Audio-Visual Emotion Recognition in Video Clips

Fatemeh Noroozi, *Student Member, IEEE*, Marina Marjanovic, *Member, IEEE*, Angelina Njegus, Sergio Escalera, and Gholamreza Anbarjafari, *Senior Member, IEEE*

Abstract—This paper presents a multimodal emotion recognition system, which is based on the analysis of audio and visual cues. From the audio channel, Mel-Frequency Cepstral Coefficients, Filter Bank Energies and prosodic features are extracted. For the visual part, two strategies are considered. First, facial landmarks' geometric relations, i.e., distances and angles, are computed. Second, we summarize each emotional video into a reduced set of key-frames, which are taught to visually discriminate between the emotions. In order to do so, a convolutional neural network is applied to key-frames summarizing videos. Finally, confidence outputs of all the classifiers from all the modalities are used to define a new feature space to be learned for final emotion label prediction, in a late fusion/stacking fashion. The experiments conducted on the SAVEE, eNTERFACE'05, and RML databases show significant performance improvements by our proposed system in comparison to current alternatives, defining the current state-of-the-art in all three databases.

Index Terms—Multimodal emotion recognition, classifier fusion, data fusion, convolutional neural networks

1 Introduction

Detecting human intentions and feelings is among the important factors for enabling robots to communicate with humans. The recognized emotional mood will be considered for determining the proper robot/machine/computer reaction [1], [2], [3], [4]. However, deciding on a reaction based on the emotional state requires recognizing human expressions. The resulting analysis has numerous applications in mobile computing [5], [6], robotics [7], health monitoring [8], [9], or gaming [10], just to mention a few. Technically, for an algorithm developed using computer vision techniques, several issues can affect the performance. For example, different subjects might express the same emotion non-identically. In addition, different viewpoints result in unequal representations of the emotion. Moreover, the presence of occlusions and changes in illumination might mislead the recognition method. If the emotion needs to be recognized based on voice, the ambient noise and the differences between voices of different subjects are significant factors that might affect final recognition performance. In order to accurately recognize emotions, humans use both audio and visual signals.

- *F. Noroozi is with the Institute of Technology, University of Tartu, Tartu 50411, Estonia. E-mail: fatemeh.noroozi@ut.ee.*
- *M. Marjanovic and A. Njegus are with the Faculty of Technical Sciences, Singidunum University, Belgrade 11000, Serbia. E-mail: {mmarjanovic, anjegus}@singidunum.ac.rs.*
- *S. Escalera is with the Department of Mathematics and Informatics, University of Barcelona, Computer Vision Center, Barcelona 08007, Spain. E-mail: sergio@maia.ub.es.*
- *G. Anbarjafari is with the iCV Research Group, Institute of Technology, University of Tartu, Tartu 50411, Estonia, and the Department of Electrical and Electronic Engineering, Hasan Kalyoncu University, Gaziantep 27410, Turkey. E-mail: shb@icv.tuit.ut.ee.*

Manuscript received 4 Nov. 2016; revised 3 May 2017; accepted 3 June 2017. Date of publication 8 June 2017; date of current version 7 Mar. 2019. (Corresponding author: Gholamreza Anbarjafari.) Recommended for acceptance by S. Zafeiriou, G. Zhao, I. Kotsia, M. Nicolaou, and J. Cohn. For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org, and reference the Digital Object Identifier below. Digital Object Identifier no. 10.1109/TAFFC.2017.2713783

According to [11], humans use coverbal signs to emphasize the implications of their speech. These include body, finger, arm and head gestures, and facial expressions such as gaze and speech prosody. This is because 93 percent of human communication is performed through nonverbal means, which consist of facial expressions, body language and voice tone. Computerized Facial Expression Recognition (FER) consists in face detection and tracking, feature extraction and recognition [12]. First, the face is detected and tracked throughout a chain of images constituting a video sequence. The examples of this solution include the (spatial) ratio template tracker [13], the enhanced Kanade-Lucas-Tomasi tracker [14], the AdaBoost learning algorithm [15], the robust face detection algorithm [16] or the piecewise Bezier volume deformation tracker [17], among others. As facial expressions are dependent on the translation, scaling or rotation of the head, both motion-based and model-based representations are considered, i.e., geometric normalization and segmentation.

The next step is to extract information from the detected face which can help to distinguish the intended emotion [18]. The main two categories of facial features are geometric and appearance features such as distances between two determined facial landmarks or angles. The geometric features consist of the shapes of specific parts of the face, such as eyes, eyebrows and mouth, and the locations of facial points, e.g., the corners of the eyes and mouth. The appearance features are based on the whole face or a certain region in it. They can be extracted by using texture filters such as Gabor. They concern the textures of the skin, which are affected by wrinkles, furrows and bulges [19].

In this paper, we propose a methodology to recognize emotions based on audio-visual data. We conduct multiclass classification, where only a single emotion is expected to be represented by each sample. For the audio data, we use the Mel-Frequency Cepstral Coefficients (MFCCs), Filter Bank Energies (FBEs) and statistics and acoustics features

Fig. 1. Four selected frames of eight different human subjects samples of RML dataset expressing frames of angry emotions.

[20], [21]. For the video, we first represent the data by using key-frames, and then facial geometric relations and convolution. We use the state-of-the-art classifiers to learn each feature space independently. The final prediction is based on a stacked second level classification. It uses the confidence outputs of each individual classifier as its inputs. The experiments are implemented on the SAVEE [22], eNTERFACE'05 [23] and RML [24] databases, showing significant improvements in the recognition rates with respect to the state-of-the-art alternatives. The rest of the paper is organized as follows: Section 2 reviews the related work in the field of audio-visual emotion recognition and video summarization based on key-frames. In the Section 3 the details of the proposed method are described followed by the experimental results and discussion provided in Section 4. Finally, conclusions are presented in Section 5.

2 Related Work

Researchers have made numerous efforts to improve emotion recognition based on the fusion of audio and visual information [25]. The fusion of multimodal data, according to [26], can be classified into data/feature level fusion, kernel based fusion, model-level fusion, score-level fusion, decision-level fusion, and hybrid approaches for audio-visual emotion recognition. Early multimodal emotion recognition results were presented by Wimmer et al. [27] which combined descriptive statistics of audio and video low-level descriptors.

In this paper, we focus on the Surrey Audio-Visual Expressed Emotion (SAVEE) [22], eNTERFACE'05 [23] and RML databases to perform audio-visual emotion recognition. The SAVEE database is based on the English language, where the speakers are from the same nationalities. The eNTERFACE'05 is also based on English, but the speakers are from seven different nationalities. However, RML includes eight human subjects, speaking six different languages. In RML, different accents of the English and Chinese languages were considered. Examples of subjects of the considered databases after key-frame selection are shown in Fig. 1.

Based on the SAVEE database, Cid et al. [28] employed the Dynamic Bayesian method as classifier on the audio mood by using the energy, pitch and tempo as audio features. For the visual part, a set of edge-based features is extracted, which is invariant to scale or distance from the user to the robot. The Gaussian Classifier with Principal Component Analysis (PCA) is applied in Sanaul Haq et al. [29]. This work used 106 audio and 240 visual features. The audio features consisted of pitch, energy, duration and MFCC features, whereas the visual features were related to positions of the 2D marker coordinates.

Gharavian et al. [30] investigated the performance of the ARTMAP neural network (FAMNN) on the audio database with the audio features such as: MFCC, pitch, Zero Crossing Rate (ZCR), energy and formants and the visual features (marker locations on the face) were extracted and the features were reduced by the PCA feature reduction algorithm. Next, the FCBF feature selection method was applied to the reduced features.

On the eNTERFACE'05 database, Datcu et al. [31] used a Hidden Markov Model (HMM) as the classifier on the audio and visual emotional moods. The audio features considered were energy, pitch, MFCCs, and formants. The video features included coordinates-based and distance-based features. The coordinates-based features were also used in the work of Paleari et al. [32]. The authors used Neural Networks (NN) to improve multimodal emotion recognition. This work increased the output performance by using features including Energy and MFCC with their delta and acceleration terms for the audio part, and applying local Binary Patterns (LBP) to estimate features in seven segments of the face for the visual part of the data. Jiang et al. [33] investigated the influence of different sets of features for recognition. For the audio part, the authors used fourteen MFCCs, together with their first-order and second-order differential coefficients, resulting in a 42-dimensional audio feature vector. The distances between specific pairs from 83 facial landmarks were considered as facial features. The authors applied HMM for classification.

Huang et al. [34] considered prosodic and frequency-domain for audio features, and geometric and appearance-based features for facial expression description. Each feature vector was used to train a unimodal classifier using a back-propagation neural network. They proposed a collaborative decision-making model using a genetic learning algorithm, which was compared to concatenated feature fusion, BPN learning-based weighted decision fusion, and equal-weighted decision fusion methods. In one of the most recent works, Gera et al. [35] surveyed the effects of changing the features and methods on the eNTERFACE'05 database. The authors applied a multiclass Support Vector Machine (SVM) for classification. A 41-dimensional feature vector was computed from Pitch, Energy, the first thirteen MFCCs and their first and second derivatives for audio-based emotion recognition. The authors also considered face tracking and geometric features for facial emotion recognition.

Related to the RML database, Wang et al. [26] investigated the influence of Kernel Matrix Fusion (KMF) by using the unsupervised Kernel Principal Component Analysis (KPCA) and supervised Kernel Discriminant Analysis (KDA) for audio-visual emotion recognition. The authors used prosodic and spectral features for audio representation, 2-D discrete Cosine transform on determined blocks of images for visual feature extraction, and a weighted linear combination for fusion at the decision level. Fadil al et. [36] employed the Deep Networks-Multilayer Perceptron (MLP) as the classifier and prosodic features such as pitch, energy and linear prediction and cepstral coefficients for the audio part and discrete Fourier coefficients and PCA projections

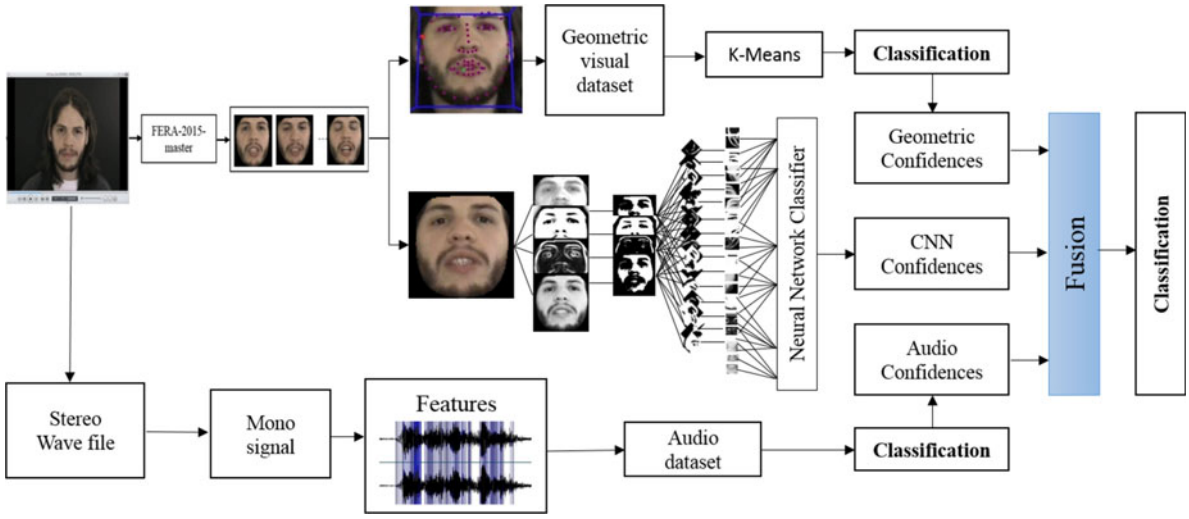


Fig. 2. Proposed audio-visual emotion recognition system.

of the face for the visual part. Seng et al. [37] benefited from the combination of a rule-based technique and machine learning methods to improve multimodal recognition. Bidirectional Principal Component Analysis (BDPCA) and Least-Square Linear Discriminant Analysis (LSLDA) were used for the visual cue. The extracted visual features used the Optimized Kernel-Laplacian Radial Basis Function (OKL-RBF) neural classifier. The audio cue was computed as a combination of prosodic and spectral features.

In this work, as in most of the state-of-the-art audio-visual approaches, we focus on MFCC-based features for audio description. On the other hand, for the visual part we summarise videos based on key-frame representations and use two sets of complementary features to train different vision-based classifiers: one based on spatial-relations of face landmarks and the other based on convolutions learnt from a Convolutional Neural Network (CNN). We will show how using the confidence output of all the set of classifiers from audio and visual data as a new feature vector representation provides a new highly discriminative descriptor, which is learnt by a second level stacked classifier in order to obtain final emotion prediction.

2.1 Key-Frame Selection Strategy

The goal of key-frame selection is to find a set of representative frames from an image sequence. However, most of the current methods are either computationally expensive or cannot effectively capture the salient content of the video. In general, there are several key-frame selection strategies, such as:

(1) Motion analysis based strategy—which compute the optical flow for each frame in order to estimate whether or not the facial expression changes [13], [38]. The main drawback is that it captures local variations and may miss important segments while longer segments might appear multiple times with similar content;

(2) Shot boundary based strategy—which takes the first, the middle, and the last frames of each shot as the key-frame [39]. Although this strategy is very simple and fast, usually these frames are not stable and do not capture the major visual content;

(3) Visual content-based strategy—which uses multiple criteria (shot based, color feature, and motion based criteria) [40]. First, the first frame is selected as key-frame, and then the current frame will be compared with other based on the similarities defined by a color histogram from my Google search or moments. If a significant content change occurs, then the new frame will be selected as the key-frame. The major drawback is that it does not effectively capture the major or significant content of the video shot;

(4) Clustering-based strategy—which attempts to group frames with a similar posture. Each frame is assigned to a corresponding cluster, and those closest to the centroid of each cluster are selected as key-frames [40], [41], [42]. The clustering methods demonstrate good performance in general; however, these methods can be easily affected by noise and motion, may end up selecting key-frames only from the dominant clusters and may overlook events which occur infrequently.

In this paper we aim to achieve automatic key-frame selection using a clustering based strategy. In order to deal with possible effects of noise and motion, we perform clustering on a set of robust detected and tracked facial landmarks. It provides a fast, simple, and accurate methodology to summarize face videos.

3 Audio-Visual Emotion Recognition System

In this section, we describe the proposed audio-visual system for emotion recognition that benefits from the fusion of several classifiers outputs, where each classifier focuses on a different and complementary set of features. The summary of our system is shown in Fig. 2. We perform fusion at the decision-level [43], which consists of processing the classification results from both the visual and audio modalities. The confidence outputs of audio and video classifiers are used as an input for the final multimodal prediction. Next, we explain the features considered by the audio and visual classifiers. Then, we describe how classifier outputs are combined, and which classifiers are considered in order to obtain final emotion prediction.

3.1 Audio Features

Emotions can be described from audio based on different features representative of each emotion. For example, anger can be described by a faster speech rate, high energy and pitch frequency. Many authors agree that the most important audio features for emotion recognition are pitch, intensity, duration, spectral energy distribution, Mel Frequency Cepstral Coefficients (MFCCs), $\Delta MFCCs$, average Zero Crossings Density (ZCD) and filter-bank energy parameters (FBEs) [20], [44], [45], [46] . These features represent the differences between the prosodic patterns of the voices of different speakers. The prosodic pattern depends on the speaker's accentuation, speaking rate, phrasing, pitch range and intonation. Spectral features that are extracted from short speech signals are useful for speaker recognition as well. The mentioned features can be summarized as follows:

- Pitch can be estimated either in the time or frequency domain, or by using a statistical approach. For a speech signal s, the pitch, $\rho_0(s)$, can be estimated as follows:

$$\rho_0(s) = \aleph\{\log|\aleph(s.{\omega_n}^H\|s\|)|\}, \quad (1)$$

  where $\aleph$ stands for the DFT function, and $\|s\|$ denotes the length of the signal. Moreover, ${\omega_n}^H$ is the Hamming window, which is calculated as follows:

$${\omega_n}^H = 0.54 - 0.46\cos\left(\frac{2\pi n}{L}\right), \quad 1 \le n \le N-1. \quad (2)$$

- Intensity measures the syllable peak, and represents the loudness of the speech signal. The syllable peak is its central part, which usually is a vowel. Intensity can be calculated as follows:

$$I_i(s) = \frac{\sum_{n=1}^{N}(s_{i+n}).w_n^H}{\sum_{n=1}^{N} w_n^H}. \quad (3)$$

- Standard deviation is one of the features of a speech signal. It is formulated as follows:

$$std = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(s_i - \alpha)^2}, \quad (4)$$

  where s_i shows the value of the speech signal at i, and α is its mean. Moreover, N is the length of the speech signal.
- ZCR can be calculated as follows:

$$ZCR = \frac{1}{N-1}\sum_{i=1}^{N-1}\mathbb{I}\{s_i s_{i-1} < 0\}, \quad (5)$$

  where $\mathbb{I}$ is the indicator function.
- If we consider a time delay τ, the autocorrelation function $r(\tau)$ maximizes the inner product of the speech signal by its shifted version, as follows:

$$r(\tau) = \frac{1}{N}\sum_{n=0}^{N-1} s(n)s(n+\tau). \quad (6)$$

- Formant frequencies, f, represent the resonating frequencies of the speaker's vocal tract. They are calculated as follows:

$$f = \frac{F_s}{2\pi}\arctan\frac{\mathrm{im}(F(s))}{\mathrm{re}(F(s))}, \quad (7)$$

  where F_s stands for the sampling frequency. Moreover, the real and imaginary parts of the speech signal in the frequency domain are shown by $\mathrm{re}(F(s))$ and $\mathrm{im}(F(s))$, respectively. In this paper, we consider the mean, standard deviation, minimum and maximum of the third formant and the fourth formant bandwidth.
- The intuitive feeling of a rough voice can be represented by using Harmonics-to-Noise Ratio (HNR). Young speakers usually produce speech signals with HNR $\approx$ 20 dB. It means that 99 percent of the energy of the signal is periodic, and the rest, i.e., 1 percent, consists of noise.
- Cepstrum Coefficients (CC) can be utilized for separating the original signal from the filter. The signal can be truncated at different frequencies, in order to extract different levels of spectral details. For example, in order to analyze the vocal tract, low coefficients should be considered. Cepstrum is calculated by finding the DFT of the log magnitude of the DFT of the signal.
- Davis et al. [47] proposed a model to calculate MFCCs, as follows:

$$MFCC_i = \sum_{\theta=1}^{N}\cos\left[i(\theta-1)\frac{\pi}{N}\right], \quad i = 1, 2, \ldots, M, \quad (8)$$

  where M stands for the number of cepstrum coefficients, X_θ, $\theta = 1, 2, \ldots, N$, is the log energy output of the θth filter, and N denotes the number of triangular band pass filters.
- FBEs and their derivatives are calculated by using a first order Finite Impulse Response (FIR) filter. The filter has a coefficient α. Short-time Fourier transform is required as well. The impulse function $\delta(n)$ is defined such that $\delta(0) = 1$. By passing it through a discrete filter, the impulse response $h(n)$ is obtained. For a given input signal $x(n)$, the output $y(n)$ is calculated as follows:

$$y(n) = \sum_{i=0}^{M} a_i x(n-1) + \sum_{j=1}^{N} b_i y(n-j), \quad (9)$$

  where $a_i = h(i)$ and M is the order of the filter function. The FBEs are calculated as follows:

$$y(m) = \sum_{\theta=0}^{L-1} h(\theta)x\left[(m-\theta)\bmod(N)\right], \quad (10)$$

  where $m = 0, 1, \ldots, N-l$ and L is the length of the filter pulse [48].
- In [49], the following equation has been proposed for calculating $\triangle MFCCs$

$$C(n) = \mathrm{DCT} * \log(y(m)), \quad (11)$$

  where DCT is the discrete cosine transform.

TABLE 1
Extracted Audio Features

Features group	Number
Pitch	1
Intensity	1
Percentile	1
Formants	20
Formant bandwidth	4
ZCR	1
average ZCD	1
Statistics	7
$MFCCs$	13
$\Delta MFCCs$	13
Filter Bank Energies (FBE)	26

In our case, para-linguistic features are the basis. These types of features are adequate for our purpose since they are independent from the lexical meaning. For each sample in the considered database, an 88-dimensional feature vector is computed with the set of features shown in Table 1.

3.2 Visual Features

In order to classify a video with a particular emotion, several hundreds (or thousands) of frames have to be processed. However, in most cases similar facial expressions appear within the same video, which are representative of a particular emotion. Here, we hypothesise that summarising an emotion video by a set of a few key-frames in terms of the variability of facial expressions will be enough in order to describe and efficiently learn the contained emotion.

3.2.1 Key-Frames Definition

In order to define a set of key-frames per video, we base this work on geometric features (areas around the mouth, eyes, eyebrows, and nose), that locate N_l = 68 landmark points, as shown in Fig. 3. Facial features are detected and tracked from the video files using FERA 2015 code.[1]

Obtaining the landmark locations from a video stream might cause the trouble of being unable to fix the head poses. Therefore, initially, we extract the frames from the recorded videos. Only the frames that contain frontal faces are kept. Afterward, we align the faces, in order to fix the orientations of the frontal faces. Then we detect the landmark positions.

We adopt the vector of extracted landmarks from one frame of video is $P = [p_1, p_2, \ldots, p_{68}]$. All landmarks are grouped into six regions. Points from 1 to 17 mark the face contour, points from 18 to 22 belong to the left eyebrow, 23 to 27 belong to the right eyebrow, 37 to 42 and 43 to 48 to the left and right eye, respectively, 28 to 36 to the nose and from 49 to 68 to the mouth region.

The inner facial landmarks $P = [p_1, p_2, \ldots, p_{49}]$ of each video are aligned with a mean shape, using landmarks such as: 20, 23, 26, 29 (eyes corners region) and 11-19 (nose region). Those points are not affected by Activation Units and they are considered as stable. The mean facial landmarks shape has been calculated before geometric features extraction. It is calculated for each database by taking mean of 10 percent randomly selected video frames from every video. Performing a non-reflective affine transformation computing, the difference between stable point coordinates of the two shapes is minimized. Then, all mean shape landmark coordinates are subtracted from the corresponding aligned shape points. Vector of aligned landmarks for each frame is presented such as $P' = [p'_1, p'_2, \ldots, p'_{49}]$ resulting in 49*2 = 98 geometric features.

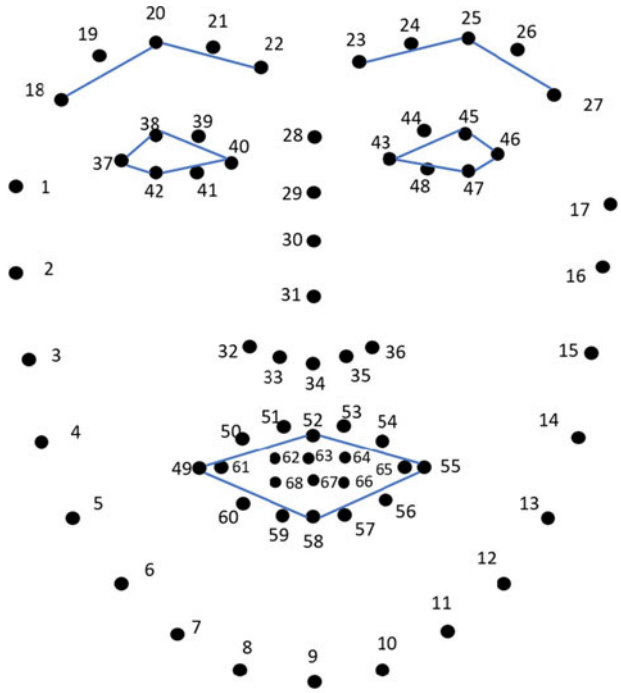


Fig. 3. Annotated facial distances for angles calculation.

Each video is sampled by 25 frames per second. In order to select the most significant frames for each video, we apply k-means clustering, where k = 4 is used for the purpose of this work. As a key-frame, we adopted the landmarks coordinates of the closest image to centroids μ_j. For each instance, we assign it to a cluster with the closest centroid $c_i := argmin\|p'_i - \mu_j\|$.

After k-means clustering, from each video we select k vectors of landmarks $C = [c_1, c_2, \ldots, c_{68}]$ and the number of samples becomes equal to N'. Since each landmark has x and y coordinates locating N_l = 68 points, in a two-dimensional euclidean plane, they can be presented as $c_i = (c_{i,x}, c_{i,y})$, where $i \in \{1, \ldots, N_l\}$. Two examples of two summarizing videos with frames of angry emotions, each one represented by 4 key-frames, are shown in Fig. 4. Note that

Fig. 4. Examples of the four key-frames representing two videos of the angry emotion of the eNTERFACE'05 database.

1. *https://github.com/TadasBaltrusaitis/FERA-2015*

TABLE 2
Visual Feature Distances Before PCA

Features group	Distances $c_i - c_j$
Eyebrows	18-19, 19-20, 20-21, 21-22, 23-24, 24-25, 25-26 26-27, 18-37, 20-38, 22-40, 23-43, 25-45, 27-46
Eyes	37-38, 38-39, 39-40, 40-41, 41-42, 37-42 43-44, 44-45, 45-46, 46-47, 47-48, 43-48
Mouth	49-50, 50-51, 51-52, 52-53, 53-54, 54-55 55-56, 56-57, 57-58, 58-59, 59-60, 49-60
Nose	32-33, 33-34, 34-35, 35-36, 32-49, 36-55
Chin	6-7, 7-8, 8-9, 9-10, 10-11, 11-12

visually we can easily discriminate the target emotion by the automatically estimated key-frames.

3.2.2 Visual Descriptors

From the set of selected key-frames representing a video, we train two classifiers, one based on a CNN (details in experimental section) and another based on geometric features.

For the geometric descriptor, we calculate the consecutive euclidean distances $d(c_i, c_{i+1})$ between the selected landmarks

$$d(c_i, c_{i+1}) = \sqrt{(c_{i+1,x} - c_{i,x})^2 + (c_{i+1,y} - c_{i,y})^2}. \quad (12)$$

After that, we normalise them by dividing them with the length of the region where the corresponding distance belongs to

$$\hat{\mathbf{d}}(c_i, c_{i+1}) = \frac{d(c_i, c_{i+1})}{\sum_j d(c_j, c_{j+1})}. \quad (13)$$

According to the Fig. 3, $j = 18, \ldots, 26$ if the distance that we want to normalize belongs to the eyebrows region, $j = 37, \ldots, 41$ or $j = 43, \ldots, 46$ if this distance belongs to eye region, $j = 49, \ldots, 59$ mouth region, $j = 32, \ldots, 35$ nose region or $j = 6, \ldots, 11$ if the distance that we want to normalize belongs to the chin region. Those distances are shown in Table 2. Additionally, for each group of landmarks, we calculate the angles between two lines defined by two pairs of landmarks that share one common landmark. Therefore, for each triplet of points $c_i - c_j - c_k$, shown in Table 3, we calculate their corresponding angle a_j according to

$$a_j = \arccos \frac{d(c_i,c_j)^2 + d(c_i,c_k)^2 - d(c_j,c_k)^2}{2d(c_i,c_j)d(c_i,c_k)}. \quad (14)$$

According to previous research papers in [50] and in [51], the region around eyes and mouth are the regions that are the most significant in recognizing specific emotions. Therefore, we selected landmarks and angles related to previous research in the topic. In the second stage, we calculated the consecutive euclidean distances between the selected landmarks, and then we normalised them by dividing with the length of the region to make them invariant to scale. For each group of landmarks, we calculated the angles between two lines defined by two pairs of landmarks that share a common landmark. Therefore, we come up with 10 angles in total. This keeps a low complexity of the methodology while providing discriminative results.

TABLE 3
Visual Feature Angles Before PCA

Features group	Angles $c_i - c_j - c_k$
Eyebrows	18-20-22, 23-25-27
Eyes	38-37-42, 37-38-40, 38-40-42, 45-46-47, 43-45-46, 45-43-47
Mouth	53-49-58, 52-55-58

There are 10 angles in total selected for classification. Thus, there are $N_f = 60$ features in total extracted from the face region.

For further description, let's consider that notation used for the training set is $\boldsymbol{x^{(i)}}, y^{(i)}$, where $\boldsymbol{x^{(i)}} \in R^{N'}$ is a vector of extracted features, $y^{(i)}$ is its associated class and $N' = 4$ frames per video, represents the number of the samples after 4-means clustering. For one sample vector, we adopt that

$$x_j = \hat{\mathbf{d}}(c_i, c_{i+1}), \quad (15)$$

where $j = 1, \ldots, 44$, and

$$x_j = a_i, \quad (16)$$

where $j = 45, \ldots, 54$, and $i \in [1, \ldots, N_l - 1]$ represents the corresponding landmark.

3.2.3 Principal Component Analysis

In order to reduce the dimensionality of data from N_f dimensional feature vector to r dimensional feature vector $\boldsymbol{z^{(i)}} \in R^{rX1}$, we apply Principal Component Analysis (PCA) as it is described in [52].

PCA is computed from the correlation matrix and used in conjunction with a Ranker search. Dimensionality reduction is accomplished by choosing thirteen eigenvectors to account for some percentage of the 95 percent variance in the original data. Attribute noise is filtered by transforming to the PC space, eliminating some of the worst eigenvectors without transforming it back to the original space.

As an input to PCA, we selected the visual features shown in Table 25 such as contour of eyes, eyebrows, top of the nose, mouth outline and chin (from 6 to 12). Apart from those consecutive distances, we included vertical distances: two distances between the region of the mouth and nose (32-49 and 36-55), and distances between the eyes and eyebrows (18-37, 20-38, 22-40, 23-43, 25-45 and 27-46).

3.2.4 Classification for Geometric Features

After applying PCA, we have obtained a new data set $\{\boldsymbol{z_j^{(i)}}, y^{(i)}\}$. $y^{(i)}$ represents the associated class. For the classification of geometric features, we adopt Multiclass-SVM (M-SVM), which includes multiple binary SVMs [53], and Random Forest (RF) [54]. RF is chosen because it is simple and efficient, and M-SVM is selected due to its fastness and reasonable performance.

3.3 Learning and Fusion

In previous sections we presented the set of features used to describe the audio channel and the generated visual

(a) KL (angry) (b) JK (happy) (c) JE (sad) (d) DC (neutral)

Fig. 5. Sample images showing different emotional states from the SAVEE database. The images have been taken from [58].

samples composed by a few key-frames that represent each video. For the visual part, from the estimated key-frames both geometric and CNN-based features are computed from key-frames. The multiclass SVM classifier was used to learn each feature space separately. Four in total: three multiclass SVM for audio, left, and mono audio channels, and one for geometric visual features. A fifth classifier is obtained by the CNN model considering as input the computed key-frames. In all five cases, we collect the output $confidences$ of the classifiers ($margin$ for SVM and probability for CNN) for all possible target emotion labels. The margin of a training set $\{\boldsymbol{x^{(i)}}, y^{(i)}\}$ (or $\{\boldsymbol{z^{(i)}}, y^{(i)}\}$ in case of PCA implementation) with respect to a corresponding classifier is presented as $y^{(i)}(\boldsymbol{\omega^{(i)}}\boldsymbol{x^{(i)}})$. The sign of the margin is positive if the classifier $\boldsymbol{\omega^{(i)}}$ correctly predicts the label $y^{(i)}$. The absolute value of the margin $\boldsymbol{m^{(i)}} = \left|y^{(i)}(\boldsymbol{\omega^{(i)}}\boldsymbol{x^{(i)}})\right| = \left|\boldsymbol{\omega^{(i)}}\boldsymbol{x^{(i)}}\right|$ represents the $confidence$ in the prediction. Finally, the output confidences of each classifier for all possible emotion labels are considered as new features to be fused in a new descriptor which is learnt again by a final multiclass SVM classifier applied on new dataset $\{\boldsymbol{m^{(i)}}, y^{(i)}\}$, where $\boldsymbol{m^{(i)}} \in R^{N'}$ is a vector of confidences, $y^{(i)}$ is its associated class and N' represents the number of the samples after 4-means clustering. In our case, the five trained initial classifiers provide six emotion confidences each, creating a final feature space of 30 dimensions to be learnt by the final multiclass SVM stacked classifier.

4 Experimental Results

In this section, we present the results of our audio-visual emotion recognition system based on the three mentioned databases, i.e., SAVEE, eNTERFACE'05 and RML. We first present the results considering the classifiers per each modality independently, and then we show the fusion results. For each experiment, we also describe the experimental setup and the obtained results in comparison with state-of-the-art alternatives. The summary of the results and the comparison of all the experiments are shown in Tables 25, 26 and 27, respectively.

All the databases contain audio-visual clips. SAVEE includes 60 samples for each of the six emotions, namely, anger, disgust, fear, sadness, surprise and happiness, and 120 for subjects neutrally speaking in English. Similarly, eNTERFACE'05 contains 44 subjects with different nationalities, but all speak in English, with the same six basic emotions. Finally, in the RML database, the six emotions are acted out by eight actors in seven different languages.

4.1 Databases Explanation

Several recent studies in the context of multimodal emotion recognition have focused on posed databases [30], [31], [32], [33], [34], [35], [36], [37], [44], [55], [56], [57], in contrast with spontaneous databases. In order to demonstrate the contribution of our approach, we used three posed databases, which have been successfully utilized in previous studies. The chosen databases are SAVEE [22], eNTERFACE'05 [23] and RML [24].

4.1.1 Surrey Audio-Visual Expressed Emotion (SAVEE) Database

The SAVEE database contains recordings of four male subjects who aged from 27 to 31, and played six basic emotions, namely, anger, disgust, fear, happiness, sadness and surprise, as well as the neutral state. In the database, 480 native British English utterances are available, which consist of 60 samples for each of the mentioned emotional states, and 120 for the neutral. The subjects were recorded while emotional and text prompts were displayed in front of them on a monitor. A video clip and three pictures were included in the prompts for each emotion. The text prompts were divided into three groups for each emotion, in order to avoid fatigue. Each round of acting was repeated if necessary, in order to guarantee that it has been performed properly. The samples were evaluated by 10 subjects, under audio, visual and audiovisual conditions. A set of sample images from the SAVEE database is shown in Fig. 5. In this work, we used all the samples from the emotional states, but only 60 from the neutral samples, i.e., 420 in total.

4.1.2 eNTERFACE'05

The eNTERFACE'05 database contains samples created from 42 subjects with different nationalities. All of them spoke English. From the subjects, 81 percent were male, and the remaining 19 percent were female. 31 percent of the subjects wore glasses, and 17 percent of them had beard. A mini-DV digital video camera with a resolution of 800,000 pixels was used to record the samples at a speed of 25 frames per second (FPS). A specialized high-quality microphone was utilized for recording uncompressed stereo voice signals at a frequency of 48,000 Hz in 16-bit format. The microphone was located around 30 cm below the subject's mouth, and outside of the camera's field of view. In order to ensure easy face detection and tracking, a solid background with a dark gray color was used, which

Fig. 6. A set of sample images from the eNTERFACE'05 database [23].

Fig. 7. A set of sample images from the RML database. The images have been taken from [59].

covered the whole area behind the subjects while recording them. Each subject acted six different emotional states, namely, anger, disgust, fear, happiness, sadness and surprise, as well as neutral that is shown in Fig. 6.

4.1.3 RML

The RML database was created at the Ryerson Multimedia Lab. It includes 720 samples containing audiovisual emotional expressions. Six basic emotions were considered, namely, anger, disgust, fear, happiness, sadness and surprise. The recordings were performed in a quiet and bright atmosphere with a plain background, by using a digital camera. Eight subjects participated in the recordings, and spoke various languages, namely, English, Madarin, Urdu, Punjabi, Persian and Italian, as well as different accents of English and Chinese. By using 16-bit single channel digitization, the samples were recorded at a frequency of 22,050 Hz. The recording speed was set to 30 FPS. The duration of each video is between 3 and 6 seconds. A set of sample images from the RML database is shown in Fig. 7.

4.2 Evaluation for Both Modalities

We applied 10 fold cross-validation, where the original database is randomly divided into 10 subsamples. Then, from those, one is treated as a test set, and the remaining nine are used as the training data. The cross-validation process is then repeated 10 times, with each of the 10 subsamples used once as the test data. At the end, the results from the 10 test folds are averaged.

TABLE 4
Audio Recognition Rates by Using SVM and RF Classifiers, with and Without Applying PCA, Based on SAVEE, RML and eNTERFACE'05 Databases

Audio	SVM	SVM-PCA	RF	RF_PCA
SAVEE	48.81	42.98	56.07	52.86
RML	43.47	41.67	65.28	47.92
eNTERFACE'05	41.32	42.04	47.11	38.37

4.3 Audio Modality Results

After extracting the 88 features described in Table 1, we performed classification by using multiclass SVM and RF. The results with and without applying PCA, by using 10-fold cross validation, are shown in Table 4. We can see that the best average recognition rates are obtained by the RF classifier without applying PCA: 56.07, 47.11 and 65.28 percent for SAVEE, eNTERFACE'05 and RML databases, respectively.

4.4 Visual Modality Results

In order to show the visual emotion recognition results, we first present the performance of the geometry-based classifier and then the performance of the CNN model.

4.4.1 Geometric Visual Recognition

In order to compute the geometric visual recognition results, we apply the same evaluation procedure as in the case of audio features. As classification methods, multiclass SVM, with the Polynomial Kernel and exponent parameter experimentally set to 1. For RF algorithm, we set the number of trees to 10. Both classifiers are applied with and without PCA, by using the 10-fold cross-validation. PCA is computed from the correlation matrix and used in conjunction with a Ranker search. Dimensionality reduction is accomplished by choosing eigenvectors to account for some percentage of the 95 percent variance in the original data. After applying PCA, we selected 4 eigenvectors, that correspond to new 4 features in the projected space as combinations of original set of distances and angles. The projected space presents the combination of 12 distances and 8 angles from the original dataset. Distances and angles after applying PCA figuring in the projected space are presented in Tables 5 and 6, respectively. New feature space is presented in Table 7

TABLE 5
Visual Feature Distances After PCA

Features group	Distances $c_i - c_j$
Eyebrows	19-20, 21-22, 23-43
Eyes	43-44, 38-39, 47-48, 41-42
Mouth	49-50, 54-55
Nose	32-36
Chin	7-8, 10-11

The geometric visual recognition results based on the three mentioned databases are represented in Table 8 by applying SVM and RF, with and without PCA. The best results are obtained by the RF classifier, which for SAVEE, eNTERFACE'05 and RML databases are 56.07, 41.59 and 73.04 percent, respectively. According to the results that are presented in Tables 4 and 8, applying PCA leads to stronger results only in the case of visual emotion recognition by using SVM. An important factor that affects the recognition performance is the number of the samples which have been used to train the system.

In the case of audio, we have a limited number of instances, i.e., 480, 720 and 1,263 samples from SAVEE, RML and eNTERFACE'05 databases, respectively. From each signal, 88 acoustics features have been extracted. However, the visual datasets contain more instances, i.e., 1,920, 2,880 and 5,052 samples from SAVEE, RML and eNTERFACE'05 databases, respectively. From each visual sample, 60 geometrical features have been extracted.

After applying PCA to the datasets for both modalities, it reduced the number of features to 30 for the audio dataset, while for the visual dataset, 20 were kept for the linear characterization of the feature space. Given that we have more visual samples due to the boosting on the data, the space that is projected by PCA could take advantage of more information for the visual dataset. This may explain why applying PCA to the visual dataset is more beneficial than the audio. About the difference between the effect of PCA on the two classifiers, we should note that in the case of RF, proportionality of the numbers of samples and features is needed in order to explain the variations between the samples. In other words, RF generally needs a larger number of samples to handle the randomization concept properly, and generalize the classifier at the level that is required for new test datasets. That is why PCA improves the visual recognition performance in the case of SVM, but not in the case of RF.

TABLE 6
Visual Feature Angles After PCA

Angle	Angles $c_i - c_j - c_k$
1.a_{20}	18-20-22
2.a_{25}	23-25-27
3.a_{37}	38-37-42
4.a_{38}	37-38-40
5.a_{40}	38-40-42
7.a_{45}	43-45-46
8.a_{43}	45-43-47
9.a_{49}	53-49-58

TABLE 7
Ranked Attributes After PCA

Ranking	Attribute
First:0.3761	0.16* d(43,44) + 0.16* d(38,39) + 0.16* d(47,48) + 0.16*(d41,42) + 0.16* d(19,20)
Second:0.2163	0.255* d(49,50) + 0.235* d(7,8) - 0.231* d(10,11) - 0.223*a_{37} + 0.215* d(21,22)
Third:0.0898	0.26*a_{43} - 0.254*a_{38} - 0.245* d(23,43) + 0.241* d(32,36) - 0.231* d(54,55)
Fourth:0.0451	-0.449*a_{40} + 0.41 *a_{20} - 0.309*a_{25} + 0.252*a_{45} + 0.233*a_{49}

4.4.2 CNN Visual Recognition

We also trained a CNN model by using the video samples represented by four key-frames. The output of the CNN is a set of six confidence values, i.e., one per emotion.

We used GoogLeNet [60]. It is a medium size network, and makes it easy to train and test with more than 100 thousand images in each fold. The network, which makes it easy based on the repetition of the inception module following the idea of a network in the network. This module is repeated nine times inside GoogLeNet, and is composed of the first level of 1×1 convolutions and a 3×3 max pooling, a second level of 1×1, 3×3, and 5×5 convolutions, and a third level of inception module with a filter concatenation step that joins all the previous results.

The width of inception modules ranges from 256 (in early modules) to 1,024 filters (in top inception modules). Given the depth of the network, the ability to propagate gradients back through all the layers is done by adding auxiliary classifiers connected to these intermediate layers on top of the output of the inception modules. During training, their loss gets added to the total loss of the network (multiplying by a factor of 0.3). At inference time, these auxiliary networks are discarded.

The original images have been resized to 256×256 pixels; then during the training phase, the network takes images as a random crop of 224×244 from the resized database.

In order to be able to avoid beginning with an empty network and thus having an expensive learning phase, we trained GoogLeNet weights with initial values coming from an Age/Gender Face classification network pre-trained from hundreds of thousands of different images, coming from a filtered mix of the Imdb-Wiki and Adience [61] databases. This previous network is specialized to detect details in faces, and was a good starting point for the first layers of network filters.

For the learning rate, we use a step-down policy with a starting value of 0.01 and an automatic decrease of 1/10 every 33 percent of the training phase.

TABLE 8
Geometric Visual Recognition Rates by Using SVM and RF as the Classifier, with and without Applying PCA, Based on SAVEE, RML and eNTERFACE'05 Databases

Visual	SVM	SVM-PCA	RF	RF_PCA
SAVEE	36.10	51.88	56.07	52.86
RML	31.67	36.91	73.04	72.92
eNTERFACE'05	30.38	30.84	41.59	40.05

TABLE 9
CNN Results for SAVEE

SAVEE	Anger	Disgust	Fear	Sadness	Surprise	Happiness	Neutral	Recognition rate (%)
Anger	13,164	144	3	3	50	5	7	98.42
Disgust	233	13,923	71	1	0	19	0	97.73
Fear	13	73	12,944	20	10	337	98	95.92
Sadness	3	2	43	13,540	10	0	100	98.85
Surprise	28	8	1	136	25,781	0	1	99.33
Happiness	67	8	233	0	0	15,834	0	98.09
Neutral	43	0	264	112	100	290	13,007	94.14
							Average rate (%)	97.50

TABLE 10
CNN Results for eNTERFACE'05

eNTERFACE'05	Anger	Disgust	Fear	Sadness	Surprise	Happiness	Recognition rate (%)
Anger	11,441	2,443	1,473	1,817	2,022	630	57.71
Disgust	2,897	11,305	1,048	169	326	570	69.29
Fear	1,390	776	7,435	2,318	2,330	1,555	47.05
Sadness	1,035	1,904	1,922	1,1641	1,288	250	64.53
Surprise	966	34	2,364	1,189	10,321	976	65.12
Happiness	203	181	1,547	346	979	11,313	77.65
						Average rate (%)	63.56

TABLE 11
CNN Results for RML

RML	Anger	Disgust	Fear	Sadness	Surprise	Happiness	Recognition rate (%)
Anger	14,375	439	384	477	768	475	84.97
Disgust	466	16,950	833	89	566	195	88.75
Fear	652	632	14198	98	724	1172	81.24
Sadness	414	59	98	16,163	293	399	92.75
Surprise	616	552	1124	198	15,543	877	82.19
Happiness	495	136	887	280	661	14,329	85.35
						Average rate (%)	85.88

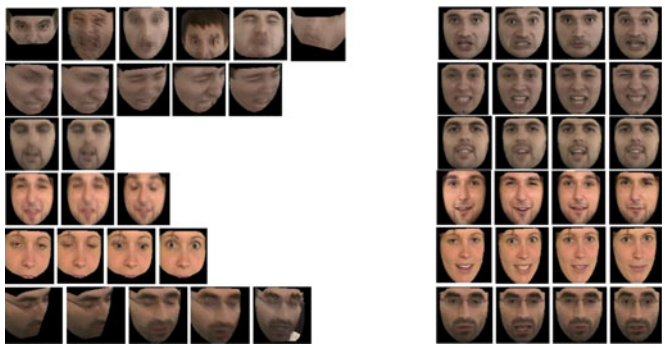
We use SGD Stochastic Gradient Descent (SGD) [62], [63] for the classification. The batch size is 90 images, and the Nvidia Titan X GPU was used for computation. 30 epochs were applied.

The CNN method is applied to the SAVEE, eNTERFACE'05 and RML databases. The final recognition rates are 97.50, 63.20 and 85.88 percent, as shown in Tables 9, 10 and 11, respectively. It can be observed that the CNN results are higher than the geometric approach. It is because CNN uses all the frames, but the geometric method only considers four frames from every video. It can be seen by comparing the results presented in Tables 8 and 10. For example, on the eNTERFACE'05 database, with the geometric approach, the highest average recognition rate has been 41.59 percent, but the CNN has increased it to 63.56 percent, i.e., by 21.97 percent.

4.5 Fusion Results

After training each classifier by using audio and video, we obtain the confidence values for each emotion in every database. We also obtain confidences values from CNN separately. Next, the confidence values are fused in order to train a stacked classifier including SVM and RF, with and without PCA, in order to obtain final emotion prediction.

In the SAVEE database, we recognize seven emotion classes. Six confidence values are available per each emotion label. As we have three sets of data, i.e., audio, visual and CNN, it results in 18 confidence values per sample. They are used as features to train new multiclass SVM and RF classifiers with and without PCA, in a stacked fashion, with the same experimental setup as in the previous experiments. These steps are represented in Fig. 9. The confusion



(a) Corrupted (b) Corrected

Fig. 8. Sample frames which have been misclassified. The images have been taken from [23].

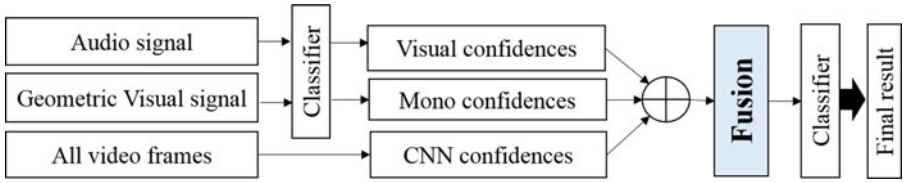


Fig. 9. Data fusion steps.

TABLE 12
Fusion by Using the SVM on the SAVEE Database

SVM	Anger	Disgust	Fear	Sadness	Surprise	Happiness	Neutral	Recognition rate (%)
Anger	60	0	0	0	0	0	0	100.00
Disgust	4	54	0	2	0	0	0	90.00
Fear	0	0	58	1	1	0	0	96.67
Sadness	0	0	0	60	0	0	0	100.00
Surprise	0	0	0	0	60	0	0	100.00
Happiness	0	0	0	0	0	60	0	100.00
Neutral	0	0	0	0	0	0	120	100.00
							Average rate (%)	98.10

TABLE 13
Fusion by Using the SVM–PCA on the SAVEE Database

SVM–PCA	Anger	Disgust	Fear	Sadness	Surprise	Happiness	Neutral	Recognition rate (%)
Anger	60	0	0	0	0	0	0	100.00
Disgust	0	60	0	0	0	0	0	100.00
Fear	0	0	59	0	1	0	0	98.33
Sadness	0	0	0	60	0	0	0	100.00
Surprise	0	0	0	0	60	0	0	100.00
Happiness	0	0	0	0	1	59	0	98.33
Neutral	0	0	0	0	0	0	120	100.00
							Average rate (%)	99.52

TABLE 14
Fusion by Using the RF on the SAVEE Database

RF	Anger	Disgust	Fear	Sadness	Surprise	Happiness	Neutral	Recognition rate (%)
Anger	60	0	0	0	0	0	0	100
Disgust	0	60	0	0	0	0	0	100
Fear	0	0	60	0	0	0	0	100
Sadness	0	0	0	60	0	0	0	100
Surprise	0	0	0	0	60	0	0	100
Happiness	0	0	0	0	0	60	0	100
Neutral	0	0	0	0	0	0	120	100
							Average rate (%)	100

TABLE 15
Fusion by Using the RF–PCA on the SAVEE Database

RF–PCA	Anger	Disgust	Fear	Sadness	Surprise	Happiness	Neutral	Recognition rate (%)
Anger	60	0	0	0	0	0	0	100
Disgust	0	60	0	0	0	0	0	100
Fear	0	0	60	0	0	0	0	100
Sadness	0	0	0	60	0	0	0	100
Surprise	0	0	0	0	60	0	0	100
Happiness	0	0	0	0	0	60	0	100
Neutral	0	0	0	1	0	0	119	99.16
							Average rate (%)	99.88

TABLE 16
Fusion by Using the SVM on the RML Database

SVM	Anger	Disgust	Fear	Sadness	Surprise	Happiness	Recognition rate (%)
Anger	118	1	1	0	0	0	98.33
Disgust	0	120	0	0	0	0	100.00
Fear	1	2	116	0	1	0	96.67
Sadness	1	0	1	117	0	1	97.50
Surprise	0	0	0	1	118	1	98.33
Happiness	0	0	0	0	0	120	100.00
						Average rate (%)	98.47

TABLE 17
Fusion by Using the SVM–PCA on the RML Database

SVM–PCA	Anger	Disgust	Fear	Sadness	Surprise	Happiness	Recognition rate (%)
Anger	119	1	0	0	0	0	99.17
Disgust	0	120	0	0	0	0	100.00
Fear	0	0	117	0	3	0	97.50
Sadness	0	0	1	117	0	2	97.50
Surprise	0	0	0	0	119	1	99.17
Happiness	0	0	0	0	0	120	100.00
						Average rate (%)	98.89

TABLE 18
Fusion by Using the RF on the RML Database

RF	Anger	Disgust	Fear	Sadness	Surprise	Happiness	Recognition rate (%)
Anger	120	0	0	0	0	0	100
Disgust	0	120	0	0	0	0	100
Fear	0	0	119	1	0	0	99.16
Sadness	0	0	1	119	0	0	99.16
Surprise	0	0	0	0	120	0	100
Happiness	0	0	0	0	0	120	100
						Average rate (%)	99.72

TABLE 19
Fusion by Using the RF–PCA on the RML Database

RF-PCA	Anger	Disgust	Fear	Sadness	Surprise	Happiness	Recognition rate (%)
Anger	120	0	0	0	0	0	100
Disgust	0	120	0	0	0	0	100
Fear	0	0	119	0	1	0	99.16
Sadness	0	0	2	118	0	0	98.33
Surprise	0	0	0	0	120	0	100
Happiness	0	0	0	0	0	120	100
						Average rate (%)	99.58

TABLE 20
Fusion by Using the SVM on the eNTERFACE'05 Database

SVM	Anger	Disgust	Fear	Sadness	Surprise	Happiness	Recognition rate (%)
Anger	203	1	0	1	6	1	95.75
Disgust	1	208	1	0	0	1	98.58
Fear	0	6	191	4	9	0	90.95
Sadness	0	0	0	203	8	0	96.21
Surprise	0	4	11	2	190	3	90.48
Happiness	1	2	0	0	2	201	97.57
						Average rate (%)	94.92

TABLE 21
Fusion by Using the SVM–PCA on the eNTERFACE'05 Database

SVM-PCA	Anger	Disgust	Fear	Sadness	Surprise	Happiness	Recognition rate (%)
Anger	211	1	0	0	0	0	99.53
Disgust	1	208	2	0	0	0	98.58
Fear	0	0	205	1	4	0	97.62
Sadness	0	0	0	209	2	0	99.05
Surprise	0	0	5	2	202	1	96.19
Happiness	0	0	0	0	2	204	99.03
						Average rate (%)	98.33

TABLE 22
Fusion by Using the RF on the eNTERFACE'05 Database

RF	Anger	Disgust	Fear	Sadness	Surprise	Happiness	Recognition rate (%)
Anger	210	2	0	0	0	0	99.06
Disgust	0	208	2	0	0	1	98.58
Fear	0	0	206	4	0	0	98.10
Sadness	0	0	2	208	1	0	98.58
Surprise	0	0	1	1	207	1	98.57
Happiness	0	1	0	0	0	205	99.51
						Average rate (%)	98.73

TABLE 23
Fusion by Using the RF–PCA on the eNTERFACE'05 Database

RF-PCA	Anger	Disgust	Fear	Sadness	Surprise	Happiness	Recognition rate (%)
Anger	204	2	0	1	3	2	96.23
Disgust	0	204	2	0	0	5	96.68
Fear	0	0	197	9	4	0	93.81
Sadness	0	0	4	204	3	0	96.68
Surprise	2	2	4	1	197	4	93.81
Happiness	3	3	0	0	1	199	96.60
						Average rate (%)	95.64

matrix of the fusion result for SVM and RF classifiers with and without PCA are represented in Tables 12, 14, 13 and 15, respectively. This procedure is repeated for eNTERFACE'05 and RML with six basic emotions, as well. The fused database for each of them is built by 18 features. There are six confidence values for each of the audio, visual and CNN signals. The results are shown in Tables 16, 17, 18, 19, 20, 21, 22, and 23, respectively.

The fusion results for all three databases and methods are summarized in Table 24. The comparison of the fusion methods for each of the databases is provided in Tables 25, 26, and 27. The best results are obtained by RF, which are 99.72, 98.73 and 100 percent for the SAVEE, eNTERFACE'05 and RML databases, respectively.

As mentioned previously, for multimodal emotion recognition, we fuse the confidence values resulted from the vocal, facial and geometric recognition stages. This results in a higher recognition rate compared to considering a single modality. After extracting the frames from each video, it could be seen that some of them would not add useful information to the system, because of defects, which reduces the recognition rate. A few example frames that have resulted in misclassification are shown in Fig. 8. The misclassification rates for the complement emotion sets are presented in Table 28. The combinations are sorted based on the descending order of the misclassification rates. It can be seen that in a few cases, the currently used key-frames

TABLE 24
Comparison of All the Fusion Results for the Three Databases

Fusion result	SVM	SVM-PCA	RF	RF-PCA
SAVEE	98.1%	99.52%	**100%**	99.88%
RML	98.47%	98.89%	**99.72%**	99.58%
eNTERFACE'05	94.92%	98.33%	**98.73%**	95.64%

TABLE 25
Comparison of all the Fusion Methods' Recognition Rates Based on the SAVEE Database

Emotion recognition system	Recognition rate (%)
Dynamic Bayesian method [28]	96.20
Gaussian method with PCA [29]	99.00
ARTMAP neural network [30]	96.88
Phoneme-Specific [64]	55.60
Our result by SVM	98.10
Our result by SVM with PCA	98.10
Our result by RF	100
Our result by RF with PCA	99.88

TABLE 26
Comparison of All the Fusion Methods' Recognition Rates Based on the eNTERFACE'05 Database

Emotion recognition system	Recognition rate (%)
Hidden Markov model [31]	56.30
Neural networks [32]	67.00
Unified hybrid feature space [55]	71.00
SVM [56]	71.30
KCFA and KCCA [26]	76.00
Bayesian network models [33]	66.54
Combinational method [34]	61.10
Local Phase Quantization [57]	76.40
Our result by SVM	94.92
Our result by SVM with PCA	98.33
Our result by RF	98.73
Our result by RF with PCA	95.64

TABLE 27
Comparison of All the Fusion Methods' Recognition Rates Based on the RML Database

Emotion recognition system	Recognition rate (%)
Deep networks MLP [36]	79.72
Kernel matrix fusion [26]	82.22
BDPCA, LSLDA, OKL and RBF [37]	90.83
Our result by SVM	98.47
Our result by SVM with PCA	98.89
Our result by RF	99.72
Our result by RF with PCA	99.58

cannot distinguish between all emotions successfully. For example, the combination "fear and happiness": has one of the most significant misclassification percentages.

The number of repetitions of each of the label combinations in the list of the first four combinations with the highest misclassification rates are shown in Table 29. One can note that the repetition of fear as a highly misclassified label is higher than the rest of the emotions. Similarly, Table 30 shows the numbers of the repetitions of each of the emotions in the first four combinations with the highest misclassification percentages, as well as their summations. Again, it can be seen that fear has the highest number of repetitions in the combinations with the highest misclassification rates.

5 Conclusion

We presented a system for audio-visual emotion recognition. Audio features included prosodic features, MFCCs and FBEs. Visual features were computed from estimated key-frames representing each video content in terms of representative facial expressions. Visual data was described both by using geometric features and by means of a CNN-based model. Four types of classification methods were used, i.e., multiclass SVM and RF with and without applying PCA. After training the first classifier for each set separately, the output confidence values were fused to define a new feature vector that was learnt by a second level classifier-with the same type as the first level—in order to obtain the final classification prediction. Experimental results were based on three different databases, namely, SAVEE, eNTERFACE'05 and RML. The RF classifier showed the best performance over all the databases. The recognition rates on the mentioned databases were 99.72, 98.73 and 100 percent, respectively. They showed improvements compared to previous state-of-the-art results on the same databases and modalities, by 0.72, 22.33 and 9.17 percent, respectively. Fear was the most repeatedly misclassified label. For future research, we plan to extend the set of key-frames to allow the work to cover additional characteristics of the emotion videos in order to better discriminate between fear and happiness, as well as between anger and disgust. In the same way, we plan to extend the CNN part of the model to include additional temporal information in the model by means of 3D convolutions and RNN-LSTM [65].

TABLE 28
Misclassificationn Percentage of Used Key-Frames by CNN Method for Each Pair Label on a Special Database

Label	SAVEE	Label	RML	Label	eNTERFACE'05
F + H	1.9703	F + H	5.9949	A + D	15.0394
A + D	1.356	F+ SU	5.0434	F+ SU	14.829
D+ F	0.5196	SU + H	4.2875	F + SA	12.6606
SA+ SU	0.2985	D+ F	3.9889	F + H	10.2289
A + SU	0.2408	A + SU	3.8985	A + SU	8.1467
F + SA	0.2311	A + F	3.0003	A + F	8.1124
A + H	0.2262	D + SU	2.9413	A + SA	7.451
D + H	0.0915	A + H	2.8781	SA+ SU	7.3206
A + F	0.0594	A + SA	2.5976	SU + H	6.4387
F+ SU	0.039	A + D	2.5174	D + SA	5.7951
A + SA	0.0222	SA + H	1.9788	D+ F	5.6668
D + SU	0.0154	SA+ SU	1.3642	D + H	2.368
D + SA	0.0108	D + H	0.9155	A + H	2.2855
SA + H	0	F + SA	0.5616	SA + H	1.8804
SU + H	0	D + SA	0.4023	D + SU	1.1063

{F: Fear, D: Disgust, H: Happiness, Sadness: SA, Surprise: SU and anger: A}

TABLE 29
The Numbers of Repetitions of the Label Combinations with the Highest Misclassification Rates in the Three Databases

Label combination (%)	F + H	A + D	D + F	F + SU	SU + H	SA + SU	F + SA
Repetition	3	2	2	2	1	1	1

TABLE 30
The Numbers of Repetitions of the Labels in the Combinations with the Highest Misclassification Rates, in Each Database, and in Total

	Fear	Happiness	Anger	Disgust	Sadness	Surprise
SAVEE	2	1	1	2	1	1
RML	3	2	0	1	0	2
eNTERFACE'05	3	1	1	1	1	1
Summation	8	4	2	4	2	4

Acknowledgments

This work has been partially supported by the Estonian Research Grant (PUT638), Spanish projects TIN2013-43478-P and TIN2016-74946-P and the European Commission Horizon 2020 granted project SEE.4C under call H2020-ICT-2015,

and the Estonian Centre of Excellence in IT (EXCITE) funded by the European Regional Development Fund and the European Network on Integrating Vision and Language (iV&L Net) ICT COST Action IC1307. The authors also gratefully acknowledge the support of NVIDIA Corporation with the donation of the Titan X Pascal GPU.

REFERENCES

[1] K. Kim, Y.-S. Cha, J.-M. Park, J.-Y. Lee, and B.-J. You, "Providing services using network-based humanoids in a home environment," *IEEE Trans. Consum. Electron.*, vol. 57, no. 4, pp. 1628–1636, Nov. 2011.

[2] A. Zaraki, D. Mazzei, M. Giuliani, and D. De Rossi, "Designing and evaluating a social gaze-control system for a humanoid robot," *IEEE Trans. Human-Mach. Syst.*, vol. 44, no. 2, pp. 157–168, Apr. 2014.

[3] I. Lüsi, et al., "Joint challenge on dominant and complementary emotion recognition using micro emotion features and head-pose estimation: Databases," in *Proc. 12th IEEE Int. Conf. Autom. Face Gesture Recognit.*, 2017, pp. 1–6.

[4] F. Noroozi, M. Marjanovic, A. Njegus, S. Escalera, and G. Anbarjafari, "Fusion of classifier predictions for audio-visual emotion recognition," in *Proc. 23rd Int. Conf. Pattern Recognit.*, 2016, pp. 61–66.

[5] Z. Lv, S. Feng, L. Feng, and H. Li, "Extending touch-less interaction on vision based wearable device," in *Proc. IEEE Virtual Reality*, 2015, pp. 231–232.

[6] Z. Zhang, E. Coutinho, J. Deng, and B. Schuller, "Cooperative learning and its application to emotion recognition from speech," *IEEE/ACM Trans. Audio Speech Language Process.*, vol. 23, no. 1, pp. 115–126, Jan. 2015.

[7] E. Russell, A. Stroud, J. Christian, D. Ramgoolam, and A. B. Williams, "SMILE: A portable humanoid robot emotion interface," in *Proc. 9th ACM/IEEE Int. Conf. Human-Robot Interaction Workshop Appl. Emotional Robots*, 2014, pp. 1–5.

[8] J. Torous, R. Friedman, and M. Keshavan, "Smartphone ownership and interest in mobile applications to monitor symptoms of mental health conditions," *JMIR mHealth uHealth*, vol. 2, no. 1, 2014, Art. no. e2.

[9] M. S. Hossain and G. Muhammad, "Cloud-assisted speech and face recognition framework for health monitoring," *Mobile Netw. Appl.*, vol. 20, no. 3, pp. 391–399, 2015.

[10] M. Szwoch and W. Szwoch, "Emotion recognition for affect aware video games," in *Image Processing & Communications Challenges*. Berlin, Germany: Springer, 2015, pp. 227–236.

[11] D. Bolinger and D. L. M. Bolinger, *Intonation and its Uses: Melody in Grammar and Discourse*. Stanford, CA: Stanford Univ. Press, 1989.

[12] T. Wu, S. Fu, and G. Yang, "Survey of the facial expression recognition research," in *Proc. Int. Conf. Brain Inspired Cognitive Syst.*, 2012, pp. 392–402.

[13] K. Anderson and P. W. McOwan, "A real-time automated system for the recognition of human facial expressions," *IEEE Trans. Syst. Man Cybern. Part B (Cybern.)*, vol. 36, no. 1, pp. 96–105, Feb. 2006.

[14] N. Ramakrishnan, T. Srikanthan, S. K. Lam, and G. R. Tulsulkar, "Adaptive window strategy for high-speed and robust KLT feature tracker," in *Proc. Pacific-Rim Symp. Image Video Technol.*, 2015, pp. 355–367.

[15] Y. N. Chae, T. Han, Y.-H. Seo, and H. S. Yang, "An efficient face detection based on color-filtering and its application to smart devices," *Multimedia Tools Appl.*, vol. 75, pp. 1–20, 2016.

[16] C. Ding, J. Choi, D. Tao, and L. S. Davis, "Multi-directional multi-level dual-cross patterns for robust face recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 3, pp. 518–531, Mar. 2016.

[17] H. Dibeklioğlu, F. Alnajar, A. A. Salah, and T. Gevers, "Combining facial dynamics with appearance for age estimation," *IEEE Trans. Image Process.*, vol. 24, no. 6, pp. 1928–1943, Jun. 2015.

[18] G. Hemalatha and C. Sumathi, "A study of techniques for facial detection and expression classification," *Int. J. Comput. Sci. Eng. Survey*, vol. 5, no. 2, 2014, Art. no. 27.

[19] M. Pantic and M. S. Bartlett, *Machine Analysis of Facial Expressions*. Vienna, Austria: I-Tech Education and Publishing, 2007.

[20] D. Kamińska, T. Sapiński, and G. Anbarjafari, "Efficiency of chosen speech descriptors in relation to emotion recognition," *EURASIP J. Audio Speech Music Process.*, vol. 2017, no. 1, 2017, Art. no. 3.

[21] F. Noroozi, T. Sapiński, D. Kamińska, and G. Anbarjafari, "Vocal-based emotion recognition using random forests and decision tree," *Int. J. Speech Technol.*, vol. 25, pp. 1–8, 2017.

[22] P. Jackson and S. Haq, "Surrey audio-visual expressed emotion (savee) database," 2015, http://kahlan.eps.surrey.ac.uk/savee/

[23] O. Martin, I. Kotsia, B. Macq, and I. Pitas, "The enterface'05 audio-visual emotion database," in *Proc. 22nd Int. Conf. Data Eng. Workshops*, 2006, pp. 8–8.

[24] Z. Xie, "Ryerson multimedia research laboratory (RML)," 2010, http://www.rml.ryerson.ca/rml-emotion-database.html

[25] S. E. Kahou, et al., "EmoNets: Multimodal deep learning approaches for emotion recognition in video," *J. Multimodal User Interfaces*, vol. 10, pp. 1–13, 2015.

[26] Y. Wang, L. Guan, and A. N. Venetsanopoulos, "Kernel cross-modal factor analysis for information fusion with application to bimodal emotion recognition," *IEEE Trans. Multimedia*, vol. 14, no. 3, pp. 597–607, Jun. 2012.

[27] M. Wimmer, B. Schuller, D. Arsic, G. Rigoll, and B. Radig, "Low-level fusion of audio, video feature for multi-modal emotion recognition," in *Proc. 3rd Int. Conf. Comput. Vis. Theory Appl.*, 2008, pp. 145–151.

[28] F. Cid, L. J. Manso, and P. Núnez, "A novel multimodal emotion recognition approach for affective human robot interaction," *Proc. FinE*, pp. 1–9, 2015.

[29] S. Haq, T. Jan, A. Jehangir, M. Asif, A. Ali, and N. Ahmad, "Bimodal human emotion classification in the speaker-dependent scenario," *Pakistan Academy Sci.*, Art. no. 27, vol. 52, no. 1, pp. 27–38, 2015.

[30] D. Gharavian, M. Bejani, and M. Sheikhan, "Audio-visual emotion recognition using FCBF feature selection method and particle swarm optimization for fuzzy ARTMAP neural networks," *Multimedia Tools Appl.*, vol. 76, pp. 1–22, 2016.

[31] D. Datcu and L. Rothkrantz, "Multimodal recognition of emotions in car environments," presented at the *Driver Car Interaction Interface*, Praag, Czech Republic, 2009.

[32] M. Paleari, B. Huet, and R. Chellali, "Towards multimodal emotion recognition: A new approach," in *Proc. ACM Int. Conf. Image Video Retrieval*, 2010, pp. 174–181.

[33] D. Jiang, Y. Cui, X. Zhang, P. Fan, I. Ganzalez, and H. Sahli, "Audio visual emotion recognition based on triple-stream dynamic Bayesian network models," in *Proc. Int. Conf. Affective Comput. Intell. Interaction*, 2011, pp. 609–618.

[34] K.-C. Huang, H.-Y. S. Lin, J.-C. Chan, and Y.-H. Kuo, "Learning collaborative decision-making parameters for multimodal emotion recognition," in *Proc. IEEE Int. Conf. Multimedia Expo*, 2013, pp. 1–6.

[35] A. Gera and A. Bhattacharya, "Emotion recognition from audio and visual data using F-score based fusion," in *Proc. 1st IKDD Conf. Data Sci.*, 2014, pp. 1–10.

[36] C. Fadil, R. Alvarez, C. Martínez, J. Goddard, and H. Rufiner, "Multimodal emotion recognition using deep networks," in *Proc. 6th Latin Amer. Congr. Biomed. Eng.*, 2015, pp. 813–816.

[37] K. Seng, L.-M. Ang, and C. Ooi, "A combined rule-based and machine learning audio-visual emotion recognition approach," *IEEE Trans. Affective Comput.*, no. 99, p. 1, 2016, doi: 10.1109/TAFFC.2016.2588488.

[38] S.-M. Guo, Y. Pan, Y.-C. Liao, C. Hsu, J. S. H. Tsai, and C. Chang, "A key frame selection-based facial expression recognition system," in *Proc. 1st Int. Conf. Innovative Comput. Inf. Control-Volume I*, 2006, pp. 341–344.

[39] A. R. Doherty, D. Byrne, A. F. Smeaton, G. J. Jones, and M. Hughes, "Investigating keyframe selection methods in the novel domain of passively captured visual lifelogs," in *Proc. Int. Conf. Content-Based Image Video Retrieval*, 2008, pp. 259–268.

[40] Q. Zhang, S.-P. Yu, D.-S. Zhou, and X.-P. Wei, "An efficient method of key-frame extraction based on a cluster algorithm," *J. Human Kinetics*, vol. 39, no. 1, pp. 5–14, 2013.

[41] Y. Zhuang, Y. Rui, T. S. Huang, and S. Mehrotra, "Adaptive key frame extraction using unsupervised clustering," in *Proc. Int. Conf. Image Process.*, 1998, pp. 866–870.

[42] S. Hasebe, M. Nagumo, S. Muramatsu, and H. Kikuchi, "Video key frame selection by clustering wavelet coefficients," in *Proc. 12th Eur. Signal Process. Conf.*, 2004, pp. 2303–2306.

[43] S. Planet and I. Iriondo, "Comparison between decision-level and feature-level fusion of acoustic and linguistic features for spontaneous emotion recognition," in *Proc. 7th Iberian Conf. Inf. Syst. Technol.*, 2012, pp. 1–6.

[44] S. Haq and P. J. Jackson, "Multimodal emotion recognition," *Mach. Audition Principles Algorithms Syst.*, pp. 398–423, 2010.
[45] F. Noroozi, T. Sapiński, D. Kamińska, and G. Anbarjafari, "Vocal-based emotion recognition using random forests and decision tree," *Int. J. Speech Technol.*, vol. 20, no. 2, pp. 239–246, 2017.
[46] I. Lüsi, S. Escarela, and G. Anbarjafari, "Human head pose estimation on SASE database using random Hough regression forests," in *Proc. Int. Workshop Face Facial Expression Recognit. Real World Videos*, 2016, pp. 137–150.
[47] S. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Trans. Acoust. Speech Signal Process.*, vol. 28, no. 4, pp. 357–366, Aug. 1980.
[48] I. Bocharova, *Compression for Multimedia*. Cambridge, U.K.: Cambridge Univ. Press, 2010.
[49] S. K. Kopparapu and M. Laxminarayana, "Choice of Mel filter bank in computing MFCC of a resampled speech," in *Proc. 10th Int. Conf. Inf. Sci. Signal Process. Appl.*, 2010, pp. 121–124.
[50] A. J. Calder, A. W. Young, J. Keane, and M. Dean, "Configural information in facial expression perception," *J. Exp. Psychology Human Perception Perform.*, vol. 26, no. 2, 2000, Art. no. 527.
[51] M. G. Calvo, A. Fernández-Martín, and L. Nummenmaa, "Facial expression recognition in peripheral versus central vision: Role of the eyes and the mouth," *Psychological Res.*, vol. 78, no. 2, pp. 180–195, 2014.
[52] S. Lindsay, "A tutorial on princial components analysis," vol. 51, no. 52, p. 65, 2002.
[53] C. Cortes and V. Vapnik, "Support-vector networks," *Mach. Learn.*, vol. 20, no. 3, pp. 273–297, 1995.
[54] L. B. Statistics and L. Breiman, "Random forests," vol. 45, pp. 5–32, 2001.
[55] M. Mansoorizadeh and N. M. Charkari, "Multimodal information fusion application to human emotion recognition from face and speech," *Multimedia Tools Appl.*, vol. 49, no. 2, pp. 277–297, 2010.
[56] R. Gajsek, V. Štruc, and F. Mihelic, "Multi-modal emotion recognition using canonical correlations and acoustic features," in *Proc. Int. Conf. Pattern Recognit.*, 2010, pp. 4133–4136.
[57] S. Zhalehpour, Z. Akhtar, and C. E. Erdem, "Multimodal emotion recognition with automatic peak frame selection," in *Proc. IEEE Int. Symp. Innovations Intell. Syst. Appl. Proc.*, 2014, pp. 116–121.
[58] Surrey audio-visual expressed emotion (SAVEE) database, (2015). [Online]. Available: http://kahlan.eps.surrey.ac.uk/savee/Database.html, Accessed on: Apr. 8, 2017.
[59] RML emotion database, (2010). [Online]. Available: http://www.rml.ryerson.ca/rml-emotion-database.html, Accessed on: Apr. 7, 2017.
[60] C. Szegedy, et al., "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 1–9.
[61] R. Rothe, R. Timofte, and L. Van Gool, "Deep expectation of real and apparent age from a single image without facial landmarks," *Int. J. Comput. Vis.*, pp. 252–257, 2016.
[62] L. Bottou, "Large-scale machine learning with stochastic gradient descent," in *Proc. Int. Conf. Comput. Statist.*, 2010, pp. 177–186.
[63] Q. V. Le, "A tutorial on deep learning part 1: Nonlinear classifiers and the backpropagation algorithm," 2015, https://cs.stanford.edu/~quocle/
[64] Y. Kim and E. Mower Provost, "Say cheese versus smile: Reducing speech-related variability for facial emotion recognition," in *Proc. 22nd ACM Int. Conf. Multimedia*, 2014, pp. 27–36.
[65] Q. Li, Z. Qiu, T. Yao, T. Mei, Y. Rui, and J. Luo, "Action recognition by learning deep multi-granular spatio-temporal video representation," in *Proc. ACM Int. Conf. Multimedia Retrieval*, 2016, pp. 159–166.

Fatemeh Noroozi received the BSc degree in computer engineering, software, from Shiraz University, Iran. She received the MSc degree in mechatronics engineering from the University of Tehran, Iran. Her thesis was entitled "Modeling of Virtual Organizations Integrating on Physical Core based on a Service-oriented Architecture". Her thesis was entitled "Developing a Real-time Virtual Environment for Connecting to a Touching Interface in Dental Applications". Currently, she is working toward the PhD degree at the University of Tartu, Estonia, working on "Multimodal Emotion Recognition based Human-robot Interaction Enhancement". She is a student member of the IEEE.

Marina Marjanovic received the graduate degree from the Faculty of Electrical Engineering of Belgrade University, Serbia in 2002. In 2007, she received the PhD diploma in the field of telecommunications (Signal Processing) with honors from the Polytechnic University of Madrid. During the PhD degree, she received a scholarship from the Spanish company "Telefonica Moviles". Since October 2013, she has been an assistant professor in the School of Electrical Engineering and Computing, Singidunum University. Her areas of expertise are: Signal processing, wireless communications, machine learning, measurements and instrumentations. She is a member of the IEEE.

Angelina Njegus received the BSc and MSc degrees in information systems from the Faculty of Organizational Sciences, University of Belgrade, and the PhD degree in information systems from the Faculty of Business Studies, Belgrade, Serbia in 1994, 1999, and 2003, respectively. She is currently an associate professor with Singidunum University, Serbia, where she teaches courses in information systems analysis and design, intelligence systems, and big data analytics systems. Her recent research is mainly on affective human computer interaction and recognition systems.

Sergio Escalera received the PhD degree in Multiclass visual categorization systems from the Computer Vision Center, Universitat de Barcelona (UAB). He is an associate professor in the Department of Mathematics and Informatics, Universitat de Barcelona. He is an adjunct professor with Universitat Oberta de Catalunya, Aalborg University, and Dalhousie University. He received the 2008 best Thesis award on Computer Science at Universitat Autònoma de Barcelona. He leads the Human Pose Recovery and Behavior Analysis Group at UB and CVC. He is also a member of the Computer Vision Center at Campus UAB. He is an expert in human behavior analysis in temporal series, statistical pattern recognition, visual object recognition, and HCI systems, with special interest in human pose recovery and behavior analysis from multi-modal data. He is vice-president of ChaLearn Challenges in Machine Learning, leading ChaLearn Looking at People events. He is vice-chair of IAPR TC-12: Multimedia and visual information systems.

Gholamreza Anbarjafari heads the intelligent computer vision (iCV) research group in the Institute of Technology at the University of Tartu. He is also deputy scientific coordinator of he European Network on Integrating Vision and Language (iV&L Net) ICT COST Action IC1307. He is an associate editor and guest lead editor of several journals, Special Issues and Book projects. He is a senior member of the IEEE and the vice chair of Signal Processing / Circuits and Systems/Solid-State Circuits Joint Societies Chapter of IEEE Estonian section. He has got Estonian Research Council Grant (PUT638) in January 2015 and has been involved in many international industrial projects such as AIDesign, A.G.E., iRental and Virtual Fitting Room by Fits. Me Rakutan. He is expert in computer vision, human-robot interaction, graphical models and artificial intelligence. His work in image super resolution has been selected for the best paper award in 2012 by Electronics and Telecommunications Research Institute (ETRI) Journal, South Korea. He has supervised 9 MSc students and 5 PhD students. He has published more than 100 scientific works. He has been in the organizing committee and technical committee of IEEE Signal Processing and Communications Applications Conference in 2013, 2014 and 2016 and TCP of conferences such as ICOSST, ICGIP, SampTA and SIU. He has been organizing challenges and workshops in FG17, CVPR17, and ICCV17.