Machine Vision and Applications (2019) 30:975–985
https://doi.org/10.1007/s00138-018-0960-9

SPECIAL ISSUE PAPER

Audiovisual emotion recognition in wild

Egils Avots[1] · Tomasz Sapiński[3] · Maie Bachmann[2] · Dorota Kamińska[3]

Received: 6 April 2018 / Revised: 2 July 2018 / Accepted: 9 July 2018 / Published online: 19 July 2018

Abstract
People express emotions through different modalities. Utilization of both verbal and nonverbal communication channels allows to create a system in which the emotional state is expressed more clearly and therefore easier to understand. Expanding the focus to several expression forms can facilitate research on emotion recognition as well as human–machine interaction. This article presents analysis of audiovisual information to recognize human emotions. A cross-corpus evaluation is done using three different databases as the training set (SAVEE, eNTERFACE'05 and RML) and AFEW (database simulating real-world conditions) as a testing set. Emotional speech is represented by commonly known audio and spectral features as well as MFCC coefficients. The SVM algorithm has been used for classification. In case of facial expression, faces in key frames are found using Viola–Jones face recognition algorithm and facial image emotion classification done by CNN (AlexNet). Multimodal emotion recognition is based on decision-level fusion. The performance of emotion recognition algorithm is compared with the validation of human decision makers.

Keywords Emotion recognition · Audio signal processing · Facial expression · Deep learning

1 Introduction

During conversation, people are constantly sending and receiving nonverbal cues, communicated through voice (para-language), body movement, facial expressions and physiological changes. The difference between the words people speak and recognizing their actual meaning comes from nonverbal communication. Understanding them enhances interaction. The ability to recognize the attitude and thoughts from ones behavior was the original system of communication preceding speech. A particular emotional state is based on verbal and nonverbal signals. Therefore, emotions are a carrier of information regarding feelings of an individual and ones expected feedback.

Although computers are now a part of human life, the relationship between a human and a machine is limited. Knowledge of the emotional state of the user would allow the machine to adapt better and generally improve cooperation between them. Emotion recognition process leverages techniques from multiple areas, such as psychology, signal processing and machine learning. Moreover, this process may utilize various input types, i.e., facial expressions [1–5], speech [6–10], gestures and body language [11–15] and physical signals such as electroencephalography (EEG) [16], electromyography (EMG) [17], electrodermal activity [18]. However, facial expressions have been studied most extensively. About 95% of literature dedicated to this topic focuses on faces as a source, at the expense of other modalities [19]. This kind of system uses a facial expression in an image as an input and returns the confidence across a set of emotions, usually consisting of anger, disgust, fear, happiness, neutral, sadness and surprise. These emotions, according to

This work is supported by the Estonian Centre of Excellence in IT (EXCITE) funded by the European Regional Development Fund. The authors also gratefully acknowledge the support of NVIDIA Corporation with the donation of the Titan XP Pascal GPU.

✉ Dorota Kamińska
dorota.kaminska@p.lodz.pl

Egils Avots
egils.avots@ut.ee

Tomasz Sapiński
tomasz.sapinski@p.lodz.pl

Maie Bachmann
maie@cb.ttu.ee

[1] Institute of Technology, University of Tartu, Tartu, Estonia

[2] Department of Health Technologies, School of Information Technologies, Tallinn University of Technology, Tallinn, Estonia

[3] Institute of Mechatronics and Information Systems, Łodz University of Technology, Lodz, Poland

Paul Ekman, are cross-culturally and universally communicated with particular facial expressions [5]. Due to the fact that speech is one of the most accessible form from the above-mentioned signals and does not require direct contact with a human body, emotion recognition based on voice prosodic parameters became a relevant trend in modern studies.

Human emotion recognition may be useful in various commercial domains such as medicine [20], job interviews [21], education [22,23], entertainment [24], robotics [25,26], or even monitoring agents in call centers [27]. However, despite good recognition performance reported in laboratory conditions, real-life applications still remain an open challenge [28–30]. Most of the existing approaches are tailored toward specific databases, which could be one of the main factors making this task tough to solve. While the system is trained on a particular database, it faces the issues of different subjects, their ethnicity, appearance, culture, sex and age [31], contextual meaning of sentences, and the background noise [32]. Consequently, the algorithm does not work well when dealing with natural environment [33].

This paper highlights how challenging the task of recognizing emotional states in natural environment might be. We analyze the use of audiovisual information to recognize human emotions in the wild, presenting how models trained on specific database deal with samples from different corpora. Our testing set consists of emotional samples extracted from movies—AFEW corpora. This database is one of the most challenging due to a large number of different speakers, differing sample quality, background noise, overlaying of sounds produced by different speakers, irrelevant face positions (e.g., partially covered) and high variety of emotional displays. Hence, it is very close to real-world environment and simulates real scenarios. Obtained results are dramatically low; they are close to guessing. This analysis is presented in order to draw attention to above-mentioned issues common in real environments. In addition, the performance of emotion recognition algorithm is compared with the validation of human decision makers.

The paper adopts the following outline: In Sect. 2, related work is reviewed. The proposed method is introduced in Sect. 3. Then, extensive experiments are described in Sect. 4. Finally, in Sect. 5, the paper concludes through providing a summary, followed by hints to possible subjects of future studies.

2 Related works

2.1 Audio–video emotion corpora

Emotional databases can be divided into three categories, taking into account their source: spontaneous, invoked and acted or simulated emotions. First type of samples can be obtained by recording in natural situations, or using movies, TV programs such as talk shows, reality shows or various types of live coverage. This type of material might not be of satisfactory quality due to background noise, artifacts, overlapping voices, etc., which may obscure the exact nature of recorded emotions. In addition, such recordings usually do not provide frontal-view facial expressions which are crucial in emotion recognition research. Moreover, collections of samples must be evaluated by human decision makers or specialists to determine the recorded emotional states. A very good example of such database is the Belfast Naturalistic Database [34], which contains 298 audiovisual samples from 125 speakers (31 males and 94 females). The main sources of those samples are talk shows and religious programs, which provided a strong emotional material, both positive and negative. The data are labeled with dimensional and categorical approaches using Feeltrace system.

Different approaches for creating this kind of database are presented in [35]. LIRIS-ACCEDE databases is composed of 9800 video excerpts (each 8–12 s long) extracted from 160 movies shared under Creative Commons licenses. Video clips are sorted along the induced valence axis, from the video representing the most negative state to the most positive. The classification was carried out by 1517 volunteers from 89 different countries.

CASIA Natural Emotional Audio-Visual Database [36] is a spontaneous, audiovisual, rich-annotated emotion database which contains two hours of spontaneous emotional segments extracted from movies, TV plays and talk shows. This database provides 219 different speakers, 52.5% male speakers, 47.5% female. Samples were labeled by three Chinese native speakers into 24 non-prototypical emotional states.

Another method of sample acquisition is provoking an emotional reaction using staged situations. Appropriate states are induced using imaging methods (videos, images), stories, or computer games. This type of recordings is preferred by psychologists, although the method cannot provide desirable effects as reaction to the same stimuli may differ. Moreover, provoking strong emotions might be ethically problematic. Similarly to spontaneous recordings, triggered emotional samples should be subjected to a process of evaluation.

An example of such corpora is the eNTERFACE'05 Audio-Visual Emotion Database [37], which consists of 1166 video sequence presented by 42 subjects (coming from 14 different nationalities, 81% men and 19% women). Each subject was asked to listen to six different stories eliciting particular emotional states: anger, disgust, fear, happiness, sadness and surprise. After that, the subject read out five utterances in English, which constitute five different reactions to the given situation. All samples were assessed by two human experts.

Fig. 1 Selected samples from audio–video emotional databases: (1) Belfast Naturalistic Database [34], (2) GEMP [41], (3) LIRIS-ACCEDE [35], (4) SAVEE [40], (5) RML [38], (6) eNTERFACE'05 [37]

The database collected at Ryerson Multimedia Lab (RML) [38] contains 720 audiovisual samples portraying six basic emotions: anger, disgust, fear, happiness, sadness and surprise. Samples were recorded in a quiet and bright environment, with a simple background. The subjects were provided with a list of emotional sentences and were directed to express their feeling by recalling the emotional happening experienced in their lives. The database is language and cultural independent; samples were collected from subjects speaking six different languages such as English, Mandarin, Urdu, Punjabi, Persian, Italian).

The Adult Attachment Interview (AAI) database, created by Roisman [39], is another example of natural audiovisual database. It consists of recordings from 60 adults acquired during the interview on which each subject was describing their childhood experiences for 30–60 min. The data are labeled using Facial Action Coding System (FACS) into six basic emotions with the addition of embarrassment, contempt, shame, and general positive and negative states.

The third source is acted out emotional samples. Subjects can be both actors and unqualified volunteers. This type of material is usually composed of high-quality recordings, with clear undisturbed emotion expression. A good example of such corpora is Surrey Audio-Visual Expressed Emotion (SAVEE) [40]. This British English database contains high-quality video recordings performed by 4 actors speaking utterances (120 per actor, 480 in total) with 7 various emotions: anger, disgust, fear, happiness, sadness, surprise and neutral. The data have been validated by 10 participants under audio, visual and audiovisual conditions.

GEMEP—The Geneva Multi-modal Emotion Portrayals database [41]—is a dynamic multimodal corpus, which consists of more than 7000 audio–video samples, representing 18 emotions. Besides the emotional labels which are commonly known and used in these types of corpora, it includes rarely studied subtle emotions such as despair, anxiety, amusement, interest or pleasure. Emotional states are portrayed by 10 professional actors, coached by a professional director.

Another example, Busso–Narayanan acted database [42], consists of recordings from an actress, who is asked to read a phoneme-balanced corpus four times, expressing sequentially anger, happiness, sadness and neutral state. A detailed description of the actress facial expression and rigid head motion is acquired by attaching 102 markers to her face. To capture the 3D position of each marker, VICON motion capture system was used. The total data consist of 612 sentences (Fig. 1).

2.2 Algorithms and models

Automatic affect recognition is a pattern recognition problem. Therefore, a standard methodology involving feature extraction and classification is usually applied. Recent work in automatic affect recognition field combines both facial expressions and acoustic information to improve recognition performance of such systems. Thus, in the majority of scientific papers two different approaches are most commonly used: feature-level fusion with single classifier and decision-level fusion with separate classifiers for each modality (see Fig. 2).

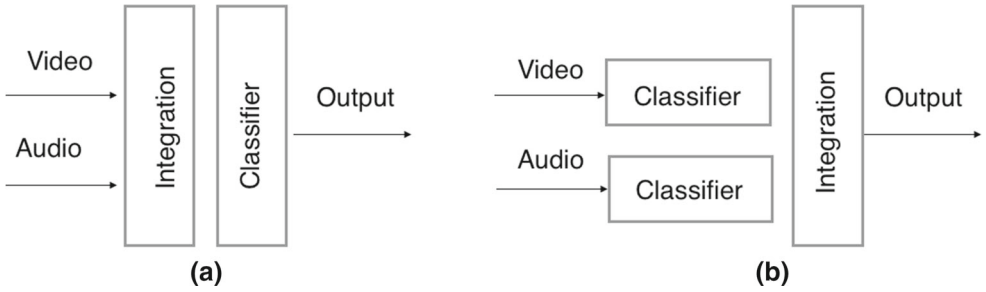

Fig. 2 Two different simplified models for multimodal emotion recognition: **a** feature-level fusion and **b** decision-level fusion

The first approach is performed by combining the audio and visual features into a single vector. This method may be supplemented with feature selection from individual modalities, either before or after combining them. For example in [43], the authors propose an audio–video emotion recognition system using convolutional neural network (CNN) to extract features from speech, combined with a deep residual network of 50 layers to extract features from video. The speech network extracts 1280-dimensional features, and the visual network extracts 2048-dimensional features, which are ultimately concatenated to form a 3328-dimensional feature vector and fed to a 2-layer recurrent network LSTM with 256 cells each. The experimental results, for prediction of arousal and valence, show that proposed models achieve significantly better performance on the test set in comparison with other models using the same database [44].

Another example of such approach is presented in [45]. The authors propose a novel Bayesian nonparametric multimodal data modeling framework using features extracted from key frames of video via convolutional neural networks (CNNs) and Mel-frequency cepstral coefficients (MFCC) features. Then, a symmetric correspondence hierarchical Dirichlet processes (Sym-cHDP) are utilized to model the multimodal data and furthermore learn the latent emotional correlations between image data and audio data. Achieved recognition rate of this method outperforms others significantly.

In the second approach, features from different modalities are processed independently and individual recognition results are combined at decision level. For instance, in [46] the authors propose a novel approach, which in addition to audio processing captures speech-related characteristics in both the upper and lower face regions. In order to create vector representations of confidence (emotional profiles—EPs) for the presence or absence of emotional expression (anger, happiness, neutrality and sadness), they use three different modalities: upper face, lower face and speech. Upper and lower face EPs are computed based on time-series similarities. Based on emotion class distribution of the k-closest training segments, the testing EPs are computed, after calculating similarity between training and testing segments. For creating speech-based EPs, the outputs of binary support vector machines (SVM) are used. The EPs calculated from the three modalities are averaged to obtain the final emotion label. The framework was tested on the IEMOCAP and SAVEE datasets, achieving a performance of 67.22 and 86.01%, respectively.

A promising audio–video emotion recognition system based on the fusion of several models is presented in [26]. The authors use five separate classifiers: three multiclass SVM for audio, left and mono audio channels, one SVM for geometric visual features and one CNN model considering as input the computed key frames. As in the previous example, the outputs of those classifiers are collected in the form of vector representations of the confidence (margin for SVM and probability for CNN) for all possible emotion labels. Finally, confidence outputs are used to define a new feature space to be learned for final emotion label prediction. The experiments are conducted on three different datasets: SAVEE, eNTERFACE'05 and RML. According to the authors, obtained results show significant performance improvements in comparison with state-of-the-art methods tested on above-mentioned three databases. (For example, the recognition rate for SAVEE was 99.72, 13.71% more than in the aforementioned example.)

In [42], Busso et al. compare separate classifiers based on acoustic and facial data with both types of fusions—on decision and feature levels. Using Busso–Narayanan acted database, four emotions are classified: sadness, anger, happiness, and neutral state. Separate classifiers based on acoustic data and facial expressions obtain accuracy performance of 70.9 and 85% respectively. Combination of audio and facial data on feature level improves the recognition rate to 90%. On the decision level, several criteria of integration are compared: maximum average, product and weight. The accuracy of decision-level integrated bimodal classifiers range from 84% to 89%, with the product integration criterion as the most efficient one. Similar conclusions are presented in [47].

As one can easily observe, the cross-corpus evaluation approach is still lacking in the state-of-the-art scientific papers. Usually, the efficiency of classifiers is measured on specific corpora, using cross-validation or by splitting fea-

Fig. 3 Selected samples from AFEW database

tures set into training and testing sets. These types of results can be overstated due to high similarity of both sets. Samples forming particular database are collected from similar sources (the same TV shows) or recorded under the same conditions. Unfortunately, one is not able to predict how specific classifier behaves in totally different conditions, how background fluctuation affects the quality of recognition. This element is crucial in real-world conditions, where there is no reproducibility. Papers using cross-corpus evaluation in speech- based emotions recognition [48–51] confirm the above concerns by indicating a significant decrease in recognition rate using this type of evaluation. Only a few studies present such approach while investigating multimodal input [52], and there is still a big demand for more comprehensive studies of this issue.

3 Methodology

3.1 Datasets description

One of the main goals of this paper is to present cross-corpus evaluation using three different type of database as a training set and database simulating real-world conditions as a testing set. For the purpose of our experiment, we decided to use SAVEE, eNTERFACE'05 and RML to train the classifier. All of them are described in Sect. 2.

As a testing set, we use Acted Facial Expressions in the Wild's (AFEW) [53], an acted facial expressions dataset in tough conditions (close to real-world environments unlike most other databases recorded in a laboratory environment). It consists of 957 videos labeled with six basic emotional states: angry, happy, disgust, fear, sad, surprise and the neutral state. The subjects (actors) belong to a wide range of ages from 1 to 70 years. The clips have various scenarios such as indoor, outdoor, nighttime, gathering of several people. Most of the samples contain background noise, which are much closer to real-world environment than laboratory-controlled conditions. Figure 3 presents selected samples from AFEW database. Due to the purpose of creating this database (facial expressions recognition), we had to remove several samples which were not suitable for our examination.

A perception test was carried out with 12 subjects (6 males and 6 females), to determine how the AFEW samples are perceived by humans. The two modalities were presented separately at the end simultaneously. They were allowed to watch or listen to each sample only once and then determine the presented emotional state. Each volunteer had to assess 36 random samples. The results are presented in Fig. 4.

Analyzing the chart one can observe that higher recognition rate occurred for facial emotion expressions. Significantly lower results were obtained in case of speech. Presenting two modalities simultaneously provided an increase in recognition performance. The average recognition rate in this case was above 69.04%.

3.2 Emotion recognition by speech

Assuming that an audio sample is given in a digital format, the vocal emotion recognition system consists of the following steps: feature extraction, dataset generation and classification. The extracted features represent non-linguistic properties of the audio signal. The speech recognition system was built using 21 audio features and support vector machine (SVM) setup. Firstly, the audio features are extracted from a mono channel. Table 1 presents a list of audio and spectral features estimated for the propose of this project.

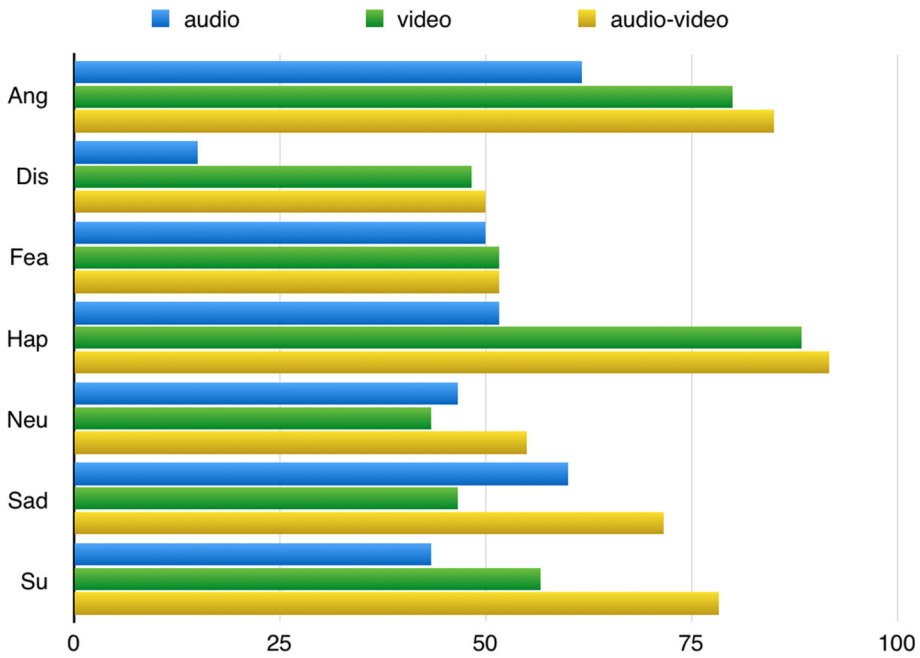

Fig. 4 Average recognition rates in % for two modalities presented and evaluated separately and simultaneously by 12 humans

Table 1 Features extracted from speech signal selected for the purpose of this research

Audio features	Spectral features
Energy entropy	Tonal power ratio
Short-time energy	Spread
Zero-crossing rate	Slope
Spectral roll-off	Skewness
Spectral centroid	Roll-off
Harmonic product spectrum (HPS)	Kurtosis
Pitch time ACF	Flux
Pitch time average magnitude difference function (AMDF)	Flatness
	Decrease
	Crest
	Centroid

In addition, we extract Mel-frequency cepstral coefficients (MFCCs), which are calculated for 400 ms sliding window with step size of 200 ms. A single audio file will have several windows with MFCC. For one audio file, we have several feature vectors, where one part of the vector represents general information about the sound sample and the second part, coefficients for a specific sliding window. One feature vector consists of 34 parameters, first 21 represent the global audio features and the remaining 13 coefficients represent the local MFCC. For MFCC feature extraction, we used the following setup: preemphasis coefficient 0.97, 20 filter bank channels, 13 cepstral coefficients, 300 Hz lower frequency limit and 3700 Hz upper frequency limit. To obtain a single prediction for a audio sample, we merged the results based on majority vote (Fig. 5).

The testing set can be considered as the most challenging dataset, because the audio clips can have other sounds in the background. As RML and eNTERFACE'05 do not have a neutral class, it is not included in training and testing sets when comparing results between databases. Testing data consists of SAVEE, RML and eNTERFACE'05 databases. The training set consists of AFEW data.

3.3 Emotion recognition by facial expressions

Facial expression consists of video preprocessing and use of Convolutional Neural Network (AlexNet) [54] for facial image classification. The videos have to be divided into separate frames which are the main source for visual-based features. Videos in the controlled databases have a fixed setup; nevertheless, it is advantageous to only focus on the face as it is the area where emotions are expressed. In preprocessing phase, we extract select frames also known as key frames from each video. This is done to avoid training the system with images where facial expression remains the same. Usually this happens at the start and end of the controlled videos, when the subject is preparing to express the emotion from a neutral state and return to neutral state after the emotion is demonstrated. The numerical frame difference is expressed as sum of absolute difference between pixels. Therefore, an image pair provides a score of similarity, for frames, that are exactly the same the score is 0. To skip frames automatically, the system averages the difference values for the last 10 processed frames, if the new frame has

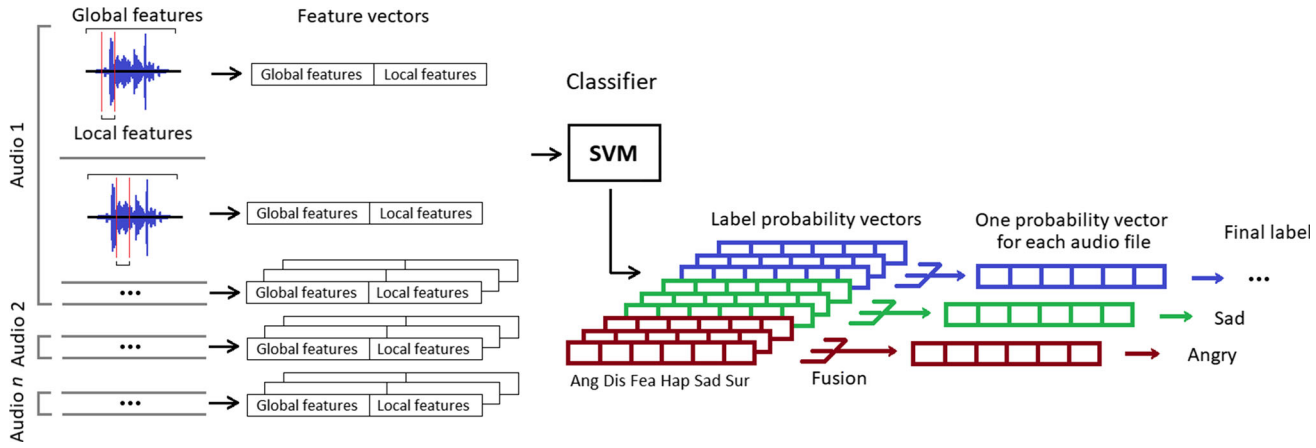

Fig. 5 Model for audio-based emotion recognition

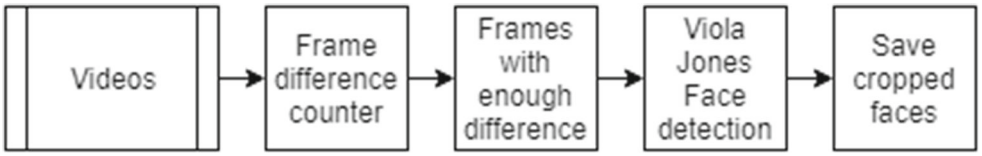

Fig. 6 Key frame extraction from video samples

difference value that is less than average value*1.5, the frame is skipped. (The value 1.5 was found empirically.) For frames where high enough difference can be observed, we applied Viola–Jones face detection to crop out the face region and save it as an image. A general pipeline can be seen in Fig. 6.

This approach worked very well for the controlled databases, the testing dataset had few wrongly classified faces, which were removed manually.

Afterward, the extracted facial images are labeled according to their respective emotion. The network was trained in Convolution Neural Network (AlexNet). We use transfer learning approach where a pretrained network is used as a starting point to learn a new task. Fine-tuning a network with transfer learning is usually much faster and easier than training a network with randomly initialized weights from scratch. To ensure that the CNN learns general features, the images are randomly translated in X and Y directions in range of −30 to 30 pixels. The presented results are obtained by using default MATLAB configuration for transferred learning, as this paper does not focus on CNN tuning. We used the recommended setup based on MATLAB documentation, most important parameters can be found in the brackets (Weight Learn Rate Factor = 20, Bias Learn Rate Factor = 20, Mini Batch Size = 10, Max Epochs = 10, Initial Learn Rate = 1e-4, Validation Frequency = 3, Validation Patience = Inf).

In order to compare the results with human participants, the frame- based prediction has to be transformed to a video-based prediction. If a video has more than one key frame, the final label is determined by majority voting. The full process can be seen in Fig. 7.

3.4 Fusion-based emotion recognition

The above-presented approaches refer to prediction for a single frame and 400 ms audio segments. In order to compare the results with human participants, the frame-based prediction was transformed to a video-based prediction (Fig. 7). Similarly, as the audio samples are also separated in small segments, the results were merged to get a single prediction for a audio file (Fig. 5). The process for obtaining single prediction was done as follows: each audio and video prediction has six score values which correspond to predicted accuracy for a specific class. The sum of all probabilities is 1, and the highest value represents the predicted label. The respective probabilities are summed together and normalized to get a final prediction. The previously mentioned procedure is necessary to obtain a single prediction for a sample file (audio + video) in a comparable manner. As audio- and video-predicted label is based on the six probabilities, they can be directly compared, which refers to decision-level fusion. The fusion results are presented for AFEW database.

4 Experimental results and discussion

To evaluate the performance of each database separately, we separated the data into training data and testing data with

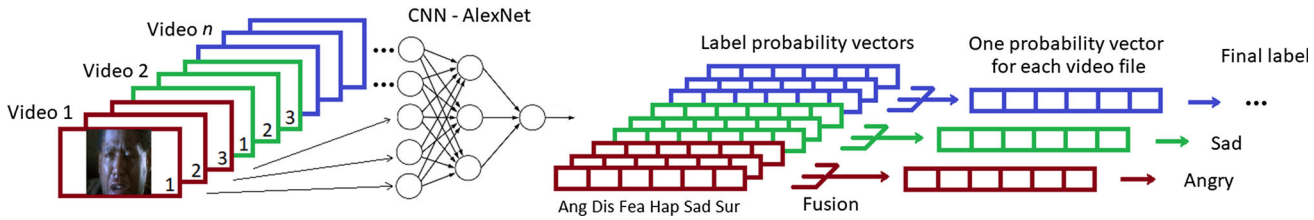

Fig. 7 Model for facial images-based emotion recognition

Table 2 SAVEE audio dataset with 650 features in test set

	Ang	Dis	Fea	Hap	Sad	Sur	RR %
Ang	**77**	1	7	1	0	4	85.6
Dis	3	**100**	2	5	12	0	81.9
Fea	1	1	**61**	6	2	11	74.4
Hap	7	4	11	**82**	0	8	73.2
Sad	0	11	1	0	**103**	0	89.6
Sur	4	6	16	10	0	**93**	72.1

Samples recognised correctly are shown in bold

Table 3 RML audio dataset with 1725 features in test set

	Ang	Dis	Fea	Hap	Sad	Sur	RR %
Ang	**218**	0	19	11	0	37	76.5
Dis	3	**221**	22	27	6	9	76.7
Fea	7	30	**142**	38	13	24	55.9
Hap	1	31	27	**156**	31	20	58.6
Sad	0	22	45	18	**234**	10	71.1
Sur	43	7	29	24	5	**195**	64.4

Samples recognised correctly are shown in bold

ratio 9:1. To get an estimate of the average performance, the classification and prediction were performed 100 times. The following data refer to average accuracy of individual databases. The highest accuracy, 77.4% was obtained for SAVEE dataset (Table 2), followed by 69.3% with RML dataset (Table 3), eNTERFACE'05 with 50.2% (Table 4) and finally AFEW with only 46.6% accuracy (Table 5). When the RML, SAVEE and eNTERFACE'05 are merged together to create a training set and AFEW is used as testing set, the accuracy of such system is 27.1% (Table 6), which is only slightly better than chance for predictions based on six classes.

For vision-based systems where the CNN was trained based on key frames, the prediction for key frames in AFEW database was 23.8% which is close to guessing. To test, if the learned model is correct, we performed analysis where only RML, SAVEE and eNTERFACE'05 are evaluated separately by splitting the data to 9:1 ratio for training and testing sets. The prediction of individual datasets can be seen in Table 7.

The results show that the trained CNN model has learned features that represent the datasets themselves and not fea-

Table 4 eNTERFACE'05 audio dataset with 1570 features in test set

	Ang	Dis	Fea	Hap	Sad	Sur	RR %
Ang	**223**	77	33	36	36	35	50.7
Dis	11	**90**	17	33	11	16	50.6
Fea	10	20	**93**	8	24	23	52.2
Hap	12	38	17	**112**	4	27	53.3
Sad	16	20	45	16	**186**	44	56.9
Sur	19	29	40	16	39	**94**	39.7

Samples recognised correctly are shown in bold

Table 5 AFEW audio dataset with 522 features in test set

	Ang	Dis	Fea	Hap	Sad	Sur	RR %
Ang	**73**	15	11	15	19	8	51.8
Dis	9	**15**	10	8	1	1	34.1
Fea	4	4	**10**	6	5	1	33.3
Hap	27	39	15	**87**	14	13	44.6
Sad	11	15	6	12	**58**	6	53.7
Sur	0	0	0	2	1	**1**	25.0

Samples recognised correctly are shown in bold

Table 6 Testing AFEW audio dataset (2609 features) with using RML, SAVEE and eNTERFACE'05 as training data

	Ang	Dis	Fea	Hap	Sad	Sur	RR %
Ang	**296**	75	64	99	168	43	39.7
Dis	120	**83**	64	98	108	17	16.9
Fea	18	17	**21**	25	11	2	22.3
Hap	86	57	29	**133**	53	40	33.4
Sad	73	181	78	238	**162**	54	20.6
Sur	21	10	3	24	26	**12**	12.5

Table 7 Accuracy of frame-based emotion recognition obtained using CNN

Database	Accuracy %
SAVEE	94.33
RML	60.20
eNTERFACE	48.31
AFEW	94.68

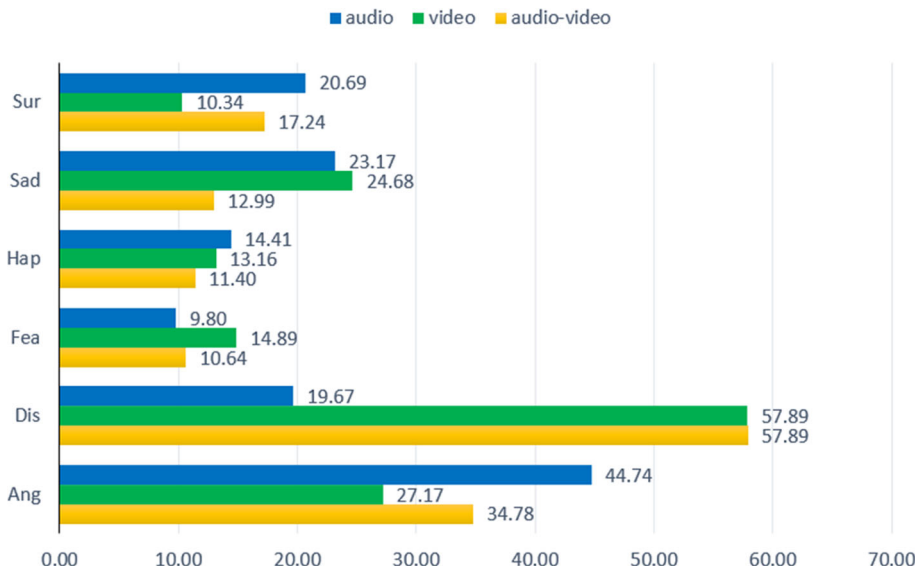

Fig. 8 Average recognition rates of decision-level fusion in %

tures that would represent emotions for a general solution. Also, it has to be pointed out that AFEW contains very expressive emotions displays, for example, yelling, screaming, crying. Such cases are not found in any of the RML, SAVEE and eNTERFACE'05 datasets.

Fusion of audio and video results at decision level, which was performed as described in Sect. 3.4, resulted in accuracy of 22.4%. When RML, SAVEE and eNTERFACE'05 are used for training and AFEW for testing, the results can be seen in Fig. 8. The results from both modalities do not compliment each other; therefore, the fused results are worse than individual results. Feature-level fusion would need frame and audio sample synchronization to create appropriate feature vectors, the current system treats audio segments and facial images in asynchronous manner; therefore, it is not suited for feature-level fusion.

5 Conclusion

In this paper, we have demonstrated the use of common audio features for emotion recognition and how such features perform on RML, SAVEE, eNTERFACE'05 and AFEW datasets, which is the most challenging dataset with high variety of emotional displays. The training and testing data used the six basic emotions and excluded *neutral*, as it is not present in all of the mentioned datasets. Also, we demonstrate that merging datasets will not necessarily improve the prediction accuracy, when testing set is a separate database, meaning that none of the samples from the separate database are used in training the classifier. The RML, SAVEE, eNTERFACE'05 datasets were merged together as a training set and evaluated against AFEW, to determine whether such approach increases the prediction accuracy. This experiment shows that when training classifiers, the classifiers work well within the dataset, but not necessarily provide a general solution, which can be applied to other datasets. This is clearly shown with speech part of the experiments. Similarly, the same can be said for visual emotion representation, merging different databases, where the training data are extracted key frames from videos, will produce classification accuracy which is close to chance when a part of testing database is not included in the training set.

One can observe that using a single database for both training and testing sets results in a significantly higher recognition results (Tables 2, 3, 5) than a complete separation of the source of both sets (Table 6 and Fig. 8). The assumption that the same features extracted from any source should produce at least similar results when applied to those features extracted for a different corpus, proved not to be valid in this case. One of the main reasons for this might be the quality of samples in the AFEW database where the background interference might obstruct the emotional message. However, the outcome resulted in much lower recognition rate than expected. As it was mentioned before, the results are close to chance, even for a combined training set of three databases which on their own produce close to baseline recognition rates. This experiment highlights how challenging the task of recognizing an emotional states in a natural environment might be, specially when compared to human recognition rate (Fig. 4), which in case of two modalities was higher than 50%. This is a result of the natural capability of the human brain to filter out the unnecessary information and focus solely on the emotional content of the samples.

In future work, we plan to look at more databases to determine which ones of them provides the most general features, when using common feature extraction and classification methods and discusses the potential flaws in benchmark databases. In addition, we plan to extend the data preparation

as to allow for feature-level fusion, where features from one audio sample and facial image are represented in one feature vector.

References

1. El Kaliouby, R., Robinson, P.: 2004 IEEE International Conference on Systems, Man and Cybernetics (IEEE, 2004), vol. 1, pp. 682–688 (2004)
2. Ofodile, I., Kulkarni, K., Corneanu, C.A., Escalera, S., Baro, X., Hyniewska, S., Allik, J., Anbarjafari, J.: Automatic recognition of deceptive facial expressions of emotion. arXiv:1707.04061 (2017)
3. Shojaeilangari, S., Yau, W.Y., Teoh, E.K.: Pose-invariant descriptor for facial emotion recognition. Mach. Vis. Appl. **27**(7), 1063 (2016)
4. Loob, C., Rasti, P., Lüsi, I., Junior, J.C.J., Baró, X., Escalera, S., Sapinski, T., Kaminska, D., Anbarjafari, G.: 2017 12th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2017) (IEEE, 2017), pp. 833–838 (2017)
5. Ekman, P., Friesen, W.V.: Facial action coding system (1977)
6. Kamińska, D., Pelikant, A.: Recognition of human emotion from a speech signal based on plutchik's model. Int. J. Electron. Telecommun. **58**(2), 165 (2012)
7. Noroozi, F., Sapiński, T., Kamińska, D., Anbarjafari, G.: Vocal-based emotion recognition using random forests and decision tree. Int. J. Speech Technol. **20**(2), 239 (2017)
8. Kamiska, D., Sapiski, T., Anbarjafari, G.: Efficiency of chosen speech descriptors in relation to emotion recognition. EURASIP J. Audio Speech Music Process. **2017**(1), 3 (2017)
9. Clavel, C., Vasilescu, I., Devillers, L., Richard, G., Ehrette, T.: Fear-type emotion recognition for future audio-based surveillance systems. Speech Commun. **50**(6), 487 (2008)
10. Noroozi, F., Marjanovic, M., Njegus, A., Escalera, S., Anbarjafari, G.: 2016 23rd International Conference on Pattern Recognition (ICPR) (IEEE, 2016), pp. 61–66 (2016)
11. Pławiak, P., Sośnicki, T., Niedźwiecki, M., Tabor, Z., Rzecki, K.: Hand body language gesture recognition based on signals from specialized glove and machine learning algorithms. IEEE Trans. Indus. Inf. **12**(3), 1104 (2016)
12. Kiforenko, L., Kraft, D.: 11th International Conference on Computer Vision Theory and Applications Computer Vision Theory and Applications (SCITEPRESS Digital Library, 2016), pp. 398–405 (2016)
13. Kleinsmith, A., Bianchi-Berthouze, N.: Affective body expression perception and recognition: a survey. IEEE Trans. Affect. Comput. **4**(1), 15 (2013)
14. Noroozi, F., Corneanu, C.A., Kamińska, D., Sapiński, T., Escalera, S., Anbarjafari, G.: Survey on emotional body gesture recognition. arXiv:1801.07481 (2018)
15. Haamer, R.E., Kulkarni, K., Imanpour, N., Haque, M.A., Avots, E., Breisch, M., Nasrollahi, K., Guerrero, S.E., Ozcinar, C., Baro, X., et al.: IEEE Conference on Automatic Face and Gesture Recognition Workshops (IEEE, 2018) (2018)
16. Jenke, R., Peer, A., Buss, M.: Feature extraction and selection for emotion recognition from eeg. IEEE Trans. Affect. Comput. **5**(3), 327 (2014)
17. Jerritta, S., Murugappan, M., Wan, K., Yaacob, S.: Emotion recognition from facial emg signals using higher order statistics and principal component analysis. J. Chin. Inst. Eng. **37**(3), 385 (2014)
18. Greco, A., Valenza, G., Citi, L., Scilingo, E.P.: Arousal and valence recognition of affective sounds based on electrodermal activity. IEEE Sens. J. **17**(3), 716 (2017)
19. Gelder, B.D.: Why bodies? twelve reasons for including bodily expressions in affective neuroscience. Philos. Trans. R. Soc. B Biol. Sci. **364**, 3475–3484 (2009). https://doi.org/10.1098/rstb.2009.0190
20. Tacconi, D., Mayora, O., Lukowicz, P., Arnrich, B., Setz, C., Troster, G., Haring, C.: Second International Conference on Pervasive Computing Technologies for Healthcare, 2008. PervasiveHealth 2008 (IEEE, 2008), pp. 100–102 (2008)
21. Gorbova, J., Lüsi, I., Litvin, A., Anbarjafari, G.: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp. 29–35 (2017)
22. Calvo, R.A., D'Mello, S.: Frontiers of affect-aware learning technologies. IEEE Intell. Syst. **27**(6), 86 (2012)
23. Noroozi, F., Akrami, N., Anbarjafari, G.: 2017 25th Signal Processing and Communications Applications Conference (SIU), (IEEE, 2017), pp. 1–4 (2017)
24. Schuller, B., Marchi, E., Baron-Cohen, S., Lassalle, A., OReilly, H., Pigat, D., Robinson, P., Davies, I., Baltrusaitis, T., Mahmoud, M., et al.: Proceedings of the of the 3rd International Workshop on Intelligent Digital Games for Empowerment and Inclusion (IDGEI 2015) as Part of the 20th ACM International Conference on Intelligent User Interfaces, IUI 2015, p. 9 (2015)
25. Marchi, E., Ringeval, F., Schuller, B.: Voice-enabled assistive robots for handling autism spectrum conditions: an examination of the role of prosody. In: Neustein, A. (ed.) Speech and Automata in the Health Care, pp. 207–236. Walter de Gruyter GmbH & Co KG, Berlin (2014)
26. Noroozi, F., Marjanovic, M., Njegus, A., Escalera, S., Anbarjafari, G.: Audio-visual emotion recognition in video clips. IEEE Trans. Affect. Comput. (2017). https://doi.org/10.1109/TAFFC.2017.2713783
27. Chakraborty, R., Pandharipande, M., Kopparapu, S.K.: Frontiers in Electronic Technologies, pp. 55–63. Springer, Berlin (2017)
28. Zhang, Z., Ringeval, F., Han, J., Deng, J., Marchi, E., Schuller, B.: 17th Annual Conference of the International Speech Communication Association (INTERSPEECH 2016), pp. 3593–3597 (2016)
29. Wan, J., Escalera, S., Baro, X., Escalante, H.J., Guyon, I., Madadi, M., Allik, J., Gorbova, J., Anbarjafari, G.: ChaLearn LaP, Action, Gesture, and Emotion Recognition Workshop and Competitions: Large Scale Multimodal Gesture Recognition and Real versus Fake expressed emotions, ICCV, vol. 4 (2017)
30. Lüsi, I., Junior, J.C.J., Gorbova, J., Baró, X., Escalera, S., Demirel, H., Allik, J., Ozcinar, C., Anbarjafari, G.: 2017 12th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2017) (IEEE, 2017), pp. 809–813 (2017)
31. Sagha, H., Matejka, P., Gavryukova, M., Povolnỳ, F., Marchi, E., Schuller, B.W.: INTERSPEECH, pp. 2949–2953 (2016)
32. Tawari, A., Trivedi, M.M.: 2010 20th International Conference on Pattern Recognition (ICPR), (IEEE, 2010), pp. 4605–4608 (2010)
33. Li, W., Tsangouri, C., Abtahi, F., Zhu, Z.: A recursive framework for expression recognition: From web images to deep models to game dataset. arXiv:1608.01647 (2016)
34. Douglas-Cowie, E., Campbell, N., Cowie, R., Roach, P.: Emotional speech: towards a new generation of databases. Speech Commun. **40**(1–2), 33 (2003)
35. Baveye, Y., Bettinelli, J.N., Dellandréa, E., Chen, L., Chamaret, C.: 2013 Humaine Association Conference on Affective Computing and Intelligent Interaction (ACII), (IEEE, 2013), pp. 13–18 (2013)

36. Bao, W., Li, Y., Gu, M., Yang, M., Li, H., Chao, L., Tao, J.: 2014 12th International Conference on Signal Processing (ICSP), (IEEE, 2014), pp. 583–587 (2014)
37. Martin, O., Kotsia, I., Macq, B., Pitas, I.: Proceedings of 22nd International Conference on Data Engineering Workshops, (IEEE, 2006), pp. 8–8 (2006)
38. RML emotion database. http://www.rml.ryerson.ca/rml-emotion-database.html. Accessed 30 Mar 2018
39. Roisman, G.I., Holland, A., Fortuna, K., Fraley, R.C., Clausell, E., Clarke, A.: The adult attachment interview and self-reports of attachment style: an empirical rapprochement. J. Pers. Soc. Psychol. **92**(4), 678 (2007)
40. Wang, W.: Machine Audition: Principles, Algorithms and Systems: Principles, Algorithms and Systems. IGI Global, Hershey (2010)
41. Glowinski, D., Camurri, A., Volpe, G., Dael, N., Scherer, K.: CVPRW (IEEE, 2008), pp. 1–6 (2008)
42. Busso, C., Deng, Z., Yildirim, S., Bulut, M., Lee, C.M., Kazemzadeh, A., Lee, S., Neumann, U., Narayanan, S.: Proceedings of the 6th International Conference on Multimodal Interfaces (ACM, 2004), pp. 205–211 (2004)
43. Tzirakis, P., Trigeorgis, G., Nicolaou, M.A., Schuller, B.W., Zafeiriou, S.: End-to-end multimodal emotion recognition using deep neural networks. IEEE J. Sel. Top. Signal Process. **11**(8), 1301 (2017)
44. Ringeval, F., Sonderegger, A., Sauer, J., Lalanne, D.: 2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG), (IEEE, 2013), pp. 1–8 (2013)
45. Xue, J., Luo, Z., Eguchi, K., Takiguchi, T., Omoto, T.: 2017 IEEE International Conference on Multimedia and Expo (ICME), (IEEE, 2017), pp. 601–606 (2017)
46. Kim, Y., Provost, E.M.: ISLA: Temporal segmentation and labeling for audio-visual emotion recognition. IEEE Trans. Affect. Comput. (2017). https://doi.org/10.1109/TAFFC.2017.2702653
47. Sidorov, M., Sopov, E., Ivanov, I., Minker, W.: 2015 12th International Conference on Informatics in Control, Automation and Robotics (ICINCO), (IEEE, 2015), vol. 2, pp. 246–251 (2015)
48. Song, P.: Transfer linear subspace learning for cross-corpus speech emotion recognition. IEEE Trans. Affect. Comput. (2017). https://doi.org/10.1109/TAFFC.2017.2705696
49. Song, P., Zheng, W., Ou, S., Zhang, X., Jin, Y., Liu, J., Yu, Y.: Cross-corpus speech emotion recognition based on transfer non-negative matrix factorization. Speech Commun. **83**, 34 (2016)
50. Schuller, B., Vlasenko, B., Eyben, F., Wollmer, M., Stuhlsatz, A., Wendemuth, A., Rigoll, G.: Cross-corpus acoustic emotion recognition: variances and strategies. IEEE Trans. Affect. Comput. **1**(2), 119 (2010)
51. Zhang, Z., Weninger, F., Wöllmer, M., Schuller, B.: 2011 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU), (IEEE, 2011), pp. 523–528 (2011)
52. Chang, C.M., Su, B.H., Lin, S.C., Li, J.L., Lee, C.C.: 2017 Seventh International Conference on Affective Computing and Intelligent Interaction (ACII), (IEEE, 2017), pp. 377–382 (2017)
53. Dhall, A., Goecke, R., Lucey, S., Gedeon, T.: Acted facial expressions in the wild database. Australian National University, Canberra, Australia, Technical Report TR-CS-11, vol. 2, p. 1 (2011)
54. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. Advances in Neural Information Processing Systems, pp. 1097–1105 (2012)