Neural Comput & Applic (2014) 24:399–412
DOI 10.1007/s00521-012-1228-3

ORIGINAL ARTICLE

Audiovisual emotion recognition using ANOVA feature selection method and multi-classifier neural networks

Mahdi Bejani · Davood Gharavian · Nasrollah Moghaddam Charkari

Received: 15 April 2012 / Accepted: 15 October 2012 / Published online: 1 November 2012

Abstract To make human–computer interaction more naturally and friendly, computers must enjoy the ability to understand human's affective states the same way as human does. There are many modals such as face, body gesture and speech that people use to express their feelings. In this study, we simulate human perception of emotion through combining emotion-related information using facial expression and speech. Speech emotion recognition system is based on prosody features, mel-frequency cepstral coefficients (a representation of the short-term power spectrum of a sound) and facial expression recognition based on integrated time motion image and quantized image matrix, which can be seen as an extension to temporal templates. Experimental results showed that using the hybrid features and decision-level fusion improves the outcome of unimodal systems. This method can improve the recognition rate by about 15 % with respect to the speech unimodal system and by about 30 % with respect to the facial expression system. By using the proposed multi-classifier system that is an improved hybrid system, recognition rate would increase up to 7.5 % over the hybrid features and decision-level fusion with RBF, up to 22.7 % over the speech-based system and up to 38 % over the facial expression-based system.

Keywords Human–computer interaction · Audiovisual emotion recognition · Feature-level fusion · Decision-level fusion · Multi-classifier

1 Introduction

Humans communicate with each other far more naturally than they do with computers. One of the main problems in human–computer interaction (HCI) systems is the transmission of implicit information. To make HCI more naturally and friendly, computers must enjoy the ability to understand human's affective states the same way as human does.

In the recent years, the emotion recognition has found many applications such as medical-emergency domain to detect stress and pain [1], interactions with robots [2, 3], computer games [4] and developing man–machine interfaces (MMI) for helping weak and old people [5].

There are many modals such as face, body gesture and speech that people use to express their feelings. Combinations of these modals depend on the place they occur and on the subjects themselves; therefore, we have a wide variety of patterns for combining [6].

Some studies in psychology and linguistics confirm the relation between affective displays and specific audio and visual signals [7, 8].

Mehrabian [9] has stated that there are basically three elements in any face-to-face communication. His studies indicated that facial expression and speech articulations in the visual channel are the most important affective cue

M. Bejani
Islamic Azad University, South Tehran Branch,
Tehran, Iran
e-mail: St_m_bejani@azad.ac.ir

D. Gharavian (✉)
EE Department, Shahid Abbaspour University,
Tehran, Iran
e-mail: gharavian@pwut.ac.ir

N. M. Charkari
Distributed Processing Lab, Tarbiat Modares University,
Tehran, Iran
e-mail: charkari@modares.ac.ir

(55 and 38 %, respectively), and words contribute only 7 % of the overall impression.

There are some approaches for quantifying and measuring emotions such as discrete categories and dimensional description [10]. In this work, we used discrete emotion categories including happiness, fear, sadness, anger, surprise and disgust. With universal emotion models, it is easy to recognize emotional states [11].

Three main fusion approaches used in the literature are feature-level fusion, decision-level fusion (combining classifiers) and hybrid of the both methods (model-level fusion) [6]. Hybrid fusion approach aims to combine the benefits of both the feature-level and decision-level fusion methods. This method may be a good choice for fusion problem. Our proposed multi-classifier, as an improved hybrid system, uses the strengths and weaknesses of individual speech and facial expression systems.

The goal of this paper is to simulate human perception of emotions by combining emotion-related information from facial expression and speech. So, in this work, we use different ways to combine audio-based and facial expression systems.

The remainder of this paper is organized as follows. Section 2 reviews the recent researches in this field. Section 3 presents the audio and visual systems and combination of them in different ways. Section 4 describes feature selection method that is used to select more relevant features for emotion recognition. Section 5 contains the experimental results. Section 6 shows the result of our proposed multi-classifier system. Finally, conclusions are drawn in Sect. 7.

2 Background and related works

Recently, audiovisual-based emotion recognition methods started to attract the attention of the research community. In the survey of Pantic and Rothkrantz [12] in 2000, only four studies were found to focus on audiovisual affect recognition. Since then, affect recognition, using audio and visual information, has been the subject of many researches. The most updated survey on affect recognition methods for audio, visual and spontaneous expressions has been carried out by Zeng et al. [11] in 2009. Here, some main works in this field are pointed out in brief.

De Silva and Pei Chi [13] used a rule-based method for decision-level fusion of speech- and visual-based systems. In speech, pitch was extracted as the feature and used in the nearest-neighbor classification method. In video, they tracked facial points with optical flow, and hidden Markov model (HMM) was trained as the classifier. The decision-level fusion improved the result of the individual systems.

Song et al. [14] used a tripled hidden Markov model (THMM) to model joint dynamics of the three signals perceived from the subject: pitch and energy as speech features; motion of eyebrow, eyelid and cheek as facial expression features; and lips and jaw as visual speech signals. The proposed THMM architecture was tested for seven basic emotions (surprise, anger, joy, sadness, disgust, fear and neutral), and its overall performance was 85 %.

Mansoorizadeh and Moghaddam Charkari [6] compared feature-level and decision-level fusion of speech and face information. They proposed an asynchronous feature-level fusion approach that improves the result of combination. For speech analysis, they used the features related to energy and pitch contour. For face analysis, the features representing the geometric characteristic of face area were used. The multimodal results showed an improvement over both of the individual systems. This result shows that hybrid fusion is a good choice for audiovisual combination.

Hoch et al. [15] presented an algorithm for bimodal emotion recognition. They used a weighted linear combination for decision-level fusion of speech and facial expression systems. They also applied a database of 840 audiovisual samples with seven speakers and three emotions. Their system classifies three emotions (positive, negative and neutral) with an average of 90.7 % recognition rate. By using a fusion model based on a weighted linear combination, the performance improvement becomes nearly 4 % compared to that of unimodal emotion recognition.

Wang and Guan [16] proposed the use of cascade audio and visual feature data to classify variant emotions. They built one-against-all (OAA) linear discriminate analysis (LDA) classifiers for each emotion state and set two rules in the decision module with several multi-class and binary classifiers to recognize emotions.

Paleari et al. [17] presented semantic affect-enhanced multimedia indexing (SAMMI) to extract real-time emotion appraisals from non-prototypical person-independent facial expressions and vocal prosody. Different probabilistic methods for fusion were compared and evaluated with a novel fusion technique called NNET. The performance has been measured using the standard precision and recall metrics, in particular the mean average precision (MAP) for the first 33 % of the responses and the positive classification rate (CR+). The results showed that NNET can improve the recognition score (CR+) by about 19 % and the MAP by about 30 % with respect to the best unimodal system.

The interdependency and correlation of the affective features are of the main advantages of feature-level fusion. The main problem of this approach is ignoring the differences in temporal structure. On the other hand, decision-level fusion ignores metrics and the above correlation as well as complementary role of the modalities [6].

According to some reports, hybrid fusion that aims at combining the benefits of both feature-level and decision-level fusion methods may be a good choice for fusion problem [6, 11]. Here, we set two experiments in hybrid fusion method. Stacked generalization method was used to fuse the output of the feature-level and decision-level ensembles. The output of the feature-level and decision-level ensembles was fed as a feature vector to MLP and RBF neural networks.

In recent years, research is focused on finding reliable informative features and combining powerful classifiers in order to improve the performance of emotion detection systems in real-life applications [12, 16, 18–28]. In this way, developing optimal design methods for combining classifiers is an active research field. Here, we propose a multi-classifier approach that improves the emotion recognition results as compared to the speech-based and facial expression systems. This proposed system is an improved hybrid system.

3 Methodology

Emotional states were recognized by the use of three different systems based on speech, facial expression and bimodal information. Speech emotion recognition system is based on mel-frequency cepstral coefficient (MFCC), pitch, energy and formant features, and facial expression recognition is based on ITMI and QIM images.

The main goal of the present work is to quantify the performance of speech-based and facial expression systems, recognize the strengths and weaknesses of these systems and compare different ways to combine these two modalities to increase the performance of the system. Figure 1 sketches an overview of the proposed recognition system. In the following, we have described the details of this hybrid system.

3.1 Speech-based system

The most widely used speech cues for audio emotion recognition are global-level prosodic features such as the statistics of pitch and intensity. Due to large number of features at the frame level, the mean value of features over a specified sentence was used for training and testing of this system. Therefore, in this research, the means, the standard deviations, the maximum values and the minimum values of the pitch and the energy were computed using Praat speech processing software [29].

In addition, MFCC was computed using Praat. MFCCs are a popular and powerful analytical tool in the field of speech recognition. In this work, we took the first 12 coefficients as the useful features. The mean, standard deviation, maximum and minimum of MFCC features were calculated, which produced a total number of 48 MFCC features.

Formant frequencies are the properties of the vocal tract system. In this paper, the first three formant frequencies and their bandwidths were calculated using Praat. The mean, standard deviation, maximum and minimum of formant features were calculated, which produced a total number of 24 formant features. In total, we extracted 80 features from speech and used them for emotion recognition.

3.2 Facial expression recognition system

For video databases, one of the important methods for describing video scene is applying space and time relation between the objects in the scene. In this paper, facial expression recognition system was based on ITMI and QIM images, which is an extension to the temporal templates introduced by Bobick and Davis [30].

Temporal templates are 2D images constructed from image sequences, which show motion history (i.e., where and when the motion in the image sequence has occurred) and reduce a 3D spatiotemporal space into a 2D representation. They are able to eliminate one dimension while retaining the temporal information; the locations where movement has occurred in an input image sequence are depicted in the related 2D image [31].

A typical stacking frame for spatiotemporal knowledge representation has been presented in [32]. In this technique, few frames of one action are combined, resulting in a kind of temporal smoothing. Combination may be performed in gray-level or transformed domain. Also, spatio-smoothing using known image filters and adding consecutive frames is a type of spatiotemporal database, which has been applied in lip reading for speech recognition [33]. In [34], motion history image (MHI) and motion flow history (MFH) are presented. MHI template includes the time of occurrence of motion, but direction of the motion is not saved:

$$\mathrm{MHI}(k,l) = \begin{cases} \tau, & \text{if } \left|m_x^{kl}(\tau)\right| + \left|m_y^{kl}(\tau)\right| \neq 0 \\ 0, & \text{elsewhere} \end{cases} \quad (1)$$

where τ is the time of action occurrence, (k, l) is the position of action occurrence in the image, and $m_x^{kl}(\tau)$ and $m_y^{kl}(\tau)$ are the components of motion vector at time τ and position (k, l) in x and y directions, respectively.

MFH includes the position and direction of action as follows:

$$\mathrm{MFH}_d(k,l) = \begin{cases} m_d^{kl}(\tau), & \text{if } E\left[m_d^{kl}(\tau)\right] < T \\ M\left(m_d^{kl}(\tau)\right), & \text{elsewhere} \end{cases} \quad (2)$$

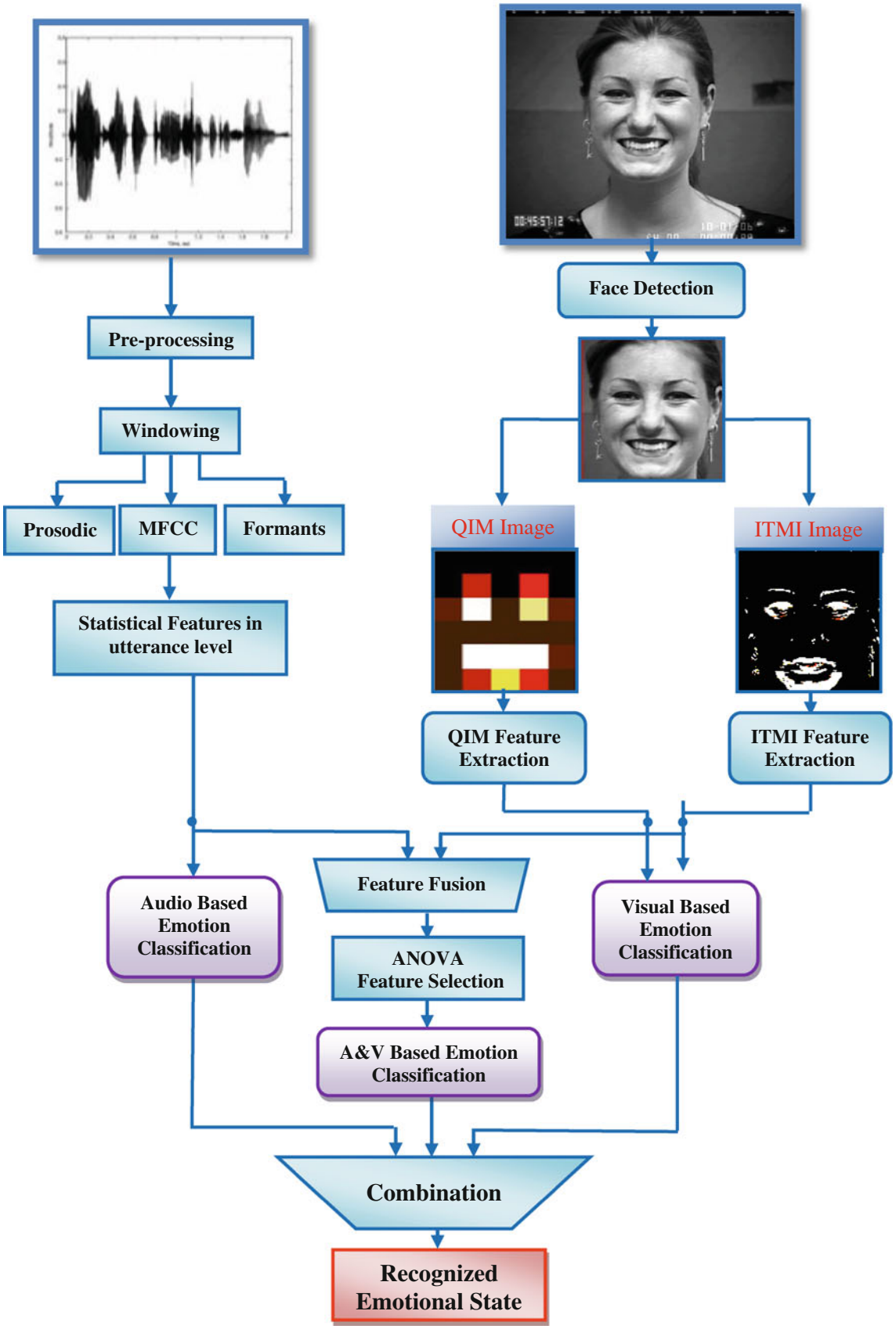

Fig. 1 Overview of the emotion recognition system

where

$$E\left[m_d^{kl}(\tau)\right] = \left\|m_d^{kl}(\tau) - \text{med}\left(m_d^{kl}(\tau), \ldots, m_d^{kl}(\tau) - \alpha\right)\right\|$$
$$M\left(m_d^{kl}(\tau)\right) = \text{med}\left(m_d^{kl}(\tau), \ldots, m_d^{kl}(\tau) - \alpha\right) \quad (3)$$

In the above equation, α is the number of old frames, which is set to between 3 and 5.

MFH and MHI are complementary temporal templates because they include spatial, temporal and directional information. In MHI, repeated motions in the same position in different times give similar results. This is a problem in storing the occurrence time of an action. This paper proposes a spatiotemporal representation that includes storing the occurrence time of each motion with an emphasis on the final action. We used spatiotemporal database in human motion recognition. Integrated time motion image (ITMI) introduced by Sadoghi Yazdi et al. [35] was used at time t and location (k, l) as follows:

$$\text{ITMI}_T(k,l) = \begin{cases} \frac{(\text{ITMI}_i(k,l)+\text{id}(k,l))}{N} & \text{if } |d(k,l)| > T \\ 0, & \text{elsewhere} \end{cases} \quad (4)$$

where i is the frame number, (k, l) is the position of action occurrence in the image, and $d(k, l)$ is the difference between frame i and primary frame [35].

T is the threshold used for motion detection, which is considered 30 in facial detection. In ITMI calculation, an average smoothing is done in order to reduce noise. The primary value for ITMI is zero [$\text{ITMI}_0\ (k, l) = 0$].

ITMI is normalized and sequence duration does not affect on it. Any change in one second is effective in ITMI calculation, and in spite of MHI calculation, pervious motion effects are still considered.

In this method, all durations are summed for each motion. The value of each motion is its frame number, and the final result is normalized to the sequence length.

ITMI is a kind of spatiotemporal database and shows the motion's history, that is, where and when the motion in the image sequence has occurred.

Adding all the events of each motion to this database, we can have more data for constructing a good database. For doing less calculation and considering the effect of unwanted motions, we use image quantization and come to a quantized matrix for motion repetition.

Quantized image matrix (QIM) increases when any pixel has $|d(k,l)| > T$:

$$\text{QIM}_t(m,n) = \text{QIM}_{t-1}(m,n) + 1 \quad (5)$$

where (k, l) is the position of action occurrence in the image, which is placed in one of the $m \times n$ regions. m and n are the number of regions that the image has divided and are set to 6 and 5, respectively.

For extracting facial expression features based on ITMI and QIM images, we should detect face and then extract the required features from the ITMI and QIM images.

3.2.1 Face detection

The first step in designing a facial expression recognition system is detecting the user's face inside the scene.

Many different techniques have been tried so far in order to solve the problem of detecting a face in a scene. In this paper, OpenCV [36] face tracker was used as an open-source implementation of a boosted face tracker. This tracker was trained on a large database of face/non-face images and produced efficient face detection in all kinds of settings, thus completely fitting to our needs [15]. Figure 2 shows the face tracker results for a sample image.

3.2.2 Features extracted from ITMI

First, we explain the features extracted from ITMI. Figure 3 shows an example of the last frame of surprise and happiness and their ITMIs.

Fig. 2 Output of the OpenCV face tracker [44]

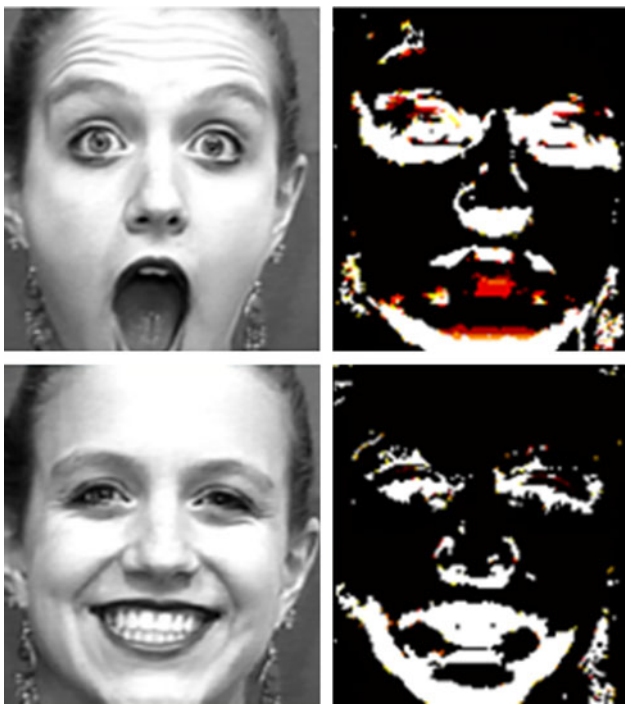

Fig. 3 Last frame of surprise and happiness and their ITMI [44]

Five features were extracted from the ITMI. The first feature extracted from the ITMI is to obtain the upper ITMI total energy to its lower half. As shown in Fig. 3, happy has asymmetric ITMI and surprise has symmetric ITMI.

Features 2–5 are a kind of action unit. An ITMI is divided into four equal horizontal regions. Average of surfaces is extracted as a feature. Figure 4 depicts these regions. These four mentioned features represent changes in face in the forehead, eyes and eyebrows, nose and mouth and chin, respectively.

3.2.3 Features extracted from QIM image

As said before, QIM is a 6×5 matrix in which each element is a sign of variations in one of the 30 areas. High vibrations in some areas lead to more brilliant areas in the QIM images. So 6th to 35th features refer to these areas. Figure 5 shows the QIM image of happiness state. As shown in Fig. 5, QIM is a good approximate for muscle changes in each area. For example, the last image in Fig. 5 shows that during facial expression of happiness, the most changes occur in the cheeks and lips.

Fig. 4 Regions of ITMI [44]

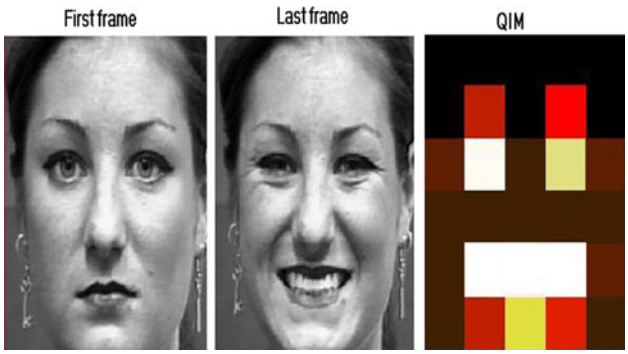

Fig. 5 QIM for happy state [44]

We extracted five features from the ITMIs and 30 features from the QIM images. Therefore, in total, 35 features were extracted to recognize facial expression.

3.3 Bimodal system

Combining is an approach to improve the performance of classification particularly for complex problems such as those involving a considerable amount of noise, limited number of training patterns, high-dimensional feature sets and highly overlapped classes [37, 38].

To combine the facial expression and speech information, three different approaches were implemented: feature-level fusion, in which a single classifier with features of both modalities is used, decision-level fusion, in which a separate classifier is used for each modality and the outputs are combined using some criteria, and, finally, hybrid of the both methods (model-level fusion) [6], in order to combine the benefits of both feature-level and decision-level fusion methods.

The block diagram of the proposed bimodal system is depicted in Fig. 6. Features of speech signal and face image sequences are extracted and used to the related individual classifiers. Furthermore, the features are mixed to use by another classifier, which is based on the joint information. Finally, the results of these systems are fused to a meta-classifier.

Classifier combination is divided into non-trainable and trainable approach. Voting, averaging and Borda counts are non-trainable. Various combiners may be used, depending on the type of output produced by the classifier. We used voting to combine the results of the classifiers. The voting method is used when each classifier produces a single class label. In this case, each classifier "votes" for a particular class, and thus, the class with the majority vote on the ensemble wins.

Weighted averaging and stacked generalization are trainable. In stacked generalization, the output of the ensemble serves as a feature vector to a meta-classifier. Stacked generalization [39] provides a way of combining trained networks together, which uses partitioning of the data set to find an overall system with usually improved generalization performance. In this work, we used MLP and RBF networks as a meta-classifier to improve generalization performance. Also, we proposed a multi-classifier that has the advantages of both the speech-based and facial expression systems. This proposed system is an improved hybrid system.

4 Feature selection using ANOVA

For dimension reduction and construction of a lower-sized feature space, an open-loop (independent of classifier) feature selection method was used in this paper.

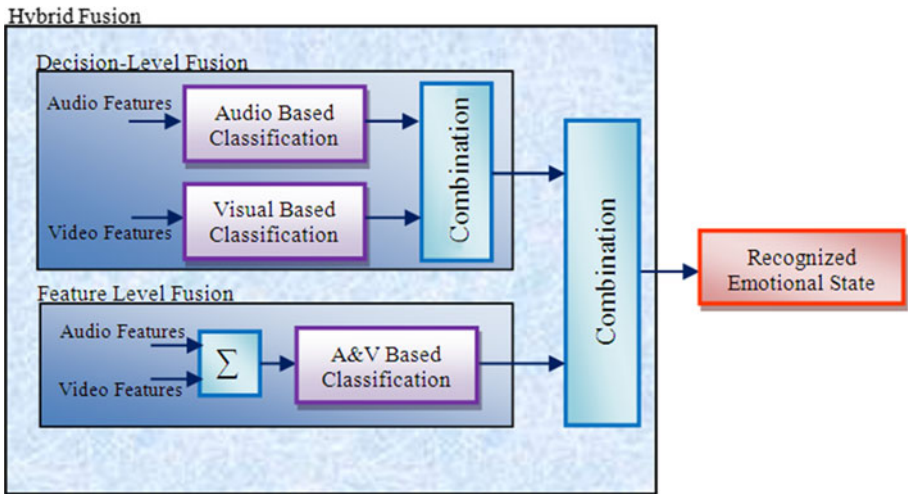

Fig. 6 *Block diagram* of the proposed bimodal system

To reduce the number of features, a feature selection method, based on the analysis of variations (ANOVA), was used. We computed the importance ranking of each feature using ANOVA, which is a technique for analyzing experimental data in which one or more response variables are measured under various conditions identified by one or more classification variables. A typical goal in ANOVA is to compare means of the response variables for various combinations of the classification variables. ANOVA was used to decide whether a feature shows a significant difference between two or more classes.

One-way ANOVA is a method for testing null hypotheses on equal means in several populations [38]. Suppose that data are sampled from k different populations, and assume the model as follows:

$$Y_{ij} = \mu_i + \varepsilon_{ij}; \quad j = 1, \ldots, n_i \quad i = 1, \ldots, k, \tag{6}$$

where Y_{ij} is the jth observation from the ith population, μ_i is the mean of the ith population, and ε_{ij} denotes the random variation in Y_{ij} away from μ_i. It is assumed that the ε_{ij}s are independent normally distributed random variables with zero mean and variance σ^2. The one-way ANOVA can only tell us whether all the means are equal, or whether there is a difference in the means of different populations.

5 Experimental results

The proposed multimodal emotion recognition system was tested over the eNterface '05 audiovisual emotional database. All the experiments were person independent. We used roughly 64 % of the data (i.e., 674 shots) for training the classifiers and the remaining (372) shots for the evaluation. In our experiments, 17.6 % of the samples were for the anger, 16.3 % for disgust, 16 % for fear, 16 % for happiness, 17 % for sadness and, finally, 17 % for surprise states. The emotion recognition was conducted through unimodal facial expression system, unimodal speech-based system, decision-level fusion of the unimodal systems, feature-level fusion and hybrid features and decision-level fusion. The results are summarized in Fig. 7.

5.1 eNterface '05 database

The eNterface '05 database [40] is the only publicly available we found for audiovisual emotional database. Forty-two non-native English-speaking subjects from 14 different nationalities posed the six basic emotions (81 % of the subjects were men). Some subjects had facial hair (17 %), some were wearing glasses (31 %), and hair of the head had covered parts of the upper face in a few subjects, and, finally, one of the subjects was bald (2 %). None of the subjects were professional actors.

The subjects were told to listen to six different short stories, each containing a particular emotion (anger, disgust, fear, happiness, sadness and surprise), and to react to each of the situations uttering five different predefined sentences. They were not given further constraints or guidelines regarding how to express the emotion.

This database contains 44 (subjects) by 6 (emotions) by 5 (sentences) shots. The average video length was about 3 s summing up to 1,320 shots and more than 1 h of videos. The videos were recorded in a laboratory environment: The subjects were recorded from frontal view with studio lightening condition and gray uniform background. Audios were recorded with a high-quality microphone placed at around 30 cm from the subject's mouth.

Paleari and Huet [41] evaluated the quality of this database and pointed out some weaknesses of it. Here, we briefly cite some of them:

1. The subjects were not trained actors possibly resulting in a mediocre emotional expression quality.
2. The quality of the encoding was mediocre.

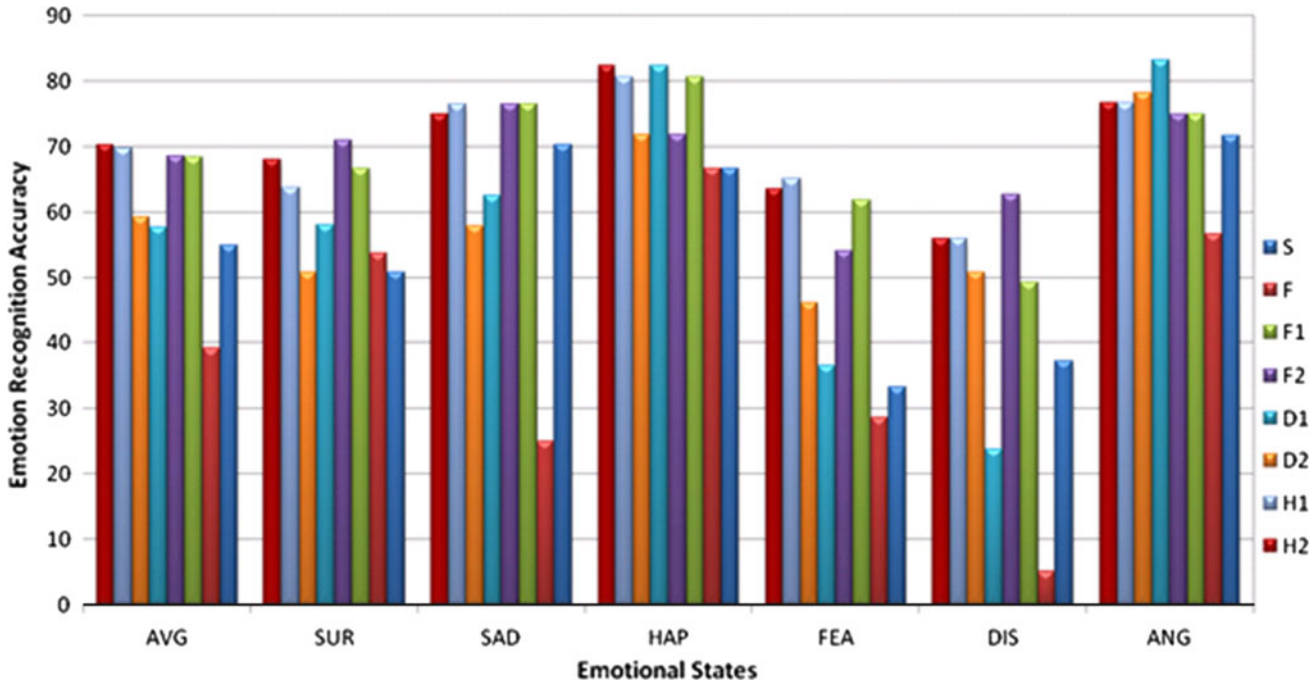

Fig. 7 Emotion recognition accuracy of the proposed systems. Each group of adjacent *columns* denotes the classification accuracy of a single class. The first group contains the average recognition rate. The vertical axis is the recognition accuracy in percentage. *S* speech, *F* face, *F1* feature-level fusion, *F2* feature-level fusion with ANOVA, *D1* decision-level fusion (max), *D2* decision-level fusion (MLP), *H1* hybrid fusion (MLP), *H2* hybrid fusion (RBF). Class *labels* are abbreviated by their *first three letters*

3. The subjects were asked to utter sentences in English, but since for some subjects English was not native language, this might result in a low quality of the prosodic emotional modulation.
4. Not all of the subjects learned their sentences by heart, resulting in a non-negligible percentage of videos starting with the subjects looking down to read their sentences.

5.2 Speech emotion classifier

Table 1 shows the confusion matrix of the emotion recognition system based on speech information. The overall performance of this classifier was 55 %. Table 1 shows that some pairs of emotions are usually confused more. For example, disgust is misclassified as happiness state by about 20.34 %, and vise versa happiness is misclassified as disgust by about 12.28 % (Table 2).

We examined our speech recognition system by Berlin database of emotional speech [42]. The result was better than the eNterface '05 database. The overall performance of this system was 79.28 % for Berlin database. This may be due to the fact that subjects of Berlin database were experienced actors and were native German speakers.

According to some reports, on eNterface '05 database [43], sadness and anger are recognized by speech very well, and disgust is poorly classified by speech.

Table 1 Confusion matrix of the emotion recognition system based on speech (eNterface '05 database)

	ANG	DIS	FEA	HAP	SAD	SUR
ANG	71.67	0	5.00	13.33	1.67	8.33
DIS	10.17	37.29	8.47	20.34	11.86	11.86
FEA	14.29	9.52	33.33	12.70	12.70	17.46
HAP	8.77	12.28	3.51	66.67	3.51	5.26
SAD	3.13	4.69	7.81	7.81	70.31	6.25
SUR	8.70	10.14	7.25	10.14	13.04	50.72

Table 2 Confusion matrix of the emotion recognition system based on speech (Berlin database)

	ANG	Boredom	DIS	FEA	HAP	SAD	Neutral
ANG	73.91	0	17.39	0	4.35	0	4.35
Boredom	8.70	86.96	0	0	4.35	0	0
DIS	2.27	0	97.73	0	0	0	0
FEA	0	0	0	72.73	4.55	13.64	9.09
HAP	6.25	12.50	12.50	0	56.25	6.25	6.25
SAD	0	0	0	4.17	0	95.83	0
Neutral	0	2.63	0	23.68	0	2.63	71.05

5.3 Facial expression-based system

Table 3 shows the confusion matrix of the emotion recognition system based on facial expressions. The overall performance of this classifier was 39.27 %.

According to some reports, on eNterface '05 database [43], disgust is poorly classified by face, because disgust is a mouth-dependent class, and during speaking, most of the facial activity in the mouth is related to lip motions.

We examined our method on a common facial expression database (Kanade et al. [44]). The overall performance of this database was 71.8 %, showing the good performance of our method. Table 4 shows the confusion matrix for this database. Disgust has been recognized relatively good in this experiment. This may be due to the fact that these data are only facial expression based and the subjects focused just on facial activity. Also, as mentioned above, during speaking, most of the facial activity in the lower face is related to lip motions, which considerably lowers the recognition rates of mouth-dependent classes such as disgust in audiovisual-based data. On the other hand, the subjects of Cohn–Kanade database were experienced actors enrolled in introductory psychology classes.

5.4 Bimodal system

The overall results of the unimodal systems suggest that, for accurate and reliable recognition of emotion classes, the modalities should be combined in a way that they benefit the interrelationships between the individual classes and the underlying modalities. In the following paragraphs, we present and compare different combination schemes. Three main fusion approaches used in the literature are feature-level fusion, decision-level fusion and, finally, a hybrid of the feature- and decision-level fusion approaches [43].

Table 5 shows the confusion matrix of the feature-level fusion. We used compound set of multimodal features as input to the classifiers. Our classifier in this experiment was MLP. The overall performance of this classifier was 68.33 %. All states, except disgust, were recognized with more than 62 %. By using ANOVA, we selected 92 features out of 115 features. Sixty-seven of the selected features were from speech-based features and 25 from facial expression features.

Table 6 shows the confusion matrix for the selected features at feature-level fusion. The overall performance of this classifier was 68.53 %.

Comparison between Tables 5 and 6 shows that emotion recognition accuracy can be improved using selected features for disgust and surprise states and deteriorated for fear and happiness states. The overall emotion recognition accuracy using feature selection algorithm improves by about 0.2 %.

Table 7 shows the confusion matrix using decision-level fusion. In this experiment, we used voting method as a combination of audio and video classifiers. In this case, each classifier "votes" for a particular class, and the class with the majority vote on the ensemble wins. The overall performance of this method was 57.75 %.

Table 3 Confusion matrix of the facial expression-based system (eNterface '05 database)

	ANG	DIS	FEA	HAP	SAD	SUR
ANG	56.67	1.67	8.33	8.33	6.67	18.33
DIS	18.64	5.08	6.78	28.81	8.47	32.20
FEA	15.87	4.76	28.57	7.94	17.46	25.40
HAP	7.02	5.26	1.75	66.67	5.26	14.04
SAD	12.50	3.13	12.50	4.69	25.00	42.19
SUR	17.39	2.90	4.35	13.04	8.70	53.62

Table 4 Confusion matrix of the facial expression-based system (Cohn–Kanade database)

	ANG	DIS	FEA	HAP	SAD	SUR
ANG	63.64	12.12	3.03	12.12	0	9.09
DIS	10.64	70.21	6.38	6.38	4.26	2.13
FEA	0	0	73.17	9.76	7.32	9.76
HAP	2.22	6.67	11.11	73.33	4.44	2.22
SAD	4.35	4.35	13.04	13.04	65.22	0
SUR	5.66	1.89	7.55	0	0	84.91

Table 5 Confusion matrix using feature-level fusion

	ANG	DIS	FEA	HAP	SAD	SUR
ANG	75.00	5.00	3.33	3.33	8.33	5.00
DIS	5.08	62.71	3.39	8.47	11.86	5.47
FEA	4.76	9.52	53.97	12.70	9.52	9.52
HAP	8.77	12.28	1.75	71.93	0	5.26
SAD	1.56	7.81	1.56	1.56	76.56	10.94
SUR	2.90	11.59	4.35	2.90	7.25	71.01

Table 6 Confusion matrix for the selected features at feature-level fusion

	ANG	DIS	FEA	HAP	SAD	SUR
ANG	75.00	5.00	3.33	3.33	8.33	5.00
DIS	5.08	62.71	3.39	8.47	11.86	8.47
FEA	4.76	9.52	53.97	12.70	9.52	9.52
HAP	8.77	12.28	1.75	71.93	0	5.26
SAD	1.56	7.81	1.56	1.56	76.56	10.94
SUR	2.90	11.59	4.35	2.90	7.25	71.01

Table 7 Confusion matrix of the voting decision-level fusion

	ANG	DIS	FEA	HAP	SAD	SUR
ANG	83.33	0	1.67	10.00	0	5.00
DIS	13.56	23.73	8.47	32.20	11.86	10.17
FEA	19.05	7.94	36.51	15.87	6.35	14.29
HAP	3.51	5.26	1.75	82.46	0	7.02
SAD	7.81	3.13	7.81	6.25	62.50	12.50
SUR	15.94	2.90	1.45	11.59	10.14	57.97

Table 8 Confusion matrix of the decision-level fusion using MLP

	ANG	DIS	FEA	HAP	SAD	SUR
ANG	78.33	5.00	6.67	1.67	1.67	6.67
DIS	11.86	50.85	11.86	6.78	8.47	10.17
FEA	12.70	12.70	46.03	6.35	4.76	17.46
HAP	0	17.54	5.26	71.93	0	5.26
SAD	3.13	9.38	20.31	0	57.81	9.38
SUR	7.25	21.74	14.49	0	5.80	50.72

In the next experiment, we used stacked generalization method. The output of the audio and video ensembles serves as a feature vector to a MLP. Table 8 shows the confusion matrix of this experiment. The overall performance of this method was 59.28 % which is better than that of individual classification and voting decision-level fusion.

As mentioned, hybrid fusion method that combines the advantages of both feature-level and decision-level methods may be a good choice for fusion of audio and visual emotion recognition.

So, in our work, we focused on hybrid fusion. In this case, the output of the feature-level and decision-level ensembles serves as a feature vector to a meta-classifier. We used MLP and RBF networks as a meta-classifier.

Table 9 shows the confusion matrix of the hybrid features and decision-level fusion using MLP as a classifier. The overall performance of this method was 69.78 %. Table 10 shows the confusion matrix of the hybrid features and decision-level fusion using RBF; the overall performance of this method was obtained as 70.28 %.

Table 9 Confusion matrix of hybrid features and decision-level fusion using MLP

	ANG	DIS	FEA	HAP	SAD	SUR
ANG	76.67	3.33	6.67	1.67	0	11.67
DIS	6.78	55.93	13.56	10.17	8.47	5.08
FEA	12.70	6.35	65.08	3.17	3.17	9.52
HAP	1.75	8.77	7.02	80.70	0	1.75
SAD	1.56	4.69	10.94	0	76.56	6.25
SUR	5.80	13.04	5.80	2.90	8.70	63.77

Table 10 Confusion matrix of hybrid features and decision-level fusion using RBF

	ANG	DIS	FEA	HAP	SAD	SUR
ANG	76.67	3.33	6.67	1.67	0	11.67
DIS	8.47	55.93	13.56	10.17	5.08	6.78
FEA	11.11	3.17	63.49	6.35	4.76	11.11
HAP	1.75	8.77	3.51	82.46	0	3.51
SAD	1.56	3.13	10.94	1.56	75.00	7.81
SUR	5.80	4.35	10.14	4.35	7.25	68.12

Table 11 Recognition rate of emotional states for various implemented systems

	S	F	F1	F2	D1	D2	H1	H2
ANG	71.67	56.67	75	75	83.33	78.33	76.67	76.67
DIS	37.29	5.08	49.15	62.71	23.73	50.85	55.93	55.93
FEA	33.33	28.57	61.9	53.97	36.51	46.03	65.08	63.49
HAP	66.67	66.67	80.7	71.93	82.46	71.93	80.7	82.46
SAD	70.31	25	76.56	76.56	62.5	57.81	76.56	75
SUR	50.72	53.62	66.67	71.01	57.97	50.72	63.77	68.12
AVG	54.99	39.27	68.33	68.53	57.75	59.28	69.78	70.28

Figure 7 and Table 11 compare the emotion recognition results obtained from the unimodal and different combining methods. As shown, combining the information of multiples modalities enhances the classification accuracy.

Combining of speech and face information in different ways enhances the performance of unimodal systems. Table 11 shows that the method of hybrid features and decision-level fusion with RBF (H2) has better performance. The mean accuracy of this system is 70.28, and these results show that this method improves the recognition rate by up to 15 % over the speech-based system and by up to 25 % over the facial expression-based system. Figure 8 compares the performance of hybrid features and decision-level fusion using RBF (H2) system with the unimodal systems.

In this research, the Clementine software [45] was used for implementing the MLP and RBF neural networks. For training the networks the gradient descent was used by this software. This software is able to estimate most of the parameters initially based on the size of the input data. Momentum rate (α) was set at 0.9 for training of MLPs to avoid local minima. The learning rate (η) was initially 0.3 and decayed exponentially to 0.01. Then it was reset to 0.1 and again decayed to 0.01 in 30 epochs.

The only remaining parameter was the architecture of the network. Clementine software uses different topologies for networks by setting various numbers of hidden layers. The work is started with training a sequence of two-layer

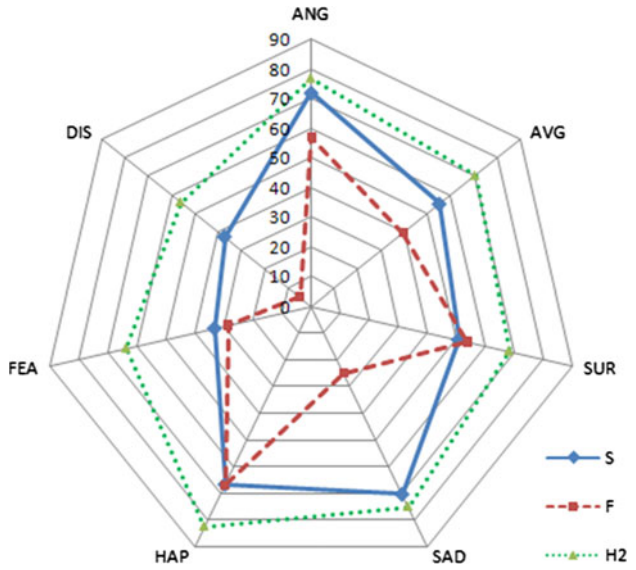

Fig. 8 Comparison of audio, video and H2 systems

Table 12 Topology of MLPs in experimented models

MLP model	Number of input layer nodes	Number of first hidden layer nodes	Number of second hidden layer nodes	Number of output layer nodes
F	36	28	–	6
S	80	67	50	6
F1	116	67	26	6
F2	91	67	–	6
D2	12	19	17	6
H1	18	12	10	6

neural networks with an increasing number of hidden nodes. As the number of hidden nodes increases, the training error decreases. For each topology, the root mean square (RMS) error is calculated, and finally, the model with the lowest error is selected. The best topologies for MLP with speech-based features (S), facial expression-based features (F), feature-level fusion (F1), feature-level fusion with ANOVA (F2), decision-level fusion (D2) and hybrid features and decision-level fusion (H1) are reported in Table 12. Topology of RBF in hybrid fusion (H2) is (18, 20 and 6).

6 Proposed multi-classifier system

According to Fig. 8 and the results of our pervious experiments, angry and sadness are recognized by speech system very well, and facial expression-based system cannot improve the result in fusion approach, so in our proposed system, we recognized these emotional states only with speech-based system. Figure 9 shows the architecture of a multi-classifier scheme. This system is an improved hybrid system that combines the advantages of both feature-level and decision-level fusion methods as well as the speech-based and facial expression systems.

In this architecture, the audio features are fed to a MLP neural network to classify the emotions. According to the results of the speech-based classifier, all sentences except for angry and sadness are fed to a facial expression-based MLP neural network.

The outputs of speech-based classifier, facial expression-based classifier and feature-level fusion classifier are fed to a RBF neural network to combine their result.

Table 13 shows the confusion matrix of the proposed system. The overall performance of this method was 77.78 %. The results showed that this method improves recognition rate by up to 7.5 % over the hybrid features and decision-level fusion (H2), by up to 22.7 % over the speech-based system and by up to 38 % over the facial expression-based system. Figure 10 compares the proposed multi-classifier system with method of hybrid features and decision-level fusion with RBF (H2) and unimodal systems. The recognition rates of angry, disgust and sadness were improved significantly in this experiment. The recognition rate of sadness was 92.66 % in this experiment.

Emotion recognition rates that were achieved by the multimodal classification in other works may be helpful for analyzing the performance of the proposed approaches. Zeng et al. [46] applied multi-stream HMMs and improved the emotion recognition rate by up to 6.5 % over the unimodal systems. Paleari and Huet [41] applied different ways to combine speech and face in eNterface '05 database and got an improvement of about 6 % over the speech-based system and about 14 % over the facial expression

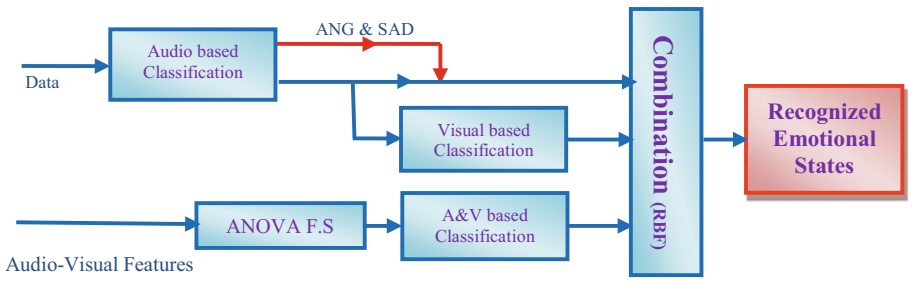

Fig. 9 The architecture of multi-classifier scheme for emotion recognition

Table 13 Confusion matrix of multi-classifier system

	ANG	DIS	FEA	HAP	SAD	SUR
ANG	84.24	2.72	0	1.09	8.15	3.80
DIS	1.17	79.53	6.43	2.92	2.92	7.02
FEA	2.37	6.51	68.64	6.51	2.37	13.61
HAP	2.40	13.17	4.19	73.65	1.20	5.39
SAD	56.0	2.82	1.13	0	92.66	2.82
SUR	1.12	14.04	8.43	8.43	0	67.98

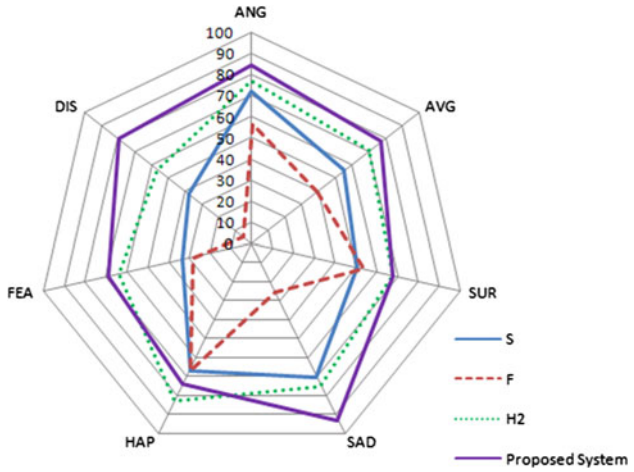

Fig. 10 Comparison of audio, video and H2 and purpose systems

system. Using feature-level fusion, Busso et al. [47] improved the emotion recognition rate by up to 5 % over the speech-based systems and by about 19 % over the facial expression system. Table 14 shows the performance of the proposed system and some other multimodal emotion recognition systems.

As already depicted, the emotion recognition system based on facial expression could not recognize the angry and sadness states very well, but the emotion recognition based on speech could do so. The main goal of this research is recognizing the strength and weakness points of the emotion recognition systems based on facial expression- and speech-based systems to use them on designing a hybrid emotion recognition system. This is one type of boosting approach. As shown in Fig. 10, the emotion recognition based on speech has the main weight in recognition of angry and sadness states.

In this research, compared with other researches that used eNterface '05 as emotional database, the emotion recognition was based on sentence level not on frames; therefore, we have less complexity and computational cost. The suitable design of speech-based emotion recognition system and the improved combination of it with other recognition systems causes the better performance of proposed system as comparing with [43] that used asynchronous feature-level fusion for recognition with more complexity and computational cost.

Table 14 Performance of typical systems for multimodal emotion recognition in the recent decade

References	Classifier	Features	Fusion	Database	Acc	Acc (fusion)
Paleari and Huet [41]	SVM, NN	Facial points, formants, prosody, MFCC, LPC	F, D	eNterface '05	Audio: 25 Video: 33	39
Busso et al. [47]	SVM	102 markers, prosody	F, D	An actress	Audio: 70.9 Video: 85.1	89
Mansoorizadeh and Moghaddam Charkari [43]	SVM	Facial points, prosody	F, D, H	eNterface '05	Audio: 33 Video: 37	71
De Silva and Pei Chi [13]	HMM, nearest neighbor	Facial points, pitch	D	2 subjects, 144 clips	Audio: 62 Video: 32	72
Cheng et al. [48]	SVM	Facial points, prosody	F	2 subjects, 350 clips	Audio: 63 Video: 75	84
Schuller et al. [49]	SVM	Face model, formants, prosody, MFCC	F	Database ABC	Audio: 74 Video: 61	81
Zeng et al. [46]	Snow and HMM	Facial points, prosody	D	20 subjects, 11 affect categories	Audio: 48 Video: 44	95
Proposed approach	NN	ITMI, QIM, MFCC, prosody, formants	F, D, H	eNterface '05	Audio: 54.99 Video: 39.27	70.28
Proposed approach	Multi-classifier	ITMI, QIM, MFCC, prosody, formants	F, D, H	eNterface '05	Audio: 54.99 Video: 39.27	77.78

7 Conclusion

This paper proposes a new multi-classifier system that is improved hybrid features and decision-level fusion architecture for multimodal emotion recognition. The system combines facial expression and speech information in feature and decision levels using stacked generalization approach. Feature-level fusion captures cross-correlations between the modalities, and decision-level fusion brings robustness into the system [6]. Also, we recognized the strength and weakness points of the emotion recognition systems based on facial expression- or speech-based systems and used them on designing a multi-classifier emotion recognition system.

Experimental results showed that the results of unimodal systems were improved by using the hybrid features and decision-level fusion. Also, by using the proposed multi-classifier system, the recognition rate was improved by about 22.7 % with respect to the speech unimodal system, and by about 38 % with respect to the facial expression system.

A number of promising methods for vision-based, audio-based and audiovisual analysis of human spontaneous behavior have so far been proposed [11]. One of the unexplored areas of researches on multimodal emotion recognition is temporal structures of the modalities (facial and vocal) and their temporal correlations. Also, developing better methods and models for multimodal fusion is one of the most important issues that lacks sufficient attention.

According to some reports ([6, 11]) and our results in this work, model-level fusion or hybrid fusion is a good choice for multimodal emotion recognition. So in this study, we focused on hybrid fusion and different ways to combine the results of audio, video and audiovisual systems. We used stacked generalization method to fuse the output of these systems. Finally, we proposed an improved hybrid system. By using this system, the recognition rate increased by up to 7.5 % over the hybrid features and decision level with RBF, by up to 22.7 % over the speech-based system and by up to 38 % over the facial expression-based system.

References

1. Devillers L, Vidrascu L (2006) Real-life emotions detection with lexical and paralinguistic cues on human call center dialogs. In: Proceedings of the interspeech, pp 801–804
2. Lee C-C, Mower E, Busso C, Lee S, Narayanan S (2009) Emotion recognition using a hierarchical binary decision tree approach. In: Proceedings of the interspeech, pp 320–323
3. Polzehl T, Sundaram S, Ketabdar H, Wagner M, Metze F (2009) Emotion classification in children's speech using fusion of acoustic and linguistic features. In: Proceedings of the interspeech, pp 340–343
4. Klein J, Moon Y, Picard RW (2002) This computer responds to user frustration: theory, design and results. Interact Comput 14:119–140
5. Oudeyer P-Y (2003) The production and recognition of emotions in speech: features and algorithms. Int J Hum Comput Interact Stud 59:157–183
6. Mansoorizadeh M, Moghaddam Charkari N (2009) Hybrid feature and decision level fusion of face and speech information for bimodal emotion recognition. In: Proceedings of the 14th international CSI computer conference
7. Ambady N, Rosenthal R (1992) Thin slices of expressive behavior as predictors of interpersonal consequences: a meta-analysis. Psychol Bull 111(2):256–274
8. Ekman P, Rosenberg EL (2005) What the face reveals: basic and applied studies of spontaneous expression using the facial action coding system (FACS), 2nd edn. Oxford University Press, Oxford
9. Mehrabian A (1968) Communication without words. Psychol Today 2:53–56
10. Greenwald M, Cook E, Lang P (1989) Affective judgment and psychophysiological response: dimensional covariation in the evaluation of pictorial stimuli. J Psychophysiol 3:51–64
11. Zeng Z, Pantic M, Roisman GI, Huang TS (2009) A survey of affect recognition methods: audio, visual, and spontaneous expressions. PAMI 31:39–58
12. Pantic M, Rothkrantz LJM (2000) Automatic analysis of facial expressions: the state of the art. IEEE Trans Patt Anal Mach Intell 22:1424–1445
13. De Silva LC, Pei Chi N (2000) Bimodal emotion recognition. In: Proceedings of the fourth IEEE international conference on automatic face and gesture recognition, vol 1, pp 332–335
14. Song M, You M, Li N, Chen C (1920) A robust multimodal approach for emotion recognition. Neurocomputing 71:1913–2008
15. Hoch S, Althoff F, McGlaun G, Rigooll G (2005) Bimodal fusion of emotional data in an automotive environment. In: Proceedings of the international conference on acoustics, speech, and signal processing, vol 2, pp 1085–1088
16. Wang Y, Guan L (2005) Recognizing human emotion from audiovisual information. In: Proceedings of the international conference on acoustics, speech, and signal processing, pp 1125–1128
17. Paleari M, Benmokhtar R, Huet B (2008) Evidence theory-based multimodal emotion recognition. In: MMM '09, pp 435–446
18. Sheikhan M, Bejani M, Gharavian D (2012) Modular neural-SVM scheme for speech emotion recognition using ANOVA feature selection method. Neural Comput Appl J. doi:10.1007/s00521-012-0814-8
19. Lee CM, Narayanan SS (2005) Toward detecting emotions in spoken dialogs. IEEE Transact Speech Audio Process 13:293–303
20. Gharavian D, Ahadi SM (2005) The effect of emotion on farsi speech parameters: a statistical evaluation. In: Proceedings of the international conference on speech and computer, pp 463–466
21. Ververidis D, Kotropoulos C (2006) Emotional speech recognition: resources, features, and methods. Speech Commun 48:1162–1181
22. Shami M, Verhelst W (2007) An evaluation of the robustness of existing supervised machine learning approaches to the classifications of emotions in speech. Speech Commun 49:201–212
23. Altun H, Polat G (2009) Boosting selection of speech related features to improve performance of multiclass SVMs in emotion detection. Expert Syst Appl 36:8197–8203
24. Gharavian D, Sheikhan M, Nazerieh AR, Garoucy S (2011) Speech emotion recognition using FCBF feature selection method and GA-optimized fuzzy ARTMAP neural network. Neural Comput Appl. doi:10.1007/s00521-011-0643-1
25. Sheikhan M, Safdarkhani MK, Gharavian D (2011) Emotion recognition of speech using small-size selected feature set and ANN-based classifiers: a comparative study. World Appl Sci J 14:616–625

26. Fersini E, Messina E, Archetti F (2012) Emotional states in judicial courtrooms: an experimental investigation. Speech Commun 54:11–22
27. Albornoz EM, Milone DH, Rufiner HL (2011) Spoken emotion recognition using hierarchical classifiers. Comput Speech Lang 25:556–570
28. López-Cózar R, Silovsky J, Kroul M (2011) Enhancement of emotion detection in spoken dialogue systems by combining several information sources. Speech Commun 53:1210–1228
29. Boersma P, Weenink D (2007) Praat: doing phonetics by computer (version 4.6.12) [computer program]
30. Bobick AF, Davis JW (2001) The recognition of human movement using temporal templates. IEEE Trans Patt Anal Mach Intell 23(3):257–267
31. Valstar MF, Patras I, Pantic M (2004) Facial action unit recognition using temporal templates. In: IEEE international workshop on human robot interactive communication
32. Osadchy M, Keren D (2004) A rejection-based method for event detection in video. IEEE Trans Circuits Syst Video Technol 14(4):534–541
33. Li N, Dettmer S, Shah M (1997) Visually recognizing speech using eigensequences. In: Motion-based recognition. Kluwer, Boston, pp 345–371
34. Babua RV, Ramakrishnanb KR (2004) Recognition of human actions using motion history information extracted from the compressed video. Image Vis Comput 22:597–607
35. Sadoghi Yazdi H, Amintoosi M, Fathy M (2007) Facial expression recognition with QIM and ITMI spatio-temporal database. In: 4th Iranian conference on machine vision and image processing, Mashhad, Iran, pp 14–15 (Persian)
36. Intel, OpenCV Open source computer vision library. http://www.intel.com/research/mrl/research/opencv/
37. Ebrahimpour R (2007) View-independent face recognition with mixture of experts. Dissertation, The Institute for Research in Fundamental Sciences (IPM)
38. Ghaderi R (2000) Arranging simple neural networks to solve complex classification problems. Dissertation, Surrey University
39. Wolpert DH (1992) Stacked generalisation. Complex Syst 5:241–259
40. Martin O, Kotsia I, Macq B, Pitas I (2006) The enterface '05 audio-visual emotion database. In: Proceedings of the 22nd international conference on data engineering workshops (ICDEW '06)
41. Paleari M, Huet B (2008) Toward emotion indexing of multimedia excerpts. In: CBMI
42. Burkhardt F, Paeschke A, Rolfes M, Sendlmeier W, Weiss B (2005) A database of German emotional speech. In: Interspeech, Lisbon, Portugal
43. Mansoorizadeh M, Moghaddam Charkari N (2009) Multimodal information fusion application to human emotion recognition from face and speech. Multimed Tools Appl
44. Kanade T, Cohn J, Tian Y (2000) Comprehensive database for facial expression analysis. In: IEEE international conference on face and gesture recognition (AFGR '00), pp 46–53
45. SPSS (2007) Clementine® 12.0 algorithms guide. Integral Solutions Limited, Chicago
46. Zeng Z, Hu Y, Roisman GI, Wen Z, Fu Y, Huang TS (2007) Audio-visual spontaneous emotion recognition. Artif Intell Hum Comput 4451:72–90
47. Busso C et al (2004) Analysis of emotion recognition using facial expressions, speech and multimodal information. In: Proceedings of the sixth ACM international conference on multimodal interfaces (ICMI '04), pp 205–211
48. Cheng-Yao C, Yue-Kai H, Cook P (2005) Visual/acoustic emotion recognition, pp 1468–1471
49. Schuller B, Arsic D, Rigoll G, Wimmer M, Radig B (2007) Audiovisual behavior modeling by combined feature spaces. In: ICASSP, pp 733–736