models. Therefore, in this work, we are focused on using categorical data, since the DL models are well-known for their hunger for tremendous amount of data.

Today, DL models dominate over typical ML models in the affect computing domain. There are several reasons for that: (1) DL models do not need sophisticated feature engineering methods, because they are able to consume data *as is*; (2) DL models are

* Corresponding author.
E-mail addresses: ryumina_ev@mail.ru (E. Ryumina), denis.dresvyanskiy@uni-ulm.de (D. Dresvyanskiy), karpov@iias.spb.su (A. Karpov).
[1] These authors contributed equally to this work.

https://doi.org/10.1016/j.neucom.2022.10.013
0925-2312/

almost infinitely scalable – and therefore can build up performance (accuracy) if needed; (3) DL models preserve the knowledge of the learned domain, opening the possibility to use transfer learning, and; (4) DL models can be constructed as end-to-end (E2E) models, omitting the necessity of breaking down the problem on several steps (feature engineering, training, and others).

However, even powerful DL models suffer from dataset biases [10]. Such problem was named as *cross-corpus problem*, and lies in model "sticking" to the specific corpus it was trained on. Applying such a model to other corpora, researchers observe dramatic decrease in the model performance due to different conditions, in which the data was acquired [10]. The current methodology in such case is the model fine-tuning on a new data, which seems to be a temporary solution, because it requires to do it every time new data comes up. Moreover, the necessity of fine-tuning basically represents the inability of the model to cover all possible variations of emotions humans express [11], which is crucial for the future HCI systems.

In our work we would like to introduce the visual-based emotion recognition (ER) end-to-end framework able to identify emotions with high performance on different datasets. Trying to train data-unbiased model, we took into account many data balancing and augmentation techniques to provide the model with as various data as possible. Moreover, we have conducted extensive experiments with many different datasets, which are diverse in the terms of lightning, rate of occlusions, noise, age, ethnicity, and head poses, analysing the strengths of the model and its efficiency.

To sum up, the main contributions of the article are:

- We propose a flexible pipeline of the facial expressions recognition (FER) system consisting of the backbone FER model and several temporal FER models. Every component of the system can be substituted with other similar models: for instance, instead of the backbone model, researchers can use other feature extractors utilized in the computer vision domain.
- We introduce an efficient feature extractor for the FER task (named backbone model) demonstrating the state-of-the-art performance on the AffectNet dataset. Moreover, we provide this model for further scientific usage and describe all fine-tuning steps done to obtain it.
- We present a large-scale visual cross-corpus study leveraging the leave-one-corpus-out experiment protocol. During the experiment, several different temporal models have been examined, while the best one was chosen based on the models' performance and generalization ability.
- We demonstrate the robustness and decent performance of the backbone model via analysis of its functioning on various complex frames from dynamic FER datasets.

The rest of the article is organized as follows: we analyze the current state of both visual emotion recognition and cross-corpus models in Section 2. Section 3 presents a developed end-to-end framework and observes utilized data. Next, Section 4 provides the setup of conducted experiments and results obtained with the proposed framework. In Section 5 we discuss the results and analyse the features of the developed framework. Lastly, Section 6 summarizes the performed work and considers the directions of future researches in cross-corpus affect computing.

2. Related Work

In this Section we firstly observe the state-of-the-art visual emotion recognition systems and then analyse the progress done in cross-corpus (data-biasing) problem elimination.

Earlier, in before-deep-learning era, affect computing researchers were forced to engineer and exploit neat hand-crafted features such as Facial Action Units (FAUs) [12], Facial Landmarks, Histogram of Oriented Gradients (HOG) feature maps [13] and many others. The emotion recognition systems were highly dependent on the quality of the extracted features, while the machine learning algorithm was being chosen according to generated features in an expert way based on the knowledge and experience of developers. Extracted and selected hand-crafted features were more important rather than machine learning techniques used. The situation was revolutionized in the last decade, approximately at the time, after the VGG [14] and ResNet [15] were introduced. Such models, fine-tuned on emotional datasets, were able to show comparable performance, while having the possibility to consume raw data (images) without the feature extraction phase.

Starting from 2015, a numerous research works using DL models have been published – H. W. Ng et al. [16] utilized the ImageNet model fine-tuned on FER [17] and EmotiW datasets. Using such a cascade fine-tuning process, they dramatically outperformed the challenge baseline on more than 15% of accuracy. To reduce the confusing factors and, thus, the amount of data needed for training, G. Levi and T. Hassner [18] have proposed a novel image transformation technique and applied it for the training of an ensemble of VGG [14] and GoogleNet [19] Convolutional Neural Networks (CNNs). Averaging the class scores of the ensemble members, the authors got more than 15% improvement over the baseline results. S. A. Bargal et al. in [20] have used VGG-based and ResNet pre-trained on the combination of emotional datasets CNNs for the feature extraction. Normalizing the features and extracting the statistics along with all frames in the video file, they trained a Support Vector Machine (SVM) to predict one emotion for the whole video file, resulting in 16% increase of the recognition rate in comparison with the baseline performance.

There were also numerous studies devoted to bringing context consideration into DL facial emotion recognition [21–23]. Besides faces, such networks take into account the context, in which the humans appear, providing complementary information for a score or feature fusion. Combining the facial and context-based features by simple NN or SVM, the authors showed a significant performance growth in considered tasks.

Over the last few years, the emotion recognition research community has concentrated on developing multi-modal emotion recognition systems, processing different information channels independently with their further aggregation. In most cases, for the feature extraction from every data channel, researchers utilize deep neural networks (DNNs), usually CNNs. The information aggregation is done, however, by different techniques: Deep Belief Networks [24,25], Attention Mechanism [26,27], Weighted Score Fusion [28], by SVM [29,30], etc.

We should note here that, although the multi-modal systems normally show better results in comparison with uni-modal systems, the most efficient component of it is the visual sub-system [31]. Therefore, in this work, we have focused on obtaining an efficient unbiased visual emotion recognition system so that other researchers could use it in further experiments with visual or multi-modal frameworks.

It is also important to underline that considered researches have been conducted without rigorous cross-corpus experiments, and therefore are applicable for concrete datasets they were experimented on.

Obtaining a system, which is not sensitive to the changes in data condition and distribution, it is a long-standing problem in almost every ML domain. In developing the affective computing frameworks it has become one of the key points, since humans express their emotions in an enormous number of ways, depending on ethnicity, culture, language, and even age and gender.

To the best of our knowledge, most of the studies in this direction are done for audio modality, because of its computational ease in processing in comparison with the video channel. However, we survey in this subsection several relevant acoustic cross-corpus works to show the general trends and features in the cross-corpus direction.

H. Kaya and A. Karpov [32] have proposed a cascaded normalization approach for minimization of the speaker- and corpus-related effects. Utilizing the Extreme Learning Machine as a classifier, they applied proposed normalization to openSMILE audio features on 5 corpora, resulting in obtaining robust features and significant improvement of the model performance in comparison with other baseline normalization techniques. In [33] the authors introduced the approach called Adversarial Discriminative Domain Generalization, which forces the deep encoding of the two (or more) datasets to be as close as possible, learning the model to generalize the representation of the emotion despite the differences in datasets. Conducting experiments on 3 datasets, they showed that such generalization increases model performance, while also decreases the variance of the model.

H. Kaya et al. in [34] investigated the application of Long Short-Term Memory (LSTM) network to the set of frame-level acoustic Low-Level Descriptors (LLDs) on 3 different corpora. Combining predictions of the LSTM-based model with predictions of the Weighted Least-Squares Kernel classifier via weighted score-level fusion, they outperformed the challenge baseline systems on the development set. However, such an approach showed worse results on the test set, likely because of the class distribution mismatch among datasets. Nevertheless, in comparison with other exploited models, the LSTM-based classifier demonstrated a decent performance improvement due to its ability to model temporal dependencies.

In [35] the authors analyzed the generalization ability of DL models (CNN, LSTM, and CNN-LSTM), training considered DL systems on 5 corpora combined altogether, and presenting the model evaluation on development out-of-domain dataset. Utilizing the t-Distributed Stochastic Neighbor Embedding (t-SNE) to visualize the learned data representations, the authors observed that increasing the variability of the data by combining corpora enhances the model performance on the development set. Moreover, the LSTM-only model was proven to be prone to overfitting, while CNN was not, making the CNN-based approach more attractive for cross-corpus training. H. Meng et al. in [36] have proposed an E2E architecture, which is consists of dilated CNN and bidirectional LSTM (B-LSTM) with an attention mechanism, taking advantage of both networks. Conducting experiments on two datasets, the authors stated that the proposed E2E DL framework has demonstrated a high generalization ability and robustness to changes in data distribution caused by switching to an unseen corpus.

Regarding the cross-corpus video-based studies, there are many fewer works, which conducted cross-corpus experiments, due to the complexity of the video processing. Moreover, some of the works presented below are not specifically directed to cross-corpus study, yet have done experiments on several datasets in cross-corpus style to show the model's ability of generalization.

In [37], the authors introduced E2E DNN Inception-based architecture, conducting comprehensive cross-corpus experiments on several datasets. Utilizing the proposed model, they outperformed state-of-the-art results, mostly presented by exploiting ML classifiers and hand-crafted features. W. Xie et al. in [38] proposed a feature sparseness-based regularization integrated into loss function, outperforming the model trained with L2 regularization. The study was done on 4 corpora in cross-corpus style, showing the superiority of the model in comparison with former state-of-the-art models. M. V. Zavarez et al. [39] fine-tuned the pre-trained on facial images VGGFace [40] model on several emotional datasets, following the leave-one-out cross-corpus experimental setup. They showed that pre-trained on related to emotion recognition domains CNN models substantially outperform the randomly initialized ones, including the cross-corpus experiments.

In [41], the authors have proposed a CNN ensembling method with the modified architecture of each CNN, exploiting so-called *maxout* layers. Carrying out the experiments on three corpora, they concluded that the developed ensemble of CNNs is able to surpass the default CNN model if the data amount is sufficient. Z. Meng et al. in [42] introduced a novel identity-aware CNN with a self-designed architecture. Training the proposed model with developed sophisticated identity-sensitive and expression-sensitive contrastive losses, they outperformed all baseline and most of the state-of-the-art models on 3 emotional datasets, following the cross-corpus protocol.

In [43] B. Hasani and M. H. Mahoor developed the facial emotion recognition framework consisted of 3D Inception-based ResNet and LSTM stacked on top of it. To emphasize the facial components instead of regions, the authors add to the input the Facial Landmarks as complementary information. Providing the results done within the cross-corpus setup, emotion recognition system shows outperforming efficiency in comparison with the state-of-the-art systems on 3 out of 4 datasets.

In our former work [44], we have conducted cross-corpus experiments with hand-crafted features (Facial Landmarks) to investigate its applicability instead of using the E2E DL approaches. Utilizing the ensemble of the classifiers, we showed that the classification accuracy highly depends on the sequence length and the diversity of the dataset expressed by number of different participants.

We should note that none of aforementioned corpora (except FER2013 and our former research [44]) are used in our study. Moreover, to make the final data as diverse as possible, we have selected datasets so that they have as less as possible intersections in terms of recording setup and conditions: the lightning, subjects' moving, obstacles, age, ethnicity, and culture.

3. Materials and Methods

In this section, we describe the methodology of our work and data used.

3.1. Experimental Data

Emotional datasets are a key element in building a reliable emotion recognition system. They can contain one to several modalities, most often visual, acoustic, linguistic, or their combinations. However, one of the main sources of information in HCI nowadays is the video channel, and therefore, in this research we focus on the Visual Emotional Datasets (VEDs), which can be divided into static or dynamic ones, depending on the type of presented images of facial expressions. Most VEDs are annotated in terms of 6 basic Ekman's emotions (*Happiness, Sadness, Surprise, Fear, Disgust, Angry*) [8] plus *Neutral* state, resulting in the 7-class task. There are also VEDs acquired with a continuous valence-arousal scheme, however, they are much less presented because of the annotation ambiguity and problems with raters' agreement. The training of the reliable E2E method requires a tremendous amount of data with high variety and therefore we have decided to focus on categorical VEDs, which are more presented nowadays.

Looking from a different point of view, VEDs can be separated depending on the recording conditions: laboratory (imitation of facial expressions) and "in the wild" (natural non-acted facial expressions). The choice of the VEDs has a significant impact on

the effectiveness of the emotion recognition systems, especially on DL models, which become more efficient with more data seen.

For the fairness of the study, we have selected 8 VEDs varied in recording conditions. They are publicly distributed and have been widely used for analysis, research, and experiments in the affect computation domain. An overview of the chosen data is presented in Table 1. Next, we describe each of the presented VEDs in detail.

To the best of our knowledge, RAMAS [45] is a sole VED with persons having Slavic appearance and Russian speech. In total, the dataset contains 564 videos annotated by 21 experts (at least 5 experts per video). The specificity of this corpus is that the raters could mark different time intervals for the presence of one to several emotions. Moreover, sometimes the intervals for different experts could overlap, causing ambiguity for overlapping regions in terms of annotated emotion. The participants were not limited in the movements of the head and arms, therefore there are frames, in which the face is completely covered by hands or is in a difficult position for visual emotion recognition. However, such frames are also labelled with one of the emotions, usually depending on the previous context. These factors make it difficult to work with RAMAS.

IEMOCAP [46] consists of 151 videos divided into chunks. Overall, 6 raters (at least 3 experts per video) were involved in the annotation process, assigning to the sections one to several emotions. The total number of sections is presented in Table 1. The dataset was acquired for recognition of 5 emotional states (*Happiness, Angry, Sadness, Frustration, Neutral*) and continuous emotional dimensions (*Valence, Arousal, and Dominance*). However, experts were free to assign other emotions (*Disgust, Fear, Surprise*) if they were presented in the sections. In addition to the videos, the authors provide the Motion Capture (MOCAP) data, which represents the information about the muscles movement of the face, head, and arms. Although the videos have a good resolution of 720×360 pixels, one-third of the frame is occupied by the participant's interlocutor (it is an interviewer, which is not annotated), resulting in only 480×360 final resolution for the rated participant. In addition, the participant's head in many frames is quite far from the camera, making it difficult to extract deep features from such a small amount of pixels.

Unlike the two former datasets, which have 10 young participants each, CREMA-D [47] provides the materials with wide diversity in terms of ethnicity and age (from 20 to 74). Each video file has one from six labeled emotional states (the emotion *Surprise* was not considered). However, in addition to the emotional label, the authors provide a confidence level of the rater for every video. Such a feature allows us to select video files with a high confidence level, discarding the noisy data.

The specificity of the RAVDESS [48] lies in containing the speech and melodic reproduction of emotions. The melodic reproduction of emotions makes the dataset suitable for nonpharmacological treatment in the rehabilitation of neurological and motor disorders. In addition to 7 common emotional state classes, the authors have annotated the emotion *Calm* as well. All emotions were recorded with a normal and strong emotional intensity regime.

The uniqueness of the SAVEE [49] dataset is that during the video recording, the authors were showing to the participants the facial expressions and text prompts on the display. The main goal of this was to convey to the participants emotion presented on screen in the most accurate way. In addition, to capture the key features of every facial expression, 60 blue markers were painted on the participants' forehead, eyebrows, cheeks, lips, and jaws.

During the work with all aforementioned datasets, we have noticed that participants of the CREMA-D, RAVDESS, and SAVEE datasets were located at equal distance to the camera, that simplified their preprocessing in comparison with RAMAS and IEMOCAP.

AffWild2 [52,53,50,54–57] was introduced within the Affective Behavior Analysis in-the-Wild (ABAW) competition. The dataset participants have different ages (from babies to elder people) and ethnicity. Since the corpus was collected "in the wild", the video frames of faces have a wide range of head poses, lighting conditions, occlusions, and a variety of emotional expressions. It was annotated for 6 basic Ekman's emotions plus *Neutral* state, time-continuous valence-arousal dimensions, and Facial Action Units (FAUs) in a frame-by-frame way. Since the dataset was used for competition, it is deliberately divided into train, development, and test sets by authors, the test set is hidden for final model efficacy evaluation. Moreover, starting from 2020, three ABAW competitions have already been introduced, resulting in different versions of the AffWild2 dataset. We should clarify that we have used the original version presented in the ABAW-2020 competition. The complexity of processing AffWild2 lies in the videos themselves: some videos contain more than one person on the frame at the same time and therefore the faces may overlap, or, even worse, the main face may be completely lost from the frame, while the annotated label remains equal to the previous one. Such a problem challenges the model to consider the context, paying attention to previous frames.

The example frames of the participants from all observed VEDs are presented in Fig. 1.

FER2013 [17] is the only gray-scale corpus selected in this work. However, despite the gray-scale and low resolution, the corpus is still actively exploited to date for training the emotion recognition systems. FER2013 is divided into train, development, and test sets and publicly available.

AffectNet [51] is the largest corpus of the image-based categorical emotions, acquired in-the-wild conditions. The authors have collected the data using three search engines (Google, Bing, and Yahoo). AffectNet contains more than 1 million facial images with extracted facial landmarks, 450,000 of which have been also annotated in terms of 8 categorical emotions (*Contempt* is taken additionally to the 6 basic Ekman's emotions and *Neutral* state). Moreover, every facial image is also annotated in terms of continuous valence-arousal dimensions. As in the FER2013 corpus, facial

Table 1
Overview of the research VEDs. St. means states, Part. – participants, Con. – the recording condition, Lab. – the laboratory recording conditions, Wild – the "in-the-wild" recording conditions, n/a – not available.

VED	# St.	# Part.	# Samples/hours	FPS	Resolution	Con.
RAMAS[45]	7	10	10848/≈36:35	25–50	1920×1080	Lab.
IEMOCAP[46]	8	10	10038/≈12:26	30	360×480	Lab.
CREMA-D[47]	6	91	7441/≈5:17	30	480×360	Lab.
RAVDESS[48]	8	24	4904/≈5:37	30	1280×720	Lab.
SAVEE[49]	7	4	480/≈0:30	60	320×256	Lab.
AffWild2[50]	7	458	564/≈26:34	7.5–30	various	Wild
FER2013[17]	7	n/a	35888	–	48×48	Wild
AffectNet[51]	8	n/a	450000	–	various	Wild

Fig. 1. Sample frames from dynamic VEDs.

regions are localized. The corpus is divided into train, development, and test sets, however, the test set is not publicly available.

Thus, for this research we have chosen both types of the VEDs (with static and dynamic facial expressions), widely covering the high diversity of participants' gender, age, and ethnicity as well as variability in head poses, lighting conditions, occlusions, and the degree of the recording control.

3.2. Methodology

The purpose of the current study is to create a robust and efficient system for the categorical emotion recognition task. We have done it in two steps: (1) implementation of the backbone emotion recognition model, which is able to predict emotion from raw image with high performance and (2) fine-tuning of the backbone emotion recognition system, adding temporal axis to it (for instance, LSTM layers) and utilizing the cross-corpus training protocol.

3.2.1. Static and Temporal Emotion Recognition Approaches

In this subsection we describe the structure of every approach we have implemented in this research, precising the training pipeline of these models. We call the static facial emotion recognition model as the *backbone model*. The training process pipeline for model is shown in Fig. 2. Hereinafter, the vectors under the facial expression images represent the emotional state in the following order: *Neutral, Happiness, Sadness, Surprise, Fear, Disgust, Anger*.

Typically, the static facial images express the peak levels of emotion. We believe that utilizing even a reliable robust model trained on the static images will not be able to demonstrate a good generalization capability for real-world applications, since in-the-wild people often do not express emotions in such a clear way.

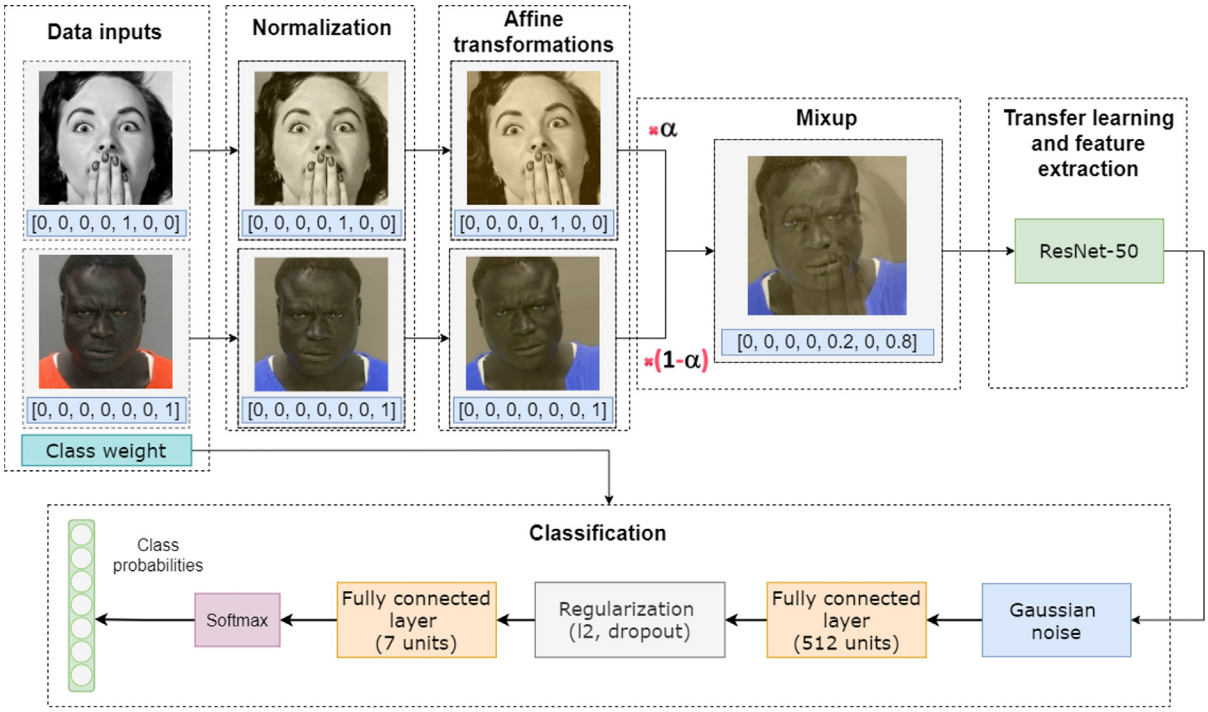

Fig. 2. The training process pipeline for the categorical emotion recognition model.

Moreover, when deciding on the state of the interlocutor, humans are well known to rely on the context information, taking into account previous emotional states as complementary information. Thus, we have decided to enhance the emotion recognition model with different temporal aggregation techniques. To learn the model catching temporal dependencies is a complex task, which requires a huge amount of data. Therefore, we have chosen 6 big different VEDs (the description is presented in Section 3.1), which represent the naturalistic way how people express emotions and which are good to learn capturing temporal context.

As a baseline, we have implemented two simple temporal approaches:

- **CNN-W**: The backbone model is used to get probability predictions for each frame in one video. Next, the emotion label of the video is predicted as a normalized sum of evaluated probability predictions.
- **CNN-S**: The backbone model is used to get probability predictions for each frame in one video. Next, the Hamming window is utilized to smooth the predictions in the window with chosen length (we have tried different lengths of the window). The formula for calculating weights in Hamming window is presented below:

$$weights_{Hamming} = 0.5 - 0.5cos(\frac{2n\pi}{l-1}), \quad (1)$$

where n is an integer vector with length l contained values from -l to l (even digits), and l is the window length. The Hamming window was applied to the entire video step-by-step starting from the first frame and with a bias equal to one frame. Then, all predictions were averaged, resulting in one probability vector.

The first baseline approach (CNN-W) is the simplest one because it takes into account only the "global" information, essentially catching the most frequent emotion that occurred in the considered video. Controversially, the CNN-S approach exploits the weighted aggregation of the frames within the fixed window, assigning the weights depending on the frame "proximity" to the central frame under consideration. This allows capturing the "local" temporal information. Both methods do not require any training and can be used directly after deep embeddings extraction done by the backbone model.

However, the CNN-W and CNN-S approaches are not flexible enough for capturing fairly complex temporal dependencies in data. Therefore, we have developed more sophisticated temporal approaches using extracted by backbone model deep embeddings:

- **CNN-SVM**: For every video, the deep embeddings are combined into sequences (windows) of 2 s. Next, we calculate *Means* and *Standard Deviations (STDs)* for every deep embedding within considered window, as we suggested in [58]. Evaluated statistics are then fed into SVM, which makes a final emotion prediction for the whole window. To make it more clear, we illustrate the scheme of the considered method in Fig. 3.
- **CNN-LSTM**: For every video, the deep embeddings are combined into sequences (windows) of 2 s. However, here we downsample every video to 5 frames per second (FPS) due to the absence of the necessity to calculate statistics. Next, an LSTM network with 512 and 256 consecutive neurons is trained on dynamic VEDs. To bound the LSTM for overfitting, after each LSTM layer we have added L2 regularization of 0.001 and dropout with rate of 0.2. The pipeline of the CNN-LSTM approach is presented in Fig. 4.
- **CNN-GRU**: To decrease the number of configurable parameters, we also tried to replace the LSTM layers with Gated Recurrent Unit (GRU) layers. All other hyperparameters are similar to the CNN-LSTM method. Due to the similarity of these two approaches, we did not present the CNN-GRU method in the pipeline depicted in Fig. 4.
- **CNN-LSTM-A**: We improved the CNN-LSTM approach by inserting between LSTM layers the Attention mechanism proposed in [59,60]. In addition, to speed up the convergence of the training, the batch normalization after every LSTM layer was inserted. The pipeline of the proposed method is presented in Fig. 4 as well.

To evaluate all proposed approaches, we utilized the leave-one-corpus-out (cross-corpus) cross-validation procedure.

3.2.2. Experimental Setup of Hyper-parameter Search for the Backbone Model

To create the efficient backbone categorical emotion recognition system, we have conducted numerous experiments, varying the following training hyper-parameters:

- **Schedulers of learning rates**: *Constant* (the learning rate is constant throughout the training process); *Time-based* (the learning rate changes at each epoch); *Piecewise Constant* (the learning rate is constant at a given iteration); *Cosine Annealing* [61].
- **Optimization algorithms** (Adam, SGD) and **Initial Learning Rates** (0.001, 0.0001, 0.00001).
- **DNN architectures**: ResNet-50 [62], SeNet-50 [63], VGG-16 [14] pre-trained on the VGG-Face2 dataset [40], EfficientNet-B0 [64], ResNet-101-V2 [62], MobileNet-V2 [65] pre-trained on Imagenet dataset [66].
- **Logarithmic Class Weighting** [58] and **Inversely Proportional Class Weighting**.

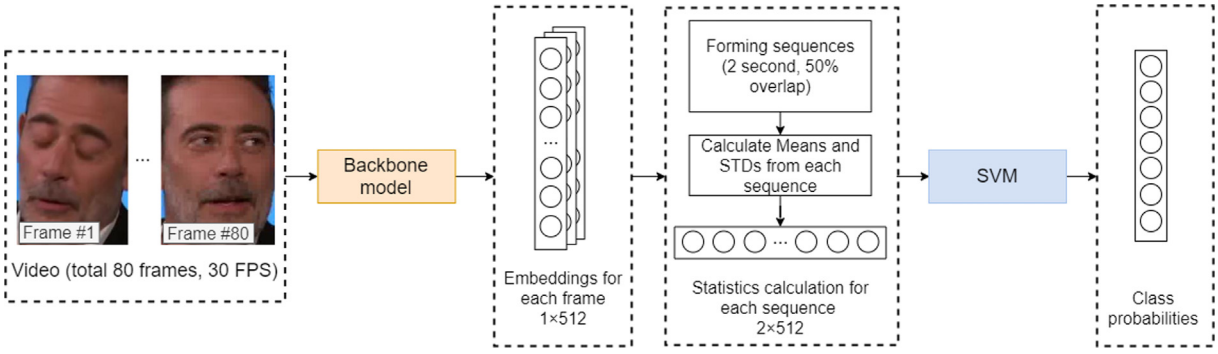

Fig. 3. The pipeline of the CNN-SVM approach.

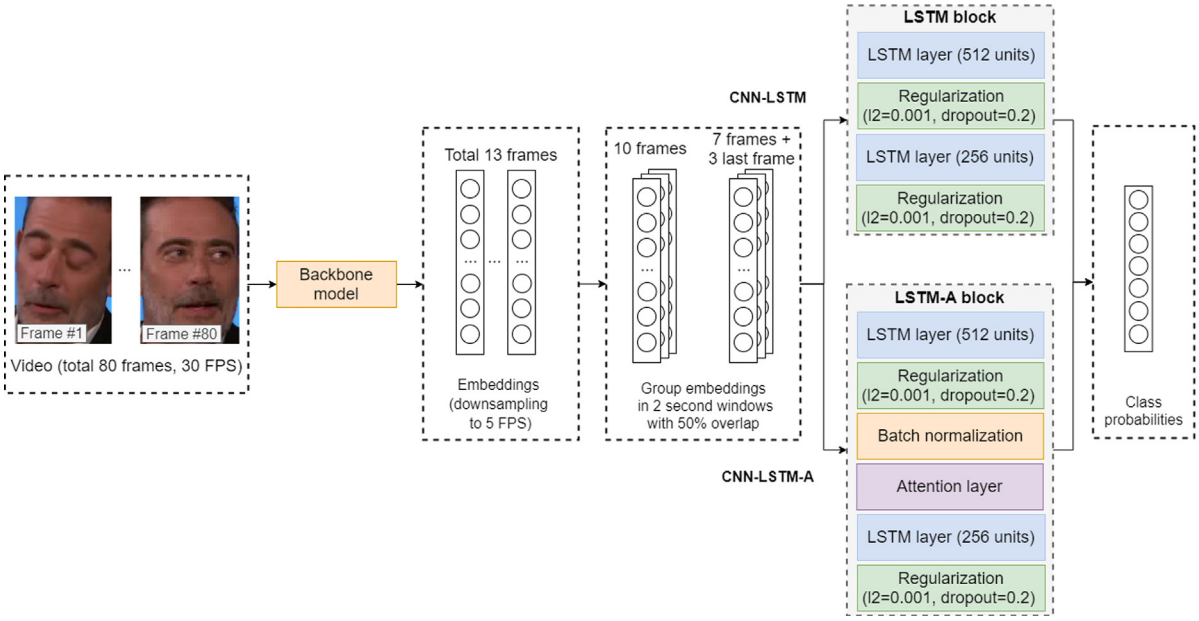

Fig. 4. The pipeline of the CNN-LSTM and CNN-LSTM-A approaches.

- **Regularization Methods**: Dropout, L2, Gaussian noise.
- **Convolutional Layers Freezing**: full freezing, without freezing, and freezing up to, but excluding the last convolutional layer.
- **Number of neurons of the last dense layer**: 256, 512, and 1048.
- **Data Augmentation Techniques**: Affine Transformations, Combining different VEDs, and Mixup[67].

Experimenting with all mentioned hyper-parameters, we have chosen the best ones by monitoring the model's performance on the development set during the training of categorical emotion recognition model. We should note that the backbone model is trained using static VEDs (AffectNet and FER2013).

4. Experimental Results

In this Section we present the data pre-processing description and experimental results, including backbone categorical model and the carried out cross-corpus analysis.

4.1. Data Pre-Processing

In contrast to static facial expressions in the AffectNet [51] and FER2013 [17], other VEDs contain dynamic video sequences without pre-detection of the faces. Therefore, the necessity to identify the face region on each video sequence has been raised. However, currently, there are many face detectors publicly available for the research, and each of them has pros and cons. To choose one for our research, we have evaluated three DL face detectors: the Single Shot Multibox Detector (SSD) [68], the RetinaFace [69], the Multi-task Cascaded CNN (MTCNN) [70]. The efficiency was scored according to the two metrics: FPS and the Intersection Over Union (IoU). The formula of the IoU is presented below:

$$IoU = \frac{TP}{TP + FP + FN}, \quad (2)$$

where TP is a number of true positive samples, FP is a number of false positives samples, and FN is a number of false negative samples.

The experiments were carried out on a randomly selected video from the AffWild2 dataset [52] with the name "99–30-720x720", which contains 1 800 frames with one person. AffWild2 is a good dataset to test the effectiveness of the face detectors, since it has many video frames with occlusions, dim or too bright lightning, and a high variety of head poses. We set the confidence threshold to be at least more than 70%, while the maximum size of the frame in width/high to be 300 pixels. The experimental results are presented in Table 2.

Table 2 shows that only the RetinaFace was able to find all 1800 faces, however, it has found 13 additional erroneous faces (FP). The analysis of the mistakes made by the RetinaFace showed that setting the confidence threshold to more than 90% eliminates all errors. However, when setting such a threshold for other face detectors, the IoU value drops significantly. Thus, since the RetinaFace showed the best IoU and a good FPS, we have selected it as a base face detector for our research. To see the more detailed research on the effectiveness of the face detectors, the reader is kindly referred to the paper [71].

In addition to the face detection, we have considered the raters' annotation confidence level, where it was possible (RAMAS, IEMOCAP, and CREMA-D). It has been experimentally proven [72] that with an increase of the annotation confidence levels, the accuracy of recognition systems grows. Therefore, currently in the emotion recognition domain, it is common to utilize the annotation confidence levels equaled to at least 60% [73]. This has been applied to the corpora mentioned above as well.

Moreover, to ensure the same processing conditions for recurrent neural networks in terms of temporality, we have equalized the FPS of every video from VED by downsampling all the video files to 5 FPS. We have utilized the simplest downsampling process – the selection of every n-th frame, while all other frames are

Table 2
The efficiency of the face detectors expressed via different metrics.

Face detector	TP	FP	FN	IoU, %	FPS
SSD	1787	41	13	99.0	≈**56**
MTCNN	1484	21	315	81.5	≈35
RetinaFace	1800	13	–	**99.3**	≈**56**

skipped. For instance, if we have a video with FPS equalled 10, to downsample it to the FPS equalled 5, we need to take every $n = 2$ (second) frame.

We should note here that there are some other, more sophisticated downsampling techniques (i. e. [74]), and we did not focus on that hyperparameter due to significant computational time of such techniques.

We present the overview of the datasets utilized in our research, including the class distribution, in Table 3. For VEDs consisted of static images (AffectNet and FER2013), the number of samples is shown in the overall number of frames. We did it for the Affwild2 as well, since it is annotated following the frame-by-frame protocol (meaning that every frame is rated separately).

Observing Table 3, one can note that most of the samples belong to the *Neutral* and *Happiness* categories, accounting for almost 76% of the overall data. It is well-known that disproportionately distributed number of samples in the training set negatively affects the performance of ML models [75]. Therefore, to eliminate this problem, we apply data augmentation techniques such as affine transformations, combining different VEDs, and Mixup. Furthermore, we have combined the AffectNet dataset with the data of several emotional categories from the FER2013, increasing the number of the minority class samples: *Sadness* at 16%, *Surprise* – 18.4%, *Fear* – 39.1%, *Disgust* – 12.2%, *Angry* – 13.8%.

4.2. Backbone Model

The main experiments on the training parameters selection for the backbone model are presented in Table 4. It is important to note that, besides all mentioned in Table 4 methods, the following training parameters for all experiments were used: random affine transformations and contrast varying; SGD optimization algorithm; two dense layers with 512 and 7 neurons are stack on top of CNN; dropout with rate equaled 0.2 after every dense layer; inversely proportional class weighting; batch size (BS) equals to 64; 30 training epoch. For experiments 1–6, the learning rate was set to constant value 0.0001. In addition, since we have exploited the pre-trained on VGGFace2 dataset model, all the images were normalized before feeding into the CNN in the same fashion, as in [40]. The normalization process consists of (1) conversion of the channel scheme from RGB to BGR; (2) centering of each channel according to the means calculated on the VGGFace2 dataset.

The best model was chosen by monitoring the model efficiency on the development set, in other words, by using *early stopping*.

To make the pre-processing procedure clearer,the batch generation process is presented in Fig. 5.

Analyzing the results from Table 4, we should mention that the Mixup data augmentation technique makes a significant contribution to improving the accuracy of the backbone model (adding it causes the accuracy increase on 1.8%, comparing experiments 5 and 9). ResNet50 showed a better performance in comparison with VGG-16 and MobileNet-V2 (see experiments 4, 5, 7, and 8). l2 regularization and adding the FER2013 data to the training slightly improved the model performance as well. Thus, utilizing all the proposed methods significantly magnified the model efficacy by 3.6% on an absolute scale (experiment 9). The confusion matrix for the experiment 9 on the AffectNet development set is presented in Fig. 6.

From the confusion matrix, one can see that the backbone model is well-balanced and mostly confuses adjacent emotions: for example, the *Fear* is correctly recognized by 60% cases, while most of the errors (16.6%) are related to the *Surprise*. These emotions are characterized by similar facial features such as wide-set eyebrows (with *Fear* they are usually lowered, while with *Surprise* they are raised). Another example lies in emotions *Disgust* and *Anger*: in both cases the eyebrows are usually lowered and pulled together, which explains why model sometimes predicts *Anger*

Table 3
Number of samples per each emotional category. NE denotes *Neutral* state, HA – *Happiness*, SA – *Sadness*, SU – *Surprise*, FE – *Fear*, DI – *Disgust* and AN – *Angry*, TS – train set, DS – development set.

VED	NE	HA	SA	SU	FE	DI	AN
RAMAS	172	496	187	341	208	214	249
IEMOCAP	1828	856	1183	155	56	8	1336
CREMA-D	907	1028	233	–	402	679	594
RAVDESS	376	752	752	384	752	384	752
SAVEE	120	60	60	60	60	60	60
AffWild2 (TS)	589215	152010	101295	39035	11155	12704	24080
AffWild2 (DS)	183636	53702	39486	23113	9754	5825	8002
AffectNet (TS)	74874	134415	25459	14090	6378	3803	24882
AffectNet (DS)	500	500	500	500	500	500	500
FER2013 (TS)	4965	7215	4830	3171	4097	436	3995
Total	856593	351034	183985	80849	33364	24613	64450
Part, %	53.71	22.01	11.54	5.07	2.09	1.54	4.04

Table 4
Experiments on the hyperparameters selection for the training process of backbone model. The figures from 1 to 9 denote the number of the experiment. GN means Gaussian noise, CA – Cosine Annealing.

Method	1	2	3	4	5	6	7	8	9
ResNet-50	+	+	+	+	+	+	–	–	+
VGG-16	–	–	–	–	–	–	+	–	–
MobileNet-V2	–	–	–	–	–	–	–	+	–
FER2013	–	+	+	+	+	+	+	+	+
l2-reg. (0.0001)	–	–	+	+	+	+	+	+	+
GN (0.1)	–	–	–	+	+	+	+	+	+
CA (5 cycles)	–	–	–	–	+	–	–	+	+
Time-based	–	–	–	–	–	+	–	–	–
Mixup	–	–	–	–	–	–	–	–	+
Accuracy, %	62.8	63.2	63.6	64.3	64.6	63.9	64.0	62.1	**66.4**
δ, %	–	0.4	0.8	1.5	1.8	1.1	1.2	−0.7	3.6

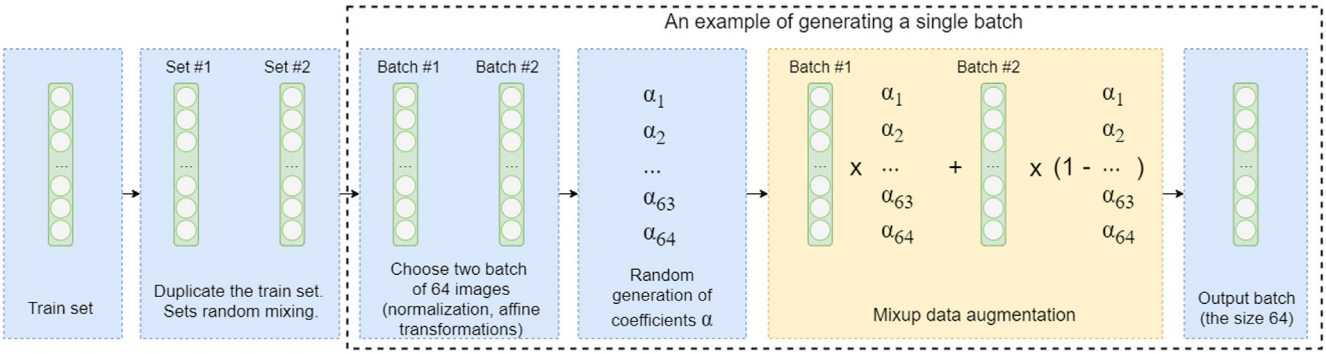

Fig. 5. The process of batch generation. α is a random value from 0 to 1.

instead of *Disgust* (in 14% cases). To read more about the manifestation of the particular emotions, the reader is kindly referred to the [8]. We would like to note that, in general, we obtained a good, well-balanced and efficient backbone emotion recognition system with a recognition rate for every emotion not less than 59%, which is quite a high result for such a subjective task as emotion recognition. Moreover, to the best of our knowledge, the developed backbone emotion recognition system obtained the highest state-of-the-art results on the AffectNet validation set. To demonstrate it, we have compared the proposed model with other known state-of-the-art results in Table 5.

4.3. Cross-Corpus Analysis

As we described before, for the cross-corpus analysis we have taken dynamic VEDs contained videos with different recording conditions. Since we have now an additional temporal dimension, the necessity to model it has raised. To accomplish it, we have chosen a window of 2 s for temporal modeling. The length of 2 s was selected because of the limitation of one VED called CREMA-D - the average length of the videos in this dataset is 2.5 s, making it difficult to set the window size bigger: the zero/same padding for short videos would only confuse the model.

For the CNN-S approach, we have experimented with different lengths of the smoothing window. The number of frames was identified in brute-force style from 3 to 99 for all odd digits. During experiments, we faced contradicting results: the optimal window size (in terms of model performance) varied starting from small

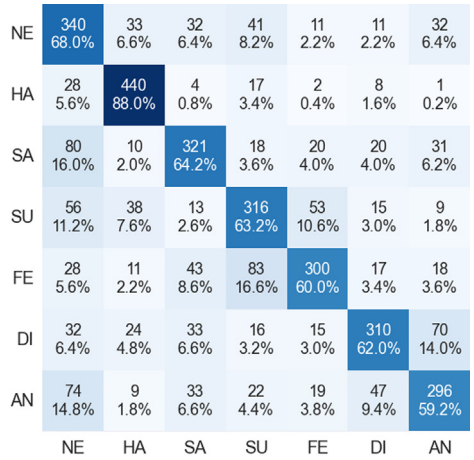

Fig. 6. Confusion matrix for the experiment 9 on AffectNet development set. NE denotes *Neutral* state, HA – *Happiness*, SA – *Sadness*, SU – *Surprise*, FE – *Fear*, DI – *Disgust* and AN – *Angry*.

to high values (61, 37, 73, 23, 97, and 97 for RAMAS, RAVDESS, CREMA-D, IEMOCAP, SAVEE, and AffWild2 respectively). This shows how important the length of the context is in dynamic affective computing. However, to generalize the system, we had to choose one window length for all datasets. It was done by averaging the Unweighted Average Recall (UAR) value for all datasets for every window length and choosing the highest one. Ultimately, the window size of 71 was selected, and therefore we present the experimental results (see Table 6) with this window size.

For all the experiments with the LSTM-based and GRU-based networks, the SGD with learning rate of 0.0001 was chosen. We have selected the following hyperparameters empirically: the number of neurons equals 512 and 256 for the first and the second recurrent layers, dropout rate 0.2, and regularization parameter 0.001 for recurrent layers. As we noted earlier, we have implemented two temporal aggregation approaches: straight recurrent network (CNN-LSTM and CNN-GRU) and attention-based LSTM network (CNN-LSTM-A), which is supposed to be more focused on significant deep embeddings extracted by trained CNN (backbone model). For the CNN-LSTM-A model we have exploited the attention mechanism developed by Z. Yang et al. in [60].

The results of the cross-corpus experiments are presented in Table 6. According to the results, the utilizing of Hamming window (CNN-W) for smoothing allows to slightly increase the models' performance on all considered datasets. Surprisingly, the CNN-SVM approach worked worse on all VEDs except CREMA-D. This can be due to the temporal complexity, which statistical parameters (Means and STDs) used by SVM were not able to encode well enough.

Regarding CNN-LSTM and CNN-LSTM-A approaches, both have demonstrated a significant improvement in the UAR in comparison with other approaches. Replacement of the LSTM layers with GRU layers has overall decreased the performance. This can be partially explained by the decrease in the number of tuning parameters that worsened the generalization ability of the GRU-based network. From the observation of the results, however, it is difficult to highlight only CNN-LSTM or CNN-LSTM-A, since it highly depends on the considered VED. We have conducted the Student's paired t-

Table 5
Comparison with the state-of-the-art results on the AffectNet validation set.

Research	Approach	Accuracy, %
Wang et al. [76]	SCN	60.2
Kervadec et al. [77]	CAKE	61.7
She et al. [78]	Res-50IBN	63.1
Georgescu et al. [79]	VGG and BOVW features + SVM	63.3
Kollias et al. [80]	FaceBehaviorNet	65.0
Savchenko [81]	EfficientNet-B2	66.3
This work	**Backbone Model**	**66.4**

Table 6
The results of the cross-corpus experiments (UAR, %).

VED	CNN-W	CNN-S	CNN-SVM	CNN-LSTM (δ)	CNN-GRU (δ)	CNN-LSTM-A (δ)	CNN-LSTM-CV
RAMAS	42.8	42.8	26.5	44.3 (1.5)	43.9 (1.1)	**45.0** (2.2)	50.2
RAVDESS	59.6	59.8	56.6	65.8 (6.2)	65.2 (5.6)	**66.2** (6.6)	69.7
CREMA-D	54.5	55.4	60.6	**60.6** (6.1)	57.5 (3.0)	59.7 (5.2)	79.0
IEMOCAP	25.7	25.9	**26.3**	25.1 (-0.6)	26.0 (0.3)	24.8 (-0.9)	28.7
SAVEE	62.1	63.7	38.3	76.1 (14.0)	**77.3** (15.2)	77.0 (14.9)	82.8
AffWild2	45.1	46.2	33.0	**51.6** (6.5)	44.3 (-0.8)	49.0 (3.9)	52.9
Average UAR	48.3	48.8	40.2	**53.9** (5.6)	52.4 (4.1)	53.6 (5.3)	60.6

test to both CNN-LSTM and CNN-LSTM-A results to figure out the difference in the efficiency statistically. With the t-score = 0.5274 and alpha value p = 0.05 the t-test found no statistically significant difference between these two methods. Thus, both of them can be used for efficient emotion recognition, yet the CNN-LSTM model contains fewer parameters and, therefore, can be more preferable. In conclusion, we should say that during the leave-one-corpus-out cross-validation procedure we have achieved on average a 5.61% increase in terms of UAR comparing to the baseline simplest CNN-W approach.

To test the model's generalization ability more, we have conducted participant-independent cross-validation as well (see Table 6, column CNN-LSTM-CV). This was carried out with 5 folds for every dataset (for SAVEE – 4 folds). We should also note that we have preserved the deliberate separation of the AffWild2 on the train and developments sets, while performed leave-one-session-out for RAMAS and IEMOCAP datasets.

We have used the CNN-LSTM approach for evaluation since it outperformed all others and is computationally more effective than CNN-LSTM-A. The results of the experiments are also presented in Table 6. Participant-independent cross-validation allows analyzing the functioning of the model in cases when it faces completely a new, unknown before person. An efficient model should "omit" the unimportant differences in faces, highlighting and exploiting only facial features salient for affective computing (in our case, emotion recognition). As one can see from Table 6, the participant-independent experiments have shown a dramatic increase (6.7% on average) in the CNN-LSTM model performance in comparison with cross-corpus experiments. While it was expected, we would like to point out that both conducted experiments prove the ability of the model to generalize the learned features regardless of the variability in faces, lightning, obstacles, and other recording conditions.

5. Discussion

In this Section, we discuss the features of the developed framework, compare it with other state-of-the-art emotion recognition works, and provide an analysis of the augmentation techniques, which dramatically influenced on the model efficacy.

5.1. Comparison with State-of-the-Art Works

The cross-corpus problem has been attracting researchers attention for a long time since without its resolving it is not possible to develop a truly general emotion recognition system. Although there are quite many studies devoted to it, most of them investigate the audio modality and use 2–3 corpora. To the best of our knowledge, our work, in turn, is the first research conducted on a large amount of considered VEDs on visual modality.

For this reason, it is difficult to compare the obtained results with other papers, where the same datasets were exploited separately. The comparison would be inadequate, since in these works the system was trained and tested on the same corpus, obviously leading to the higher performance. However, we can collate the results of our former [44] work, where we have conducted a study on two corpora (CREMA-D and RAVDESS), with the current one. The comparison in terms of the *Accuracy* performance measure is presented in Table 7. We can note that the current model efficiency is comparable with our former study despite the bigger amount of in-the-wild training data, which can "confuse" the model during the evaluation phase on the laboratory-obtained data.

To make the comparison choices wider, we have included in Table 7 the results of the participant-independent cross-validation studies [82–84] in terms of UAR and F1-score (the measure was chosen depending on scores reported by authors). Such works can be viewed as cross-corpus research to some extent due to the fact that the validation part of the data does not contain participants included in the training data, nevertheless, it is much easier than the original cross-corpus problem. From the results, we can mention that the developed system has lower UAR on CREMA-D and SAVEE datasets, while higher F1-score on the AffWild2 dataset. It can be caused, as we noted earlier, by the "nature" of the data: the model was trained on more diverse and biased towards in-the-wild data (represented by AffWild2), while the participant-independent cross-validation studies utilized solely laboratory-controlled data.

5.2. Backbone Model

Currently, the facial expression recognition models in the affective computing field are still far from perfection in terms of model accuracy. This is because of two main things: (1) the model imperfection and (2) the corruption in the data annotation (mislabelling, raters' subjectivity) used for training. Although we cannot state that our backbone model obtained during the study has no shortcomings at all, we would like to demonstrate its robustness and decent performance via analysis of its functioning on various complex frames from dynamic VEDs.

To show where the model is "looking at" during the decision-making process, we have exploited the GradCAM [85] technique, obtaining the gradient heatmaps for every frame under the consideration. Fig. 7 presents the heatmaps for eight correctly and incorrectly recognized facial expressions, which were randomly chosen from dynamic VEDs. The redder the area of the image, the more attention "pays" the model to this area.

Table 7
The comparison with other cross-corpus and participant-independent cross validation studies. *CC* means cross-corpus, *PI* – participant-independent

#	Type	VED	Metric	CNN-LSTM	Other studies
1	CC	RAVDESS	Accuracy	67.3	**69.4** [44]
2	CC	CREMA-D	Accuracy	**50.9**	49.9 [44]
3	PI	CREMA-D	UAR	66.6	**73.9** [82]
4	PI	SAVEE	UAR	82.8	**86.5** [83]
5	PI	AffWild2	F1-score	**44.0**	37.0 [84]

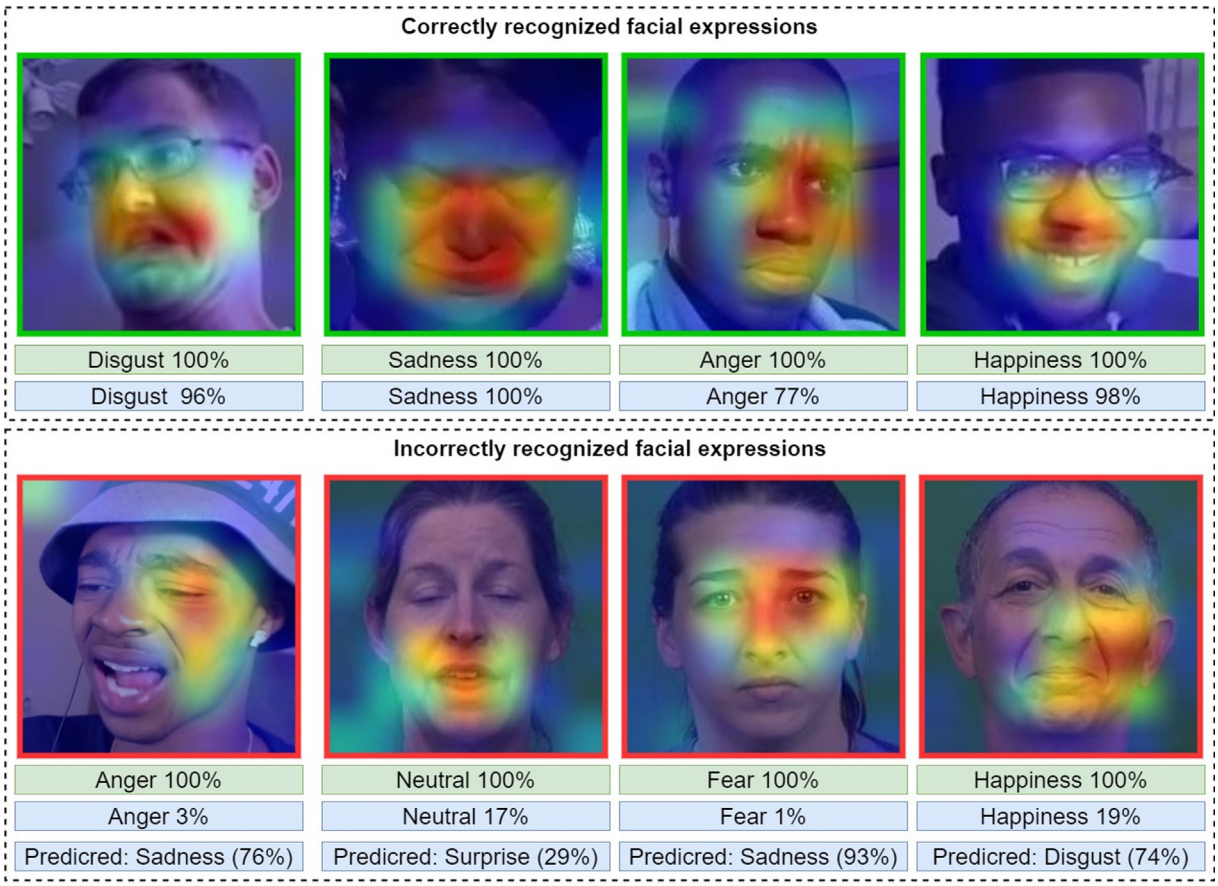

Fig. 7. GradCAM heatmaps for the backbone model with correctly/incorrectly recognized facial expressions.

Analyzing Fig. 7, we can note that the model correctly identifies the important regions of the faces for certain emotions despite the difficulties caused by head tilt. For instance, on the picture with the *Disgust*, it correctly (according to Ekman) pays attention to raised upper lip and wrinkles around the nose. When we take a look at a picture with *Sadness*, we can observe that model is taking into account the dropped down eyelids and pulled together brows.

Nevertheless, there were also cases, where the model made mistakes. For example, the model wrongly predicts the state of the man depicted in the down left corner of Fig. 7. Looking at the almost closed eyes and noticeable bags under the eyes, the model is being confused, predicting an emotion of *Sadness* instead of the annotated emotion *Anger*. Observing another participant's face in the down-right corner (*Happiness*), we suppose that the model was confused by the lacking of a smile on the face. Taking into account the wrinkles, it has wrongly predicted the emotion *Disgust*.

Because of the subjectivity of emotions, frequently datasets contain corrupted annotations, which can highly bias the model, confusing it and decreasing its efficiency. To examine it, we have chosen several frames, which show, in our opinion (and based on Ekman's rules), emotions different from annotated and depicted them in Fig. 8.

For instance, the woman with annotated emotion *Neutral* has all features of the emotion *Sadness*, except for the lips: the eyebrows are raised and pulled together, the eyelids almost fully dropped down, tears on the face, and the wrinkles on the sides of the nose. That is why the model has almost no response in terms of a gradient for the *Neutral* emotion. On the other hand, when the emotion *Sadness* for the GradCAM is chosen, the model correctly takes into account all the aforementioned facial areas, predicting *Sadness* with 96%.

Considering another instance with initially annotated *Anger* (Fig. 8, upper right corner), we can note similar things: the jaw dropped open, raised and pulled together eyebrows, raised upper eyelids, and tensed lower eyelids indicate that it is rather the emotion *Fear*. The raters likely have chosen the *Anger*, because of the widely opened eyes and pulled together eyebrows. It could also be caused by the interpolation of the annotations: while two frames (separated by the annotation frequency) were expressing, for example, *Anger*, the participant's state *Fear* between these two time points was missed.

Nevertheless, we can partially eliminate such problem as corrupted labels by using the Mixup technique, which mixes two images and, therefore, weakens the influence of the corrupted labels on the model. In the next subsection, we present its analysis, demonstrating the functioning of the model on separated and mixed images.

5.3. The Mixup Technique Analysis

As we described earlier, the backbone model was trained using to Mixup data augmentation technique (see Section 4.2), which has given a significant gain in the model performance expressed in UAR. This is because Mixup allows us to weaken (smooth) labels by combining images with different categories, forcing network to refuse from the prediction *overconfidence*. Moreover, besides regularizing the model, it increases its robustness by resisting corrupted labels (labels, which were annotated incorrectly) [67]. To show aforementioned Mixup features, we have randomly taken two images from AffectNet development set, applied to them Mixup technique with randomly generated mixing ratio $\alpha = 0.64$,

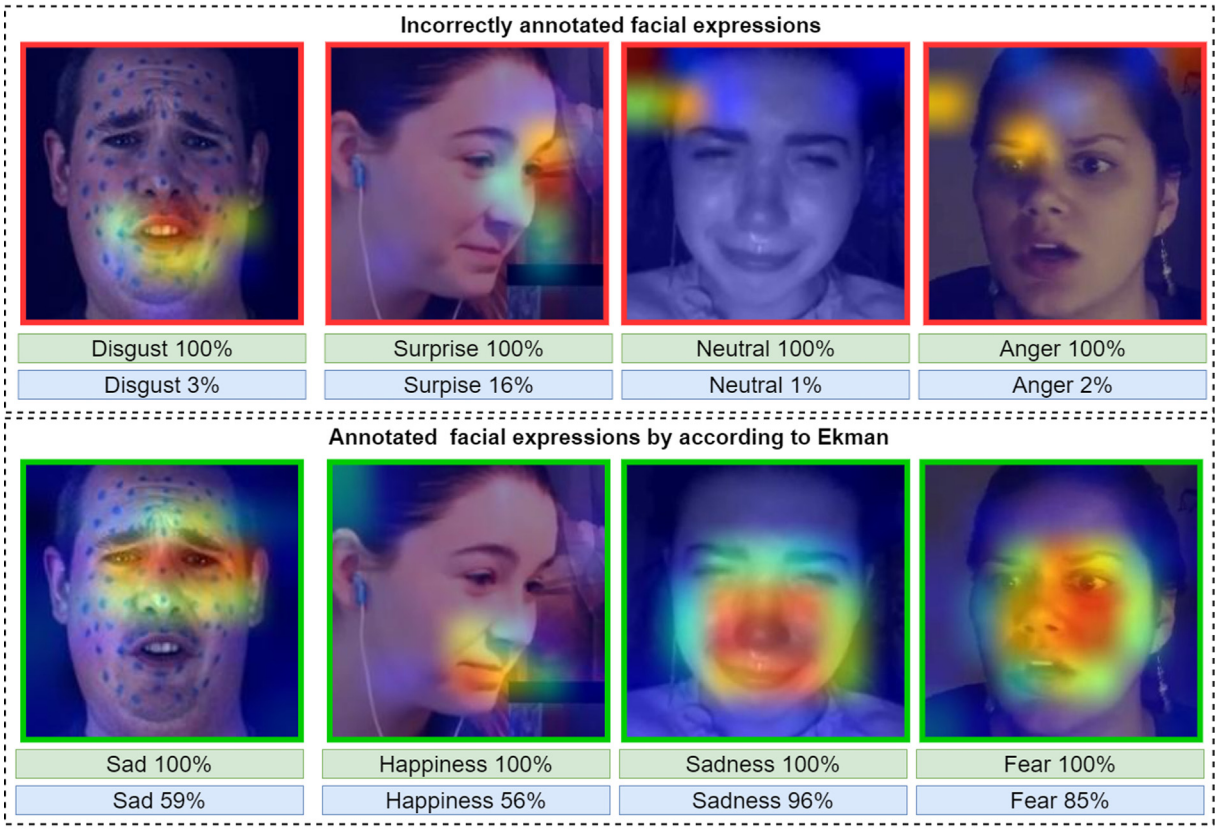

Fig. 8. GradCAM heatmaps for the backbone model with incorrectly/correctly annotated facial expressions.

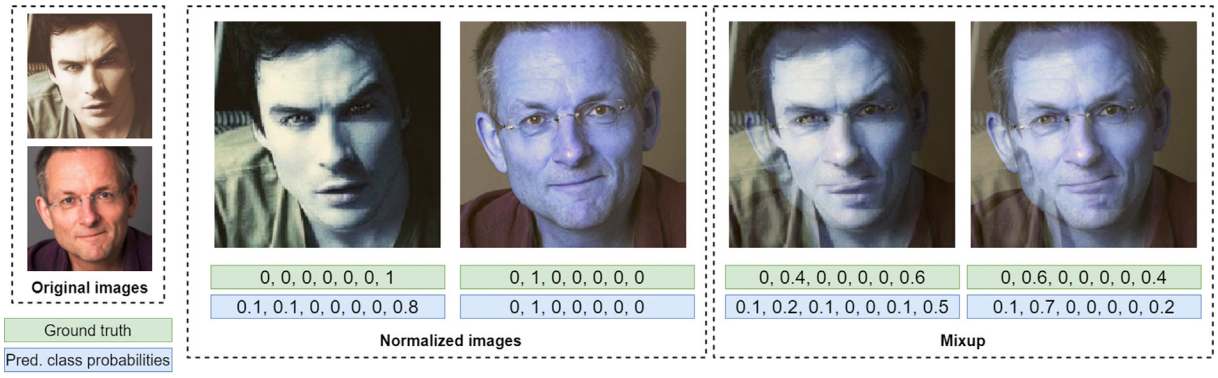

Fig. 9. The evaluation of the Backbone model on examples with and without applied Mixup.

and obtained a class probability prediction for trained backbone model. The result is depicted in Fig. 9.

Vectors in green and blue are probability vectors for 7 emotional classes (the order is *Neutral, Happiness, Sadness, Surprise, Fear, Disgust, Anger*). In case of ground truth labels (green color), they are one-hot vectors, while in row with the blue color the probability prediction generated by model is presented. Analyzing Fig. 9, we can observe that the backbone model functions well both with one-hot and smoothed labels. For instance, the model predicted the right one of the normalized images as a *Happiness* state with 100 % probability, which is very close to the one-hot encoding. On the other hand, when mixing this image with another one (the block Mixup, right image), the model correctly forecasts the *Happiness* state (the error is only 0.1), while predicts also the *Anger* state of another imposed image with probability of 0.2. Thus, one can see that the model during the training was adapted to both one-hot and smoothed images, which allowed it to overcome the problems with prediction overconfidence.

During the training, we have considered the erroneously classified images presented in AffectNet development set. After the analysis, we have found that some of them corrupted in the same way as we noted in the previous section. However, it is interesting to see how the model works on mixed images with corrupted and non-corrupted labels. Fig. 10 shows an illustrative example. Images framed in red/green color show an incorrect/correct predicted class label.

From Fig. 10 one can see that the first normalized image is labeled as *Disgust*, while it contains all the features of the *Surprise* emotion (eyebrows raised, but not drawn together; upper eyelids raised, lower eyelids neutral; jaw dropped down), turning out a

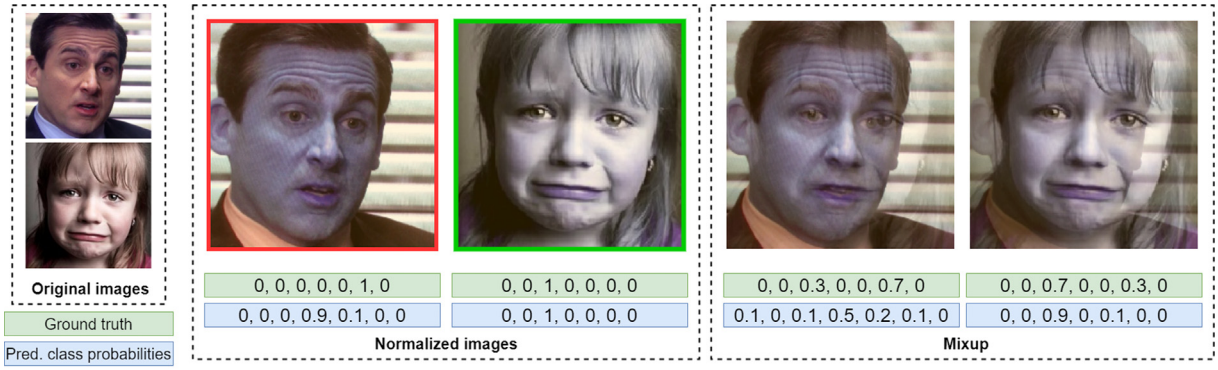

Fig. 10. The evaluation of the Backbone model on incorrectly annotated example with and without applied Mixup.

corrupted label. After mixing this image with *Sadness* annotated image, we got the probability 0.3 of *Sadness* and 0.7 of *Disgust* for the first generated sample, and vise versa for the second one. In case the corrupted label prevails (first generated image), one can see that the model is confused, trying to focus simultaneously on two presented emotions, although preferring to stand still on emotion *Surprise*. Such confusion forces the model to be doubtful and does not make strict decisions (like one-hot encoding), which helps not to focus entirely on the corrupted label. Regarding the second generated image, we can observe that in case the emotion is expressed in a really strong way, the model prefers to give such emotion a high probability, nevertheless, softening it a little bit. Thus, the Mixup technique allows us to soft the biasing of the image with a corrupted label by mixing it with a non-corrupted one, obtaining two new samples (less corrupted in comparison with previous ones), and making the model to doubt and cope with complex "two-emotional" states.

In addition, we would like to note that figures presented in this subsection indicate once more that the backbone model is adapted to different ages, genders, lightning conditions, head movements, and obstacles, making it a good base for efficient emotion recognition model.

6. Conclusions

In this article, we have presented probably one of the largest cross-corpus studies on visual emotion recognition at the moment. We suggested a novel and effective E2E emotion recognition framework consisting of two key elements, which are employed for different functions: (1) the backbone emotion recognition model, which is based on the VGGFace2 ResNet50 model, trained in a balanced way, and able to predict emotion from the raw image with high performance and (2) the temporal block stacked on top of the backbone model and trained with dynamic VEDs using the cross-corpus protocol in order to show its reliability and effectiveness.

During the research, the backbone model was fine-tuned on the largest facial expression corpus AffectNet containing static images. Our backbone model achieved the accuracy of 66.4 % on the AffectNet validation set, outperforming all currently known state-of-the-art works.

Models trained on static images are usually biased since such images demonstrate the peak levels of emotions, that are rarely expressed in-the-wild conditions. Moreover, utilizing only the backbone model, it is not possible to take into account temporal information, yet it is extremely important in-the-wild circumstances: exploiting context information (former video frames), the model can overcome such problems as bad lightning, occasional occlusion, noises, etc. Therefore, we have conducted extensive experiments on six datasets (RAMAS, IEMOCAP, CREMA-D, RAVDESS, SAVEE, AffWild2) with various temporal aggregation techniques (CNN-SVM, CNN-LSTM, CNN-GRU, and CNN-LSTM-A), leveraging the leave-one-corpus-out protocol to figure out the best method in terms of its performance and ability to generalize.

Analysis of the results shows that CNN-LSTM and CNN-LSTM-A approaches provide the best efficiency on most research datasets (45%, 66.2%, 60.6%, 25.1%, 77%, and 51.6% on RAMAS, IEMOCAP, CREMA-D, RAVDESS, SAVEE, and AffWild2 datasets accordingly). Moreover, we have achieved and, in some cases, outperformed the state-of-the-art results. However, there is no statistically significant difference in performance between these two approaches. Thus, we suggest using the CNN-LSTM model, since it has fewer parameters and more computationally efficient.

Additionally, to show the predictive balance of the model, we have studied model functioning on the images randomly selected from dynamic VEDs, analyzing the model behavior in case of mixed labels and utilizing the GradCAM technique. We found that the backbone model demonstrates a good performance not only on the images expressing one emotion, but also with complex combined emotional states. Moreover, the usage of the GradCAM allowed showing that the model "looks" on right salient areas of the facial expressions even when they have occlusions, bad lightning conditions or turned apart from the camera.

For the possibility to reproduce our results, we provide trained models and the code for testing them as well on GitHub[2].

In our future research, we plan to perform a cross-corpus analysis for both acoustic and linguistic modalities in order to fuse different information channels and systems that process them into one multi-modal emotion recognition system, which is a promising approach based on some previous studies [86,87].

Funding

This work was supported by the Analytical Center for the Government of the Russian Federation (IGK 000000D730321P5Q0002), agreement No. 70–2021-00141.

CRediT authorship contribution statement

Elena Ryumina: Conceptualization, Methodology, Investigation, Validation, Visualization, Writing - original draft. **Denis Dresvyanskiy:** Conceptualization, Methodology, Investigation, Writing - original draft, Visualization. **Alexey Karpov:** Supervision, Writing - review & editing, Funding acquisition.

[2] https://github.com/ElenaRyumina/EMO-AffectNetModel

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] J. Yang, R. Wang, X. Guan, M.M. Hassan, A. Almogren, A. Alsanad, AI-enabled emotion-aware robot: The fusion of smart clothing, edge clouds and robotics, Future Generation Computer Systems 102 (2020) 701–709, https://doi.org/10.1016/j.future.2019.09.029.

[2] Z. Liu, M. Wu, W. Cao, L. Chen, J. Xu, R. Zhang, M. Zhou, J. Mao, A facial expression emotion recognition based human-robot interaction system, IEEE/CAA Journal of Automatica Sinica 4 (4) (2017) 668–676, https://doi.org/10.1109/JAS.2017.7510622.

[3] A. Shukla, S.S. Gullapuram, H. Katti, K. Yadati, M. Kankanhalli, R. Subramanian, Affect recognition in ads with application to computational advertising, in: 25th ACM International Conference on Multimedia, 2017, pp. 1148–1156, https://doi.org/10.1145/3123266.3123444.

[4] S. Cosentino, E.I. Randria, J.-Y. Lin, T. Pellegrini, S. Sessa, A. Takanishi, Group emotion recognition strategies for entertainment robots, in: IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), 2018, pp. 813–818, https://doi.org/10.1109/IROS.2018.8593503.

[5] Z. Fei, E. Yang, D.D.-U. Li, S. Butler, W. Ijomah, X. Li, H. Zhou, Deep convolution network based emotion analysis towards mental health care, Neurocomputing 388 (2020) 212–227, https://doi.org/10.1016/j.neucom.2020.01.034.

[6] M.S. Hossain, G. Muhammad, Emotion-aware connected healthcare big data towards 5G, IEEE Internet of Things Journal 5 (4) (2017) 2399–2406, https://doi.org/10.1109/JIOT.2017.2772959.

[7] D. Yang, A. Alsadoon, P.C. Prasad, A.K. Singh, A. Elchouemi, An emotion recognition model based on facial recognition in virtual learning environment, Procedia Computer Science 125 (2018) 2–10, https://doi.org/10.1016/j.procs.2017.12.003.

[8] P. Ekman, W. Friesen, Nonverbal leakage and clues to deception, Psychiatry 32 (1) (1969) 88–106, https://doi.org/10.1080/00332747.1969.11023575.

[9] J.A. Russell, A circumplex model of affect, Journal of Personality and Social Psychology 39 (6) (1980) 1161–1178, https://doi.org/10.1037/h0077714.

[10] J. Parry, D. Palaz, G. Clarke, P. Lecomte, R. Mead, M. Berger, G. Hofer, Analysis of deep learning architectures for cross-corpus speech emotion recognition, Interspeech (2019) 1656–1660, https://doi.org/10.21437/Interspeech.2019-2753.

[11] S. Zhang, S. Zhang, T. Huang, W. Gao, Q. Tian, Learning affective features with a hybrid deep model for audio–visual emotion recognition, IEEE Transactions on Circuits and Systems for Video Technology 28 (10) (2018) 3030–3043, https://doi.org/10.1109/TCSVT.2017.2719043.

[12] E. Friesen, P. Ekman, Facial action coding system: a technique for the measurement of facial movement, Palo Alto 3 (2) (1978) 5.

[13] C. Shu, X. Ding, C. Fang, Histogram of the oriented gradient for face recognition, Tsinghua Science and Technology 16 (2) (2011) 216–224, https://doi.org/10.1016/S1007-0214(11)70032-3.

[14] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, in: 3rd International Conference on Learning Representations (ICLR), 2015, pp. 1–14.

[15] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 770–778, https://doi.org/10.1109/CVPR.2016.90.

[16] H.-W. Ng, V.D. Nguyen, V. Vonikakis, S. Winkler, Deep learning for emotion recognition on small datasets using transfer learning, in: 17th ACM on International Conference on Multimodal Interaction, 2015, pp. 443–449, https://doi.org/10.1145/2818346.2830593.

[17] I.J. Goodfellow, D. Erhan, P.L. Carrier, A. Courville, M. Mirza, B. Hamner, W. Cukierski, Y. Tang, D. Thaler, D.H. Lee, et al., Challenges in representation learning: A report on three machine learning contests, in: International Conference on Neural Information Processing, 2013, pp. 117–124, https://doi.org/10.1007/978-3-642-42051-1_16.

[18] G. Levi, T. Hassner, Emotion recognition in the wild via convolutional neural networks and mapped binary patterns, in: 17th ACM on International Conference on Multimodal Interaction, 2015, pp. 503–510, https://doi.org/10.1145/2818346.2830587.

[19] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, A. Rabinovich, Going deeper with convolutions, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 1–9, https://doi.org/10.1109/CVPR.2015.7298594.

[20] S.A. Bargal, E. Barsoum, C.C. Ferrer, C. Zhang, Emotion recognition in the wild from videos using images, in: 18th ACM International Conference on Multimodal Interaction, 2016, pp. 433–436, https://doi.org/10.1145/2993148.2997627.

[21] P. Balouchian, H. Foroosh, Context-sensitive single-modality image emotion analysis: A unified architecture from dataset construction to cnn classification, in: 25th IEEE International Conference on Image Processing (ICIP), 2018, pp. 1932–1936, https://doi.org/10.1109/ICIP.2018.8451048.

[22] M.-C. Sun, S.-H. Hsu, M.-C. Yang, J.-H. Chien, Context-aware cascade attention-based RNN for video emotion recognition, in: First Asian Conference on Affective Computing and Intelligent Interaction (ACII Asia), 2018, pp. 1–6. doi:10.1109/ACIIAsia.2018.8470372.

[23] J. Lee, S. Kim, S. Kim, J. Park, K. Sohn, Context-aware emotion recognition networks, IEEE/CVF International Conference on Computer Vision (2019) 10143–10152, https://doi.org/10.1109/ICCV.2019.01024.

[24] D. Nguyen, K. Nguyen, S. Sridharan, A. Ghasemi, D. Dean, C. Fookes, Deep spatio-temporal features for multimodal emotion recognition, in: IEEE Winter Conference on Applications of Computer Vision (WACV), 2017, pp. 1215–1223, https://doi.org/10.1109/WACV.2017.140.

[25] S. Zhang, S. Zhang, T. Huang, W. Gao, Q. Tian, Learning affective features with a hybrid deep model for audio-visual emotion recognition, IEEE Transactions on Circuits and Systems for Video Technology 28 (10) (2017) 3030–3043, https://doi.org/10.1109/TCSVT.2017.2719043.

[26] T. Mittal, U. Bhattacharya, R. Chandra, A. Bera, D. Manocha, M3er: Multiplicative multimodal emotion recognition using facial, textual, and speech cues, in: AAAI Conference on Artificial Intelligence, Vol. 34, 2020, pp. 1359–1367. doi:10.1609/aaai.v34i02.5492.

[27] J. Huang, J. Tao, B. Liu, Z. Lian, M. Niu, Multimodal transformer fusion for continuous emotion recognition, in: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2020, pp. 3507–3511, https://doi.org/10.1109/ICASSP40776.2020.9053762.

[28] H. Kaya, F. Gürpınar, A.A. Salah, Video-based emotion recognition in the wild using deep transfer learning and score fusion, Image and Vision Computing 65 (2017) 66–75, https://doi.org/10.1016/j.imavis.2017.01.012.

[29] E. Avots, T. Sapiński, M. Bachmann, D. Kamińska, Audiovisual emotion recognition in wild, Machine Vision and Applications 30 (2019) 975–985, https://doi.org/10.1007/s00138-018-0960-9.

[30] F. Noroozi, M. Marjanovic, A. Njegus, S. Escalera, G. Anbarjafari, Audio-visual emotion recognition in video clips, IEEE Transactions on Affective Computing 10 (1) (2017) 60–75, https://doi.org/10.1109/TAFFC.2017.2713783.

[31] M. Wu, W. Su, L. Chen, W. Pedrycz, K. Hirota, Two-stage fuzzy fusion based-convolution neural network for dynamic emotion recognition, IEEE Transactions on Affective Computing 1 (2020) 1–13, https://doi.org/10.1109/TAFFC.2020.2966440.

[32] H. Kaya, A.A. Karpov, Efficient and effective strategies for cross-corpus acoustic emotion recognition, Neurocomputing 275 (2018) 1028–1034, https://doi.org/10.1016/j.neucom.2017.09.049.

[33] B. Zhang, E.M. Provost, G. Essl, Cross-corpus acoustic emotion recognition with multi-task learning: Seeking common ground while preserving differences, IEEE Transactions on Affective Computing 10 (1) (2017) 85–99, https://doi.org/10.1109/TAFFC.2017.2684799.

[34] H. Kaya, D. Fedotov, A. Yesilkanat, O. Verkholyak, Y. Zhang, A. Karpov, LSTM based cross-corpus and cross-task acoustic emotion recognition, Interspeech (2018) 521–525, https://doi.org/10.21437/Interspeech.2018-2298.

[35] J. Parry, D. Palaz, G. Clarke, P. Lecomte, R. Mead, M. Berger, G. Hofer, Analysis of deep learning architectures for cross-corpus speech emotion recognition, Interspeech (2019) 1656–1660, https://doi.org/10.21437/Interspeech.2019-2753.

[36] H. Meng, T. Yan, F. Yuan, H. Wei, Speech emotion recognition from 3D log-mel spectrograms with deep learning network, IEEE Access 7 (2019) 125868–125881, https://doi.org/10.1109/ACCESS.2019.2938007.

[37] A. Mollahosseini, D. Chan, M.H. Mahoor, Going deeper in facial expression recognition using deep neural networks, in: IEEE Winter Conference on Applications of Computer Vision (WACV), 2016, pp. 1–10, https://doi.org/10.1109/WACV.2016.7477450.

[38] W. Xie, X. Jia, L. Shen, M. Yang, Sparse deep feature learning for facial expression recognition, Pattern Recognition 96 (2019), https://doi.org/10.1016/j.patcog.2019.106966.

[39] M.V. Zavarez, R.F. Berriel, T. Oliveira-Santos, Cross-database facial expression recognition based on fine-tuned deep convolutional network, in: 30th IEEE Conference on Graphics, Patterns and Images (SIBGRAPI), 2017, pp. 405–412, https://doi.org/10.1109/SIBGRAPI.2017.60.

[40] Q. Cao, L. Shen, W. Xie, O. Parkhi, A. Zisserman, Vggface2: A dataset for recognising faces across pose and age, in: 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG), 2018, pp. 67–74, https://doi.org/10.1109/FG.2018.00020.

[41] G. Wen, Z. Hou, H. Li, D. Li, L. Jiang, E. Xun, Ensemble of deep neural networks with probability-based fusion for facial expression recognition, Cognitive Computation 9 (2017) 597–610, https://doi.org/10.1007/s12559-017-9472-6.

[42] Z. Meng, P. Liu, J. Cai, S. Han, Y. Tong, Identity-aware convolutional neural network for facial expression recognition, in: 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG), 2017, pp. 558–565, https://doi.org/10.1109/FG.2017.140.

[43] B. Hasani, M.H. Mahoor, Facial expression recognition using enhanced deep 3D convolutional neural networks, in: IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 2017, pp. 30–40, https://doi.org/10.1109/CVPRW.2017.282.

[44] E. Ryumina, A. Karpov, Facial expression recognition using distance importance scores between facial landmarks, CEUR Workshop Proceedings 2744 (2020) 1–10, https://doi.org/10.51130/graphicon-2020-2-3-32.

[45] O. Perepelkina, E. Kazimirova, M. Konstantinova, RAMAS: Russian multimodal corpus of dyadic interaction for affective computing, in: 20th International Conference on Speech and Computer, 2018, pp. 501–510, https://doi.org/10.1007/978-3-319-99579-3_52.

[46] C. Busso, M. Bulut, C.C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J.N. Chang, S. Lee, S.S. Narayanan, IEMOCAP: Interactive emotional dyadic motion capture

database, Language Resources and Evaluation 42 (2008) 335–359, https://doi.org/10.1007/s10579-008-9076-6.
[47] H. Cao, D.G. Cooper, M.K. Keutmann, R.C. Gur, A. Nenkova, R. Verma, CREMA-D: Crowd-sourced emotional multimodal actors dataset, IEEE Transactions on Affective Computing 5 (4) (2014) 377–390, https://doi.org/10.1109/TAFFC.2014.2336244.
[48] S.R. Livingstone, F.A. Russo, The ryerson audio-visual database of emotional speech and song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in north american english, PLoS One 13 (5) (2018), https://doi.org/10.1371/journal.pone.0196391.
[49] S. Haq, P. Jackson, J.R. Edge, Audio-visual feature selection and reduction for emotion classification, in: International Conference on Auditory-Visual Speech Processing, 2008, pp. 185–190.
[50] D. Kollias, S. Zafeiriou, Expression, affect, action unit recognition: Aff-Wild2, multi-task learning and ArcFace, ArXiv abs/1910.04855 (2019) 1–15.
[51] A. Mollahosseini, B. Hasani, M.H. Mahoor, Affectnet: A database for facial expression, valence, and arousal computing in the wild, IEEE Transactions on Affective Computing 10 (1) (2017) 18–31, https://doi.org/10.1109/TAFFC.2017.2740923.
[52] D. Kollias, A. Schulc, E. Hajiyev, S. Zafeiriou, Analysing affective behavior in the first ABAW 2020 competition, in: 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG), 2020, pp. 794–800, https://doi.org/10.1109/FG47880.2020.00126.
[53] D. Kollias, S. Zafeiriou, A multi-task learning & generation framework: Valence-arousal, action units & primary expressions, ArXiv abs/1811.07771 (2018) 1–9.
[54] D. Kollias, S. Zafeiriou, Aff-Wild2: Extending the Aff-Wild database for affect recognition, ArXiv abs/1811.07770 (2018) 1–8.
[55] D. Kollias, P. Tzirakis, M.A. Nicolaou, A. Papaioannou, G. Zhao, B. Schuller, I. Kotsia, S. Zafeiriou, Deep affect prediction in-the-wild: Aff-Wild database and challenge, deep architectures, and beyond, International Journal of Computer Vision 127 (2019) 907–929, https://doi.org/10.1007/s11263-019-01158-4.
[56] S. Zafeiriou, D. Kollias, M.A. Nicolaou, A. Papaioannou, G. Zhao, B. Schuller, I. Kotsia, Aff-wild: Valence and arousal 'in-the-wild' challenge, in: IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 2017, pp. 1980–1987, https://doi.org/10.1109/CVPRW.2017.248.
[57] D. Kollias, M.A. Nicolaou, I. Kotsia, G. Zhao, S. Zafeiriou, Recognition of affect in the wild using deep neural networks, in: IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 2017, pp. 1972–1979, https://doi.org/10.1109/CVPRW.2017.247.
[58] D. Dresvyanskiy, E. Ryumina, H. Kaya, M. Markitantov, A. Karpov, W. Minker, End-to-end modeling and transfer learning for audiovisual emotion recognition in-the-wild, Multimodal Technologies and Interaction 6 (2) (2022) 1–23, https://doi.org/10.3390/mti6020011.
[59] G. Winata, O. Kampman, in: F. P, Attention-based LSTM, for psychological stress detection from spoken language using distant supervision, in IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2018, pp. 6204–6208, https://doi.org/10.1109/ICASSP.2018.8461990.
[60] Z. Yang, D. Yang, C. Dyer, X. He, A. Smola, H.E., Hierarchical attention networks for document classification, in: Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2016, pp. 1480–1489. doi:10.18653/v1/N16-1174.
[61] I. Loshchilov, F. Hutter, SGDR: Stochastic gradient descent with warm restarts, ArXiv abs/1608.03983 (2016) 1–16.
[62] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 770–778, https://doi.org/10.1109/CVPR.2016.90.
[63] J. Hu, L. Shen, G. Sun, Squeeze-and-excitation networks, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 7132–7141.
[64] M. Tan, Q.V. Le, EfficientNet: Rethinking model scaling for convolutional neural networks, in: International Conference on Machine Learning (ICML), 2019, pp. 6105–6114.
[65] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, L. Chen, Mobilenetv 2: Inverted residuals and linear bottlenecks, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 4510–4520, https://doi.org/10.1109/CVPR.2018.00474.
[66] J. Deng, W. Dong, R. Socher, L.J. Li, K. Li, L. Fei-Fei, Imagenet: A large-scale hierarchical image database, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2009, pp. 248–255, https://doi.org/10.1109/CVPR.2009.5206848.
[67] H. Zhang, M. Cissé, Y.N. Dauphin, D. Lopez-Paz, Mixup: Beyond empirical risk minimization, in: 3rd International Conference on Learning Representations (ICLR), 2018.
[68] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.Y. Fu, A.C. Berg, SSD: Single shot multibox detector, in: European Conference on Computer Vision, Amsterdam, 2016, pp. 21–37. doi:10.1007/978-3-319-46448-0_2.
[69] J. Deng, J. Guo, E. Ververas, I. Kotsia, S. Zafeiriou, RetinaFace: Single-shot multi-level face localisation in the wild, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 5203–5212, https://doi.org/10.1109/CVPR42600.2020.00525.
[70] K. Zhang, Z. Zhang, Z. Li, Y. Qiao, Joint face detection and alignment using multitask cascaded convolutional networks, IEEE Signal Processing Letters 23 (10) (2016) 1499–1503, https://doi.org/10.1109/LSP.2016.2603342.
[71] E. Ryumina, D. Ryumin, D. Ivanko, A. Karpov, A novel method for protective face mask detection using convolutional neural networks and image histograms, in: International Archives of the Photogrammetry Remote Sensing and Spatial Information Sciences XLIV-2/W1-2021, 2021, pp. 177–182, https://doi.org/10.5194/isprs-archives-XLIV-2-W1-2021-177-2021.
[72] E. Ryumina, O. Verkholyak, A. Karpov, Annotation confidence vs. training sample size: Trade-off solution for partially-continuous categorical emotion recognition, Interspeech (2021) 3690–3694, https://doi.org/10.21437/Interspeech.2021-1636.
[73] S. Poria, N. Majumder, D. Hazarika, E. Cambria, A. Gelbukh, A. Hussain, Multimodal sentiment analysis: Addressing key issues and setting up the baselines, IEEE Intelligent Systems 33 (6) (2018) 17–25, https://doi.org/10.1109/MIS.2018.2882362.
[74] L. Wang, Y. Xiong, Z. Wang, Y. Qiao, D. Lin, X. Tang, L.V. Gool, Temporal segment networks: Towards good practices for deep action recognition, in: European conference on computer vision, Springer, 2016, pp. 20–36.
[75] E. Ryumina, A. Karpov, Comparative analysis of methods for imbalance elimination of emotion classes in video data of facial expressions, Scientific and Technical Journal of Information Technologies, Mechanics and Optics 20 (5 (129)) (2020) 683–691, https://doi.org/10.17586/2226-1494-2020-20-5-683-691.
[76] K. Wang, X. Peng, J. Yang, S. Lu, Y. Qiao, Suppressing uncertainties for large-scale facial expression recognition, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 6897–6906.
[77] C. Kervadec, V. Vielzeuf, S. Pateux, A. Lechervy, F. Jurie, CAKE: a compact and accurate k-dimensional representation of emotion, British Machine Vision Association (2018) 1–12.
[78] J. She, Y. Hu, H. Shi, J. Wang, Q. Shen, T. Mei, Dive into ambiguity: latent distribution mining and pairwise uncertainty estimation for facial expression recognition, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 6248–6257.
[79] M.-I. Georgescu, R.T. Ionescu, M. Popescu, Local learning with deep and handcrafted features for facial expression recognition, IEEE Access (2019) 64827–64836, https://doi.org/10.1109/ACCESS.2019.2917266.
[80] D. Kollias, V. Sharmanska, S. Zafeiriou, Distribution matching for heterogeneous Multi-Task learning: a Large-Scale face study, ArXiv abs/2105.03790 (2021) 1–15.
[81] A.V. Savchenko, Facial expression and attributes recognition based on Multi-Task learning of lightweight neural networks, in: 2021 IEEE 19th International Symposium on Intelligent Systems and Informatics (SISY), 2021, pp. 119–124, https://doi.org/10.1109/SISY52375.2021.9582508.
[82] E. Ghaleb, M. Popa, S. Asteriadis, Multimodal and temporal perception of audio-visual cues for emotion recognition, in: 8th IEEE International Conference on Affective Computing and Intelligent Interaction (ACII), 2019, pp. 552–558, https://doi.org/10.1109/ACII.2019.8925444.
[83] L.N. Do, H.J. Yang, H.D. Nguyen, S.H. Kim, G.S. Lee, I.S. Na, Deep neural network-based fusion model for emotion recognition using visual data, J Supercomputing 77 (2021) 10773–10790, https://doi.org/10.1007/s11227-021-03690-y.
[84] D. Gera, S. Balasubramanian, Affect expression behaviour analysis in the wild using spatio-channel attention and complementary context information, ArXiv abs/2009.14440 (2020) 1–8.
[85] R.R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, D. Batra, Grad-CAM: Visual explanations from deep networks via gradient-based localization, in: IEEE International Conference on Computer Vision, 2017, pp. 618–626, https://doi.org/10.1109/ICCV.2017.74.
[86] M. Gogate, A. Adeel, A. Hussain, A novel brain-inspired compression-based optimised multimodal fusion for emotion recognition, in: IEEE Symposium Series on Computational Intelligence (SSCI), 2017, pp. 1–7, https://doi.org/10.1109/SSCI.2017.8285377.
[87] S. Yoon, S. Dey, H. Lee, K. Jung, Attentive modality hopping mechanism for speech emotion recognition, in: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2020, pp. 3362–3366, https://doi.org/10.1109/ICASSP40776.2020.9054229.

Elena Ryumina received the M.E. degree from the ITMO University in 2021. She is currently a joint Ph.D. student at ITMO University and works as the junior researcher at the St. Petersburg Institute for Informatics and Automation of the Russian Academy of Sciences, St. Petersburg Federal Research Center of the Russian Academy of Sciences (SPCRAS) in the Speech and Multimodal Interfaces Laboratory. Her main research focus is affective computing, audiovisual emotion recognition, computational paralinguistics, machine learning, neural networks, human-machine interfaces.

Dresvyanskiy Denis received his BE and MS degrees in system analysis from Reshetnev Siberian State University of Science and Technology in 2017 and 2019, respectively. He is currently a joint Ph.D. student at Ulm University and ITMO University. He has wide research interests mainly including human-computer interaction, signal processing, computer vision, deep and transfer learning, and paralinguistics analysis. Particularly, his study lies in the evaluation and analysis of conversational characteristics such as engagement, dominance, and affective states of interlocutors.

Karpov Alexey is Doctor of Technical Sciences (2013), Professor. Currently he works as the chief researcher and head of the Speech and Multimodal Interfaces Laboratory at the St. Petersburg Institute for Informatics and Automation of the Russian Academy of Sciences of the St. Petersburg Federal Research Center of the Russian Academy of Sciences (SPC RAS). He also works as a Full Professor (part-time) at the ITMO University. His research interests are speech technology, automatic speech recognition, audio-visual speech processing, multimodal human-computer interfaces, and computational paralinguistics. He is a general chair of International Conferences on Speech and Computer (SPECOM).