SIViP (2016) 10:827–834
DOI 10.1007/s11760-015-0822-0

ORIGINAL PAPER

# Multimodal emotion recognition based on peak frame selection from video

**Sara Zhalehpour**[1] · **Zahid Akhtar**[2] · **Cigdem Eroglu Erdem**[3]

Received: 12 November 2014 / Revised: 18 September 2015 / Accepted: 19 September 2015 / Published online: 29 September 2015

**Abstract** We present a fully automatic multimodal emotion recognition system based on three novel peak frame selection approaches using the video channel. Selection of peak frames (i.e., apex frames) is an important preprocessing step for facial expression recognition as they contain the most relevant information for classification. Two of the three proposed peak frame selection methods (i.e., MAXDIST and DEND-CLUSTER) do not employ any training or prior learning. The third method proposed for peak frame selection (i.e., EIFS) is based on measuring the "distance" of the expressive face from the subspace of neutral facial expression, which requires a prior learning step to model the subspace of neutral face shapes. The audio and video modalities are fused at the decision level. The subject-independent audio-visual emotion recognition system has shown promising results on two databases in two different languages (eNTERFACE and BAUM-1a).

**Keywords** Affective computing · Facial expression recognition · Apex frame · Audio-visual emotion recognition

## 1 Introduction

Automatic human emotional states recognition is an important problem in human–computer interaction [21,29]. Recognition of a person's emotional state has many applications in diverse areas such as psychology [24], security [20], health care [14], to name a few. Below, we briefly review the literature on emotion recognition and then motivate the proposed method for audio-visual emotion recognition based on three different novel peak frame selection approaches.

### 1.1 Prior work

Humans express their emotions through various channels including facial expressions, head gestures, and speech. The majority of existing automatic emotion recognition approaches focus on either visual or audio modality. The visual modality is the most widely used channel, and state-of-the-art methods usually utilize 2D facial features. 2D facial features can be broadly grouped as *geometric features* and *appearance-based features*. Geometric features localize the salient facial points and detect the emotion based on the deformation of these facial points [26]. Appearance-based features represent the change in the texture of the expressive face [8,14,20,30]. The audio modality methods use prosodic features along with spectral, cepstral, and voice quality features [2,3,22].

Although research on the audio and visual channels has considerably progressed in the recent years, performance of uni-modal systems is not very high for both acted [19,27] and spontaneous emotions [6,7]. Hence, integration of audio and visual channels to increase the emotion recognition performance is still an open research problem. Recent studies have shown many advantages of fusing audio and video channels for emotion recognition [5,9,13,16,19,27,28]. Man-

This work was supported by the Turkish Scientific and Technical Research Council (TÜBİTAK) under project EEAG-110E056.
Sara Zhalehpour, Zahid Akhtar: The authors should be considered as co-first authors.

✉ Cigdem Eroglu Erdem
cigdem.eroglu@eng.bahcesehir.edu.tr

[1] INRS-EMT, Montreal, Canada

[2] University of Udine, Udine, Italy

[3] Bahcesehir University, Istanbul, Turkey

soorizadeh et al. [16] proposed an asynchronous feature-level fusion approach, which uses both feature- and decision-level fusion. This method assumes that the features of the audio and video modalities are temporally overlapped. Gajsek et al. [9] presented an audio-visual emotion recognition system, which uses prosodic and cepstral coefficients as audio features and Gabor wavelets as video features, followed by stepwise feature selection and a multiclass classifier system. Datcu et al. [5] outlined a multimodal semantic data fusion model. This method considers the presence or absence of speech and utilizes only eye- and eyebrow-related features. Paleari et al. [19] presented a framework for multimodal emotion recognition based on different possible fusion approaches. Wang et al. [27] introduced a kernel cross-modal factor analysis method using nonlinear transformations for multimodal emotion recognition. Kuan-Chieh et al. [13] devised a multimodal emotion recognition technique that automatically learns the weighted sum of decision-making parameters for each modality.

### 1.2 Problem statement and motivation

An emotional video of facial expressions consists of hundreds of frames, where the emotion is expressed with a different intensity at each frame. Therefore, when it comes to emotion recognition from a video, it is a challenge to decide how to use these frames so that the facial expression recognition rate is maximized. One promising approach is to utilize a single frame or a subset of frames, which best represent the emotional content of the sequence. These are the frames at which the emotional expression is at its apex, assuming that the facial parts reach the peak point at the same time. We call such frames "*peak frames*" of the video. The problem of selecting the peak frames automatically is an issue that has been addressed by very few researchers so far. In this work, we propose three novel methods for automatically selecting the peak frames from an emotional video, which are used in a multimodal emotion recognition system.

Two of the three proposed peak frame selection methods (i.e., MAXDIST and DEND-CLUSTER) have several advantages. First, they can select the peak frames without employing any training or prior learning. Second, the facial features that are used for peak frame selection can also be used for facial expression recognition. Third, neutral face image of the test subject is not required. The third method proposed for peak frame selection (i.e., EIFS) is based on measuring the "distance" of the expressive face from the subspace of neutral faces. Thus, a prior learning step is needed to model the subspace of neutral face images. All the proposed approaches are based on the assumption that there is a single emotion expressed in the given video segment. The audio-visual emotion recognition method presented in this paper utilizes decision-level fusion to merge the information from the audio and video channels. The method has shown promising results on two databases.

The paper is organized as follows: The proposed approaches are described in Sect. 2 and evaluated in Sect. 3. A conclusion is drawn in Sect. 4.

## 2 Multimodal emotion recognition system

In this section, we present our methods for extraction of the video and audio features, classification, and multimodal decision-level fusion.

### 2.1 Feature extraction from video

Below, we detail the three proposed peak frame selection methods: (i) *maximum dissimilarity-based method* (MAX DIST), (ii) *clustering-based method* (DEND-CLUSTER), and (iii) *emotion intensity-based method* (EIFS). In all the methods, as a preprocessing step, the faces in all frames of the video sequence are detected, aligned, and cropped to eliminate unnecessary regions such as the background and hair. Since some face regions (e.g., upper forehead) generally do not carry information about the facial expression, we divide the face region into sub-blocks and discard the sub-blocks that are irrelevant, which are shown with black in Fig. 1. In the remaining (relevant) blocks, we extract the LPQ features to get an enhanced feature vector of the facial expression. More details of the preprocessing steps are provided in Sect. 3.

#### 2.1.1 *Maximum dissimilarity-based peak frame selection method (MAXDIST)*

This method of peak frame selection is based on the assumption that candidate peak frames are maximally dissimilar from the other frames in the sequence. Therefore, first the dissimilarity between successive frames is computed by comparing the facial features. The method sorts the frames based on their average dissimilarity score with respect to other frames and selects those frames (i.e., peak frames) that correspond to the $K$ largest average dissimilarity scores. We refer to this method as MAXDIST since peak frames are selected using a maximum dissimilarity criteria.

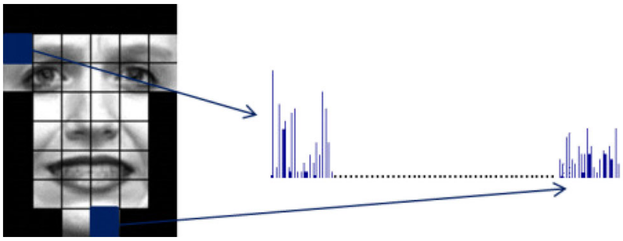

**Fig. 1** Relevant and irrelevant (*darkened*) sub-blocks of the face used for facial feature extraction in this work

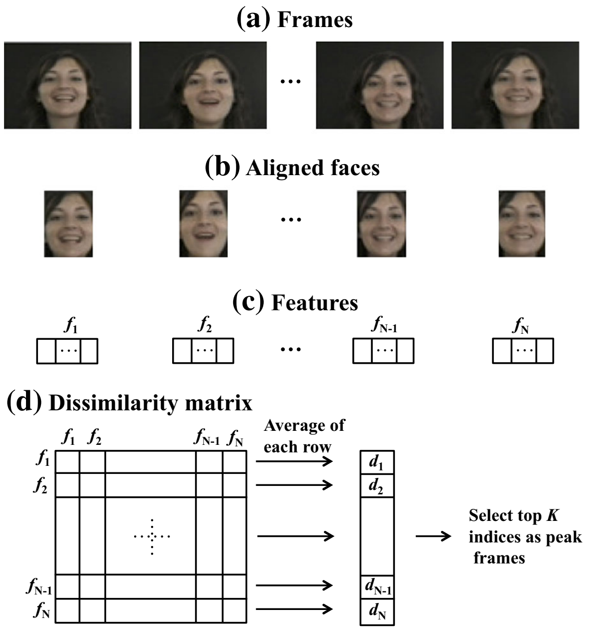

**Fig. 2** Overview of proposed MAXDIST peak frame selection method. **a** Face detection and alignment. **b** The cropped face region. **c** Extraction of facial features for each frame. **d** Calculation of dissimilarity matrix and selection of peak frame based on *top* $K$ average dissimilarity scores

Let $V = \{v_1, v_2, \ldots, v_N\}$ represent a video sequence with $N$ frames, and let $F = \{f_1, f_2, f_3, \ldots, f_N\}$ denote the facial representation features (e.g., LPQ features) of each frame. The steps of the proposed peak frame selection algorithm are as follows:

1. Generate an $N \times N$ dissimilarity matrix, $M$, where each element $M(i, j), i, j \in \{1, 2, \ldots, N\}$ is the Chi-squared distance between LPQ features of frames $i$ and $j$.
2. For the $j$th frame, compute the average dissimilarity score $(d_j)$ with respect to the remaining $(N-1)$ frames by finding the average of the elements in the $j$th row of $M$.
3. Arrange the average values obtained in step 2 in descending order and select the top $K$ frames that have the largest average dissimilarity scores as the peak frames, since they are most "dissimilar" frames in the video.

The choice for the value of $K$ is application dependent. In Fig. 2, MAXDIST peak frame selection method is illustrated. In Fig. 3, an example from subject 14 of the eNTERFACE [17] database is shown, where almost all the selected six peak frames (except frame 17) reflect the emotion disgust at its apex.

### 2.1.2 Clustering-based peak frame selection method (DEND-CLUSTER)

In this peak frame selection method, the $N$ frames of the video $V$ are grouped into $K$ clusters, so that frames within a cluster are more similar to each other than the frames belonging to different clusters. Then, for each cluster a prototype (representative) frame that typifies members of that cluster is chosen, resulting in $K$ peak frames. We refer to the above method as DEND-CLUSTER since it uses a dendrogram [11].

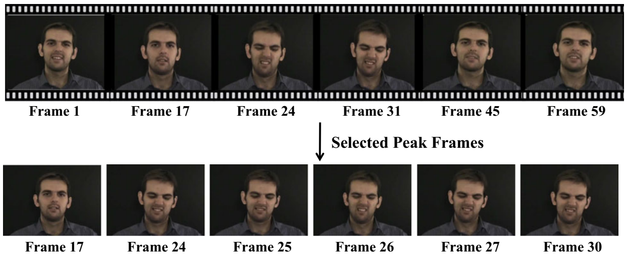

**Fig. 3** *Top row* an example sequence (with 66 frames) from eNTERFACE05 dataset for subject 14. *Bottom row* selected six peak frames by the proposed MAXDIST method reflect the emotion disgust at its apex

In order to perform the clustering, it is required to compute the dissimilarity scores between frames. First, the dissimilarity between successive frames is computed by comparing the facial features (i.e., LPQ features). We then use a hierarchical clustering method [11], since our representation of the $N$ frames is in the form of an $N \times N$ dissimilarity matrix instead of an $N \times d$ pattern matrix, where $d$ is the dimension of feature vector. In particular, we use an agglomerative complete link clustering algorithm [11]. The output of this algorithm is a dendrogram, which is a binary tree where each terminal node corresponds to a frame and the intermediate nodes indicate the formation of clusters. The set of peak frames with $K$ elements is formed as follows:

1. Find the pairwise distance scores between all $N$ frames to form the dissimilarity matrix $M$.
2. Apply the complete link clustering algorithm on $M$, and generate the dendrogram, $D$. Use the dendrogram $D$ to identify $K+1$ clusters. Since some clusters may represent low-intensity frames, we aim at $K + 1$ clusters.
3. In each of the clusters identified in step 2, select a frame whose average distance from the rest of the frames in the cluster is minimum. If a cluster has only 2 frames, choose any one of the two frames at random.
4. The frames selected in step 3 are arranged in descending order on the basis of their ideal selection measure (ISM) value, which is computed as:

$$ISM(v) = \sum_{x=1}^{P-2} \sum_{y=1}^{Q-2} G(x, y) \tag{1}$$

where $v$ is a frame of size $P \times Q$ and $G(x, y)$ is the image gradient at location $(x, y)$. The top $K$ frames are selected as peak frames. Through extensive empirical investigation, we observed that the $ISM$ value approximately increases as the emotion intensity in the frames increases.

In step 2, DEND-CLUSTER method automatically determines the threshold distance to cut the dendrogram and identify exactly $K$ clusters.

### 2.1.3 Emotion intensity-based peak frame selection method (EIFS)

This peak frame selection method is based on estimation of the unknown neutral face for a given expressive face, which will serve as a baseline to assess the intensity of the expression. The difference between the feature vectors of the expressive face and the estimated neutral face is calculated to represent the "emotion intensity" of a frame. Finally, the frames which have high "emotion intensities" are selected as peak frames. The proposed method is completely subject independent since it eliminates the requirement of a neutral expression frame of the same subject whose expression is to be recognized.

The neutral face estimation method applies KL transform [23,25] to formulate a *neutral face subspace* by eigenvector decomposition of available neutral images containing variations present in subspace of neutral faces. The neural face images are collected from other databases, which contain sufficient examples of the neutral expression such as the Cohn–Kanade database [15]. Whenever an expressive face image is projected onto this subspace, it is expressed as a linear combination of the eigenfaces modeling the space of neutral frame images. Therefore, we can synthesize a virtual neutral frame of the subject whose frame with some expression is given. Once a single virtual neutral frame image is estimated for the video, it is subtracted from each given expressive frame to estimate the emotion intensity.

The steps of the proposed emotion intensity-based peak frame selection algorithm (EIFS) are as follows:

1. First, we construct the neutral face subspace. Let $A = [\phi_1, \phi_2, \ldots, \phi_m]$ be the matrix containing neutral faces $\phi_i, i \in \{1, \ldots, m\}$ in its columns. If the image has a size of $P \times P$, then the size of $\phi_i$ is $P^2 \times 1$ and size of matrix $A$ is therefore $P^2 \times m$. The covariance matrix is formed as $C = AA^T$, and the eigenvalues $\lambda_i$ and eigenvectors are estimated using a computationally efficient method [25]. Note that since there are $m$ columns in $A$, there are at most $m$ nonzero eigenvalues of the covariance matrix $C$.
2. Project all the neutral face images onto the neutral subspace to obtain the weight vectors $\mathbf{n}_i, i = 1, \ldots, m$, where $\mathbf{n}_i$ has a size of $m \times 1$.
3. Given a video $V$ with $N$ frames $\{v_i\}, i = 1, 2, \ldots, N$, which reflect a single expression with different intensities at each frame, we want to select a single neutral image from the set of images $\phi_i$ that represents all the frames in the sequence in the best way. This is done using a two-stage approach. In the first stage, we first project each frame $v_i$ onto the neutral subspace to obtain the corresponding weight vectors $\mathbf{t}_k, k = 1, 2, \ldots, N$. Then, we select the nearest neutral image for each expressive frame by minimizing the following Euclidean distance:

$$i_k^* = \min_i \|\mathbf{t}_k - \mathbf{n}_i\|, \quad i = 1, 2, \ldots, m. \tag{2}$$

   In the second stage, the neutral frame which has been selected the most over all frames is determined as the single neutral frame "closest" to the whole sequence. Let us denote that neutral frame as $\mathbf{n}_s$.
4. Determine the peak frame(s) of the sequence based on the Euclidean distance from the neutral face $\mathbf{n}_s$. That is, we calculate

$$d(k) = \mathbf{t}_k - \mathbf{n}_s, \quad k = 1, 2, \ldots, N. \tag{3}$$

   We order $d(k)$ in descending order and select the top $K$ frames as the peak frames of the sequence, as they are expected to have highest emotion intensities.

## 2.2 Feature extraction from audio

In order to extract audio features for emotion recognition, we used mel-frequency cepstral coefficients (MFCC) and relative spectral features (RASTA) based on perceptual linear prediction (PLP) [10]. Before extraction of audio features, we first detected the starting and ending points of the speech in a given audio file. Then, we calculated the MFCC and RASTA-PLP features using filters with an order of 12 and 20, respectively, using a window of length 25 ms and a 50 % overlap ratio. Then, we appended the 12 MFCC and 13 RASTA-PLP coefficients with their first and second time derivatives. The final audio-based feature vector was extracted by applying nine statistical functions (max, min, maximum position, minimum position, mean, variance, range, kurtosis, and skewness) to the 75 elements of the MFCC and RASTA-PLP vector, thus resulting in a vector of length $75 \times 9 = 675$ as our audio feature to be classified.

## 2.3 Classification and multimodal fusion

In order to classify the audio features, we used an SVM classifier [4] with a radial basis kernel function using one-against-all method. Before the classification, we normalized audio features to the interval [0, 1]. We used an SVM classifier with a linear kernel for the video features to avoid the curse of dimensionality problem, since the dimension of video features is high (i.e., 7680).

We utilized a decision-level fusion technique, where the decision values (i.e., probabilities) for each class were estimated using each modality, separately. We tested several approaches for combining the probabilities [1,12] obtained

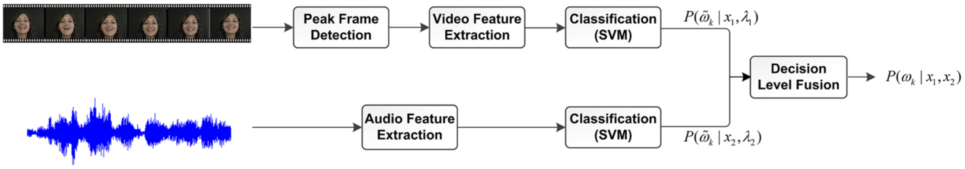

**Fig. 4** Overview of the multimodal affective and mental state recognition system

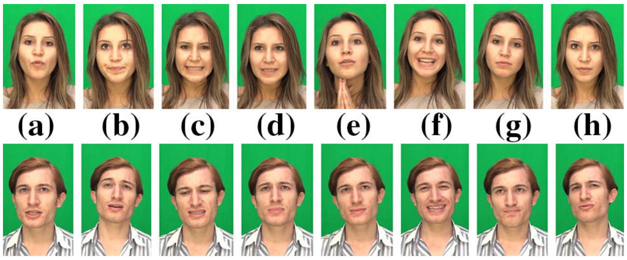

**Fig. 5** Example frames from the BAUM-1a acted database for subject 22 (*top row*) and subject 11 (*bottom row*). The *numbers in parentheses* are (sequence number-frame number of *top row*/sequence number-frame number of *bottom row*). **a** Anger (6-75/7-83), **b** boredom (9-118/10-12), **c** disgust (7-10/8-11), **d** fear (4-108/6-53), **e** interest (10-68/11-2), **f** happiness (1-49/3-74), **g** sadness (3-46/4-74), **h** unsure (8-10/11-11)

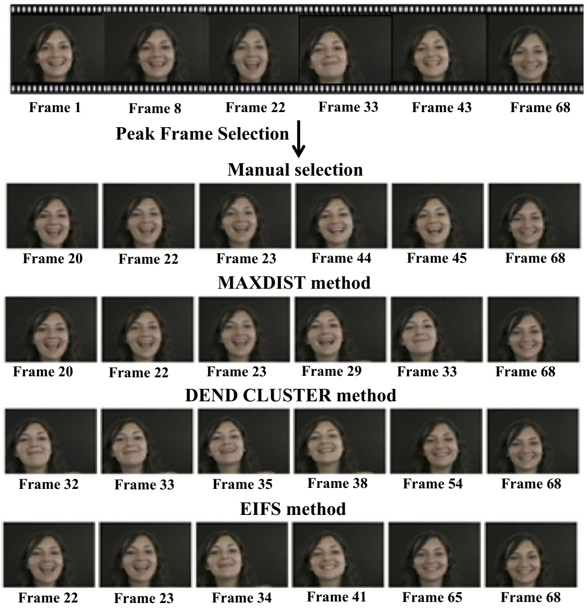

**Fig. 6** Three peak frame selection methods are compared with manual selection for an example sequence from the eNTERFACE dataset (subject 7, happiness). *Top row* shows example frames from the original sequence with 68 frames. Six peak frames have been selected manually (*second row*), using MAXDIST method (*third row*), DEND-CLUSTER method (*fourth row*), and EIFS method (*last row*)

from the SVM classifiers. The **weighted product rule** gave us the best results [27], in which the probabilities of each modality for given a test vector are multiplied. The final label is the one which gives maximum product.

Let $x_1$ and $x_2$ denote feature vectors for video and audio channels of a test video, respectively. Then, given trained SVM classifiers $\lambda_1$ and $\lambda_2$ for the audio and visual modalities, we represent the probability of class $k$ for each individual classifier as $P(\tilde{\omega}_k | x_i, \lambda_i)$ and fuse at the decision level as follows:

$$P(\omega_k | x_1, x_2) = \prod_{i=1}^{2} \left[ P(\tilde{\omega}_k | x_i, \lambda_i) \right]^{W_i}, \quad k = 1, 2, \ldots, 6 \tag{4}$$

$$\omega^* = \max_k P(\omega_k | x_1, x_2), \quad k = 1, 2, \ldots, 6 \tag{5}$$

where $\tilde{\omega}_k$ and $\omega_k$ represent the label of class $k$ before and after fusion, $\omega^*$ is the final estimated class of the test video, and $W_i$ is the weight assigned to the $i$th modality. In Fig. 4, the overview of multimodal decision fusion method is shown.

## 3 Experimental results

We conducted experiments on two databases: eNTERFACE [17] and BAUM-1a [18]. The eNTERFACE dataset contains audio-visual clips in English of 44 subjects from 14 different nationalities. One of six basic emotional states: anger, disgust, fear, happiness, sadness, and surprise is expressed in each video clip of the database in an acted way by uttering a given sentence with a target emotion by imagining a certain scenario. In the experiments, we used a total of 1287 clips from 43 subjects by omitting subject 6 due to lack of enough samples. The video clips were provided in 720 × 576 Microsoft AVI format, where the video frame rate was 25 frames/sec and the audio sampling rate was 48KHz.

The audio-visual BAUM-1a database [18] contains acted recordings of 31 subjects recorded in Turkish. This database contains recordings reflecting the five basic emotions (anger, disgust, fear, happiness, and sadness) as well as boredom and also two mental states interest and unsure (confusion). There are a total of 273 short acted recordings, and several examples are shown in Fig. 5.

During preprocessing, we first used Zhu's face tracker [31] to locate the eyes in all frames. Then, face images are scaled to obtain an inter-ocular distance of 64 pixels and cropped such that the face region has a size of 168 × 126 pixels. The face region is then divided into sub-blocks of size $8 \times 6 = 48$, and 18 sub-blocks are discarded as shown with black in Fig. 1, since they do not carry emotion-related information. In the remaining (relevant) blocks, we extract the LPQ features of length 256. Finally, the LPQ features of the 30 sub-blocks are concatenated into a single feature vector, which has a length of $30 \times 256 = 7680$.

**Table 1** Single- and multimodal emotion recognition accuracies on eNTERFACE database for all proposed peak frame selection methods and the manual peak frame selection based on leave-one-subject-out cross-validation technique

| Method | Recognition accuracy | | |
|---|---|---|---|
| | Audio (%) | Video (%) | Fused (%) |
| Manual selection | 72.95 | 47.05 | 79.57 |
| MAXDIST | 72.95 | 38.22 | 76.94 |
| DEND-CLUSTER | 72.95 | 40.00 | 78.26 |
| EIFS | 72.95 | 39.38 | 76.08 |

The decision-level fusion uses the weighted product rule with $W_1 = 1, W_2 = 2$

**Table 2** Confusion matrix for the 6 basic emotions using eNTERFACE database for the **audio** modality

| | A | D | F | H | Sa | Su |
|---|---|---|---|---|---|---|
| A | **88.37** | 1.40 | 3.26 | 2.33 | 2.33 | 2.33 |
| D | 5.58 | **71.16** | 6.98 | 4.19 | 6.05 | 6.05 |
| F | 7.44 | 9.77 | **64.19** | 4.65 | 7.44 | 6.51 |
| H | 6.98 | 4.65 | 1.86 | **75.35** | 6.98 | 4.19 |
| Sa | 3.26 | 6.51 | 5.12 | 6.05 | **72.56** | 6.51 |
| Su | 4.19 | 5.12 | 7.44 | 6.51 | 10.70 | **66.05** |

Numbers are given in percentages, and the average accuracy is 72.95 %. The row and column labels indicate true and estimated emotion labels, respectively. *A* anger, *D* disgust, *F* fear, *H* happiness, *Sa* sadness, *Su* surprise

The bold numbers indicate the correct accuracies for each emotion

**Table 3** Confusion matrix for the 6 basic emotions using eNTERFACE database for the **video** modality with DEND-CLUSTER peak frame selection method

| | A | D | F | H | Sa | Su |
|---|---|---|---|---|---|---|
| A | **25.58** | 13.02 | 11.16 | 15.35 | 12.56 | 22.33 |
| D | 5.58 | **63.72** | 7.91 | 15.81 | 4.19 | 2.79 |
| F | 17.67 | 15.81 | **13.95** | 9.77 | 23.26 | 19.53 |
| H | 6.98 | 14.42 | 6.05 | **53.49** | 3.26 | 15.81 |
| Sa | 12.56 | 11.63 | 14.88 | 6.05 | **37.67** | 17.21 |
| Su | 14.42 | 5.12 | 11.63 | 13.49 | 9.77 | **45.58** |

Numbers are given in percentages, and the average accuracy is 40.00 %

The bold numbers indicate the correct accuracies for each emotion

### 3.1 Experiments on eNTERFACE database

In Fig. 6, an example sequence expressing happy emotion is provided to visually compare the three peak frame selection methods and the manually selected peak frames. We can see that the automatic peak frame selection methods are mostly successful at selecting the frames at which emotions have maximum intensity. However, sometimes frames with lower intensities are also seen (e.g., frame 33 in DEND-CLUSTER method).

**Table 4** Confusion matrix for the 6 basic emotions using eNTERFACE database after **audio-visual** decision-level fusion

| | A | D | F | H | Sa | Su |
|---|---|---|---|---|---|---|
| A | **89.30** | 0.93 | 4.19 | 2.33 | 1.40 | 1.86 |
| D | 2.79 | **80.93** | 7.91 | 3.72 | 2.33 | 2.33 |
| F | 6.51 | 8.37 | **68.37** | 3.26 | 7.44 | 6.05 |
| H | 3.26 | 3.72 | 0.93 | **84.19** | 4.19 | 3.72 |
| Sa | 2.79 | 4.19 | 6.98 | 1.86 | **76.28** | 7.91 |
| Su | 3.26 | 2.33 | 7.44 | 5.12 | 11.16 | **70.70** |

Numbers are given in percentages, and the average accuracy is 78.26 %. The decision-level fusion uses the weighted product rule with $W_1 = 1, W_2 = 2$

The bold numbers indicate the correct accuracies for each emotion

The experimental results on eNTERFACE database for three peak frame selection methods (MAXDIST, DEND-CLUSTER, and EIFS) along with the manual peak frame selection are given in Table 1, using leave-one-subject-out cross-validation method for subject independence. The maximum and minimum number of frames per clip are 171 and 26, respectively (i.e., clip length is between 1 and 7 s). In all peak frame selection methods, we chose $K = 6$ peak frames from each video clip and used the average of the feature vectors of peak frames. Selecting multiple peak frames and averaging their visual feature vectors also help to reduce the effects of emotion-independent lip motion corresponding to speech. We can see that DEND-CLUSTER method gives the highest average emotion recognition rate for the video modality (40.0 %) even though it is not as high as emotion recognition rate when peak frames are selected manually (47.05 %). We also tested smaller or larger values of $K$ (e.g., 4, 10, 15) and received 1–2 % lower accuracies. Therefore, using $K = 6$ peak frames was reasonable for video clips of length 1–7 s. The audio modality gives an average emotion recognition rate of 72.95 % and is more successful as compared to video modality (40.0 %) on the eNTERFACE database.

The confusion matrices for the audio, video, and fused results are given in Tables 2, 3, and 4, respectively. It is easy to see that fusion always improves the recognition accuracy; the happiness and disgust benefit the most (with almost 10 % increase in their recognition rates). The comparison of our method with the existing methods is given in Table 5. We can observe that among the methods which use subject-independent cross-validation (last four rows), our method performs the best in terms of the audio-visual emotion recognition accuracy and the second best in terms of visual emotion recognition accuracy.

### 3.2 Experiments on BAUM-1a database

In this section, we present our experimental results on the BAUM-1a Turkish audio-visual database [18], which has

**Table 5** Comparison of the proposed audio-visual emotion recognition method and other methods in the literature using eNTERFACE database

| Method | Audio accuracy (%) | Video accuracy (%) | Audio-visual accuracy (%) |
|---|---|---|---|
| Paleari [19] | 35.0 | 25.0 | 67.0 |
| Wang [27] | 38.0 | 58.0 | 76.0 |
| Mansoori. [16] | 33.0 | 37.0 | 71.0 |
| Gajsek [9] | 62.9 | 54.7 | 71.3 |
| Datcu* [5] | 55.9 | 37.7 | 56.3 |
| Schuller* [22] | 72.5 | None | None |
| Kuan* [13] | 56.4 | 52.3 | 61.1 |
| Our approach* | 72.95 | 40.00 | 78.26 |

Subject-independent methods have been marked with *

**Table 6** Single- and multimodal emotion recognition accuracies on BAUM-1a database for all proposed peak frame selection methods and the manual peak frame selection based on fivefold cross-validation technique for **5 basic emotions**

| Method | Recognition accuracy | | |
|---|---|---|---|
| | Audio (%) | Video (%) | Fused (%) |
| Manual selection | 71.70 | 55.61 | 74.18 |
| MAXDIST | 71.70 | 46.60 | 74.42 |
| DEND-CLUSTER | 71.70 | 55.70 | 74.64 |
| EIFS | 71.70 | 52.06 | 74.61 |

The decision-level fusion uses the parameters $W_1 = 1$, $W_2 = 3$

**Table 7** Confusion matrix for the 5 basic emotions using BAUM-1a database for the **audio** modality

| | A | D | F | H | Sa |
|---|---|---|---|---|---|
| A | **87.88** | 6.26 | 1.82 | 0.00 | 4.04 |
| D | 10.00 | **73.33** | 5.00 | 3.33 | 8.33 |
| F | 23.69 | 8.19 | **46.62** | 7.00 | 14.50 |
| Ha | 10.00 | 6.19 | 12.38 | **60.57** | 10.86 |
| Sa | 2.00 | 2.50 | 5.33 | 0.00 | **90.17** |

Numbers are given in percentages, and the average accuracy is 71.70 %
The bold numbers indicate the correct accuracies for each emotion

**Table 8** Confusion matrix for the 5 basic emotions using BAUM-1a database for the **video** modality with DEND-CLUSTER peak frame selection method

| | A | D | F | H | Sa |
|---|---|---|---|---|---|
| A | **53.38** | 17.95 | 13.91 | 12.93 | 1.82 |
| D | 5.83 | **75.33** | 0.00 | 11.50 | 7.33 |
| F | 27.02 | 7.83 | **41.90** | 6.67 | 16.57 |
| H | 17.05 | 9.52 | 6.67 | **62.76** | 4.00 |
| Sa | 17.67 | 18.44 | 18.78 | 0.00 | **45.11** |

Numbers are given in percentages, and the average accuracy is 55.70 %
The bold numbers indicate the correct accuracies for each emotion

**Table 9** Confusion matrix for the 5 basic emotions using BAUM-1a database after **audio-visual** decision-level fusion

| | A | D | F | H | Sa |
|---|---|---|---|---|---|
| A | **87.47** | 6.26 | 2.22 | 0.00 | 4.04 |
| D | 5.00 | **80.83** | 5.00 | 5.83 | 3.33 |
| F | 27.40 | 4.50 | **59.10** | 4.50 | 4.50 |
| H | 10.00 | 6.67 | 13.99 | **59.63** | 9.71 |
| Sa | 2.00 | 2.50 | 5.33 | 4.00 | **86.17** |

Numbers are given in percentages, and the average accuracy is 74.64 %.
The decision-level fusion uses the weighted product rule with $W_1 = 1$, $W_2 = 3$
The bold numbers indicate the correct accuracies for each emotion

been recorded in our laboratories from 31 subjects. The experiments have been conducted in two parts. The first part of the experiments were done using five basic emotions (anger, disgust, fear, happiness, and sadness) so that results can be compared with the eNTERFACE database. The second part of the experiments additionally includes boredom and two mental states, interest and unsure (including confusion). In all the experiments, we used a fivefold subject-independent cross-validation strategy.

The single- and multimodal emotion recognition accuracies on BAUM-1a database for five basic emotions are given in Table 6. We can observe that the audio-based emotion recognition accuracy is 71.70 %, which is quite similar to the eNTERFACE database (73 %). The highest video-based emotion recognition rate is achieved using the DEND-CLUSTER peak frame selection method (55.70 %). The highest audio-visual emotion recognition accuracy is 74.64 % on the BAUM-1a database. The confusion matrices for the audio-only, video-only, and fusion experiments are given in Tables 7, 8, and 9, respectively. We can see that there are significant increases in the recognition rates of fear and disgust after fusion as compared to the audio modality (13 % increase in fear and 7 % increase in disgust).

The single- and multimodal emotion recognition accuracies on BAUM-1a database for 8 emotions and mental states are given in Table 10. We can observe that the audio channel has a much higher average recognition rate (63.53 %) than the video channel (36.33 %). It is worth noting that DEND-CLUSTER peak frame selection method is more successful than the manual peak frame selection method. This may be due to the fact that the subjects in BAUM-1a database were not very successful in reflecting the target emotions and mental states on their facial expressions, since they were not professional actors and were not instructed to exaggerate. They generally have subtle facial expressions, which make it difficult to recognize the expression even for humans.

## 4 Conclusion

We presented a fully automatic multimodal emotion recognition framework based on three novel peak frame selection

**Table 10** Single- and multimodal emotion recognition accuracies on BAUM-1a database for all proposed peak frame selection methods and the manual peak frame selection based on fivefold cross-validation technique for **8 basic emotions/mental states** ($W_2 = 3$ in decision-level fusion)

| Method | Recognition accuracy | | |
|---|---|---|---|
| | Audio (%) | Video (%) | Fused (%) |
| Manual selection | 63.53 | 31.32 | 65.44 |
| MAXDIST | 63.53 | 26.30 | 65.06 |
| DEND-CLUSTER | 63.53 | 36.33 | 64.80 |
| EIFS | 63.53 | 29.55 | 64.51 |

approaches from the video channel. Selection of peak frames (i.e., apex frames) is an important preprocessing step for facial expression recognition, since other frames have less information relevant to emotion to be recognized and may decrease the recognition rate. Experimental results provided on two audio-visual emotional databases: eNTERFACE [17] and BAUM-1a [18] are promising. We can achieve a subject-independent audio-visual emotion recognition rate of 78.26 % on eNTERFACE dataset, which is, to the best of our knowledge, higher than the audio-visual results reported in the literature. The visual recognition rate can be improved by using other appearance-based features for facial expression representation at the selected peak frames. The investigation of the effects of lip motion due to speech on the facial expression recognition performance is another direction for future research. We also plan to test the proposed method on spontaneous databases [7,18].

## References

1. Atrey, P.K., Hossain, M.A., Saddik, A.E., Kankanhalli, M.S.: Multimodal fusion for multimedia analysis: a survey. Multimed. Syst. **16**, 345–379 (2010)
2. Ayadi, M.E., Kamel, M.S., Karray, F.: Survey on speech emotion recognition: features, classification schemes, and databases. Pattern Recognit. **44**, 572–587 (2011)
3. Bozkurt, E., Erzin, E., Erdem, C.E., Erdem, A.T.: Formant position based weighted spectral features for emotion recognition. Speech Commun. **53**, 1186–1197 (2011)
4. Chang, C.C., Lin, C.J.: LIBSVM: a library for support vector machines. ACM Trans. Intell. Syst. Technol. **2**, 27 (2011)
5. Datcu, D., Rothkrantz, L.J.: Emotion recognition using bimodal data fusion. In: Proceedings of the international conference on computer systems and technologies, pp 122–128 (2011)
6. Erdem, C.E., Bozkurt, E., Erzin, E., Erdem, A.T.: Ransac-based training data selection for emotion recognition from spontaneous speech. In: AFFINE (2010)
7. Erdem, C.E., Turan, C., Aydin, Z.: BAUM-2: a multilingual audio-visual affective face database. Multimed. Tools Appl. **74**, 7429–7459 (2014)
8. Fasel, B., Luettin, J.: Automatic facial expression analysis: a survey. Pattern Recognit. **36**, 259–275 (2003)
9. Gajsek, R., Struc, V., Mihelic, F.: Multi-modal emotion recognition using canonical correlations and acoustic features. In: International Conference on Pattern Recognition (2010)
10. Hermansky, H., Morgan, N.: RASTA processing of speech. IEEE Trans. Speech Audio Process. **2**, 578–589 (1994)
11. Jain, A.K., Dubes, R.C.: Algorithms for Clustering Data. Prentice-Hall, Upper Saddle River (1988)
12. Kittler, J., Duin, M.H.R.P., Matas, J.: On combining classifiers. IEEE TPAMI **20**(3), 226–239 (1998)
13. Kuan-Chieh, H., et al.: Learning collaborative decision-making parameters for multimodal emotion recognition. In: IEEE International Conference on Multimedia and Expo (2013)
14. Littlewort, G.C., et al.: Automatic coding of facial expressions displayed during posed and genuine pain. Image Vis. Comput. **27**(12), 1797–1803 (2009)
15. Lucey, P., et al.: The extended cohn-kanade dataset (CK+): a complete dataset for action unit and emotion-specified expression. In: IEEE CVPR Workshop (2010)
16. Mansoorizadeh, M., et al.: Multimodal information fusion application to human emotion recognition from face and speech. Multimed. Tools Appl. **49**, 277–297 (2010)
17. Martin, O., Kotsia, I., Macq, B., Pitas, I.: The eNTERFACE05 audio-visual emotion database. In: Proceedings of the IEEE Workshop on Multimedia Database Management (2006)
18. Onder, O., Zhalehpour, S., Erdem, CE.: A Turkish audio-visual emotional database. In: IEEE signal processing and applications conference (SIU), pp. 1–4. http://www.baum1.bahcesehir.edu.tr (2013)
19. Paleari, M., Huet, B.: Toward emotion indexing of multimedia excerpts. In: Proceedings of the CBMI, pp. 425–432 (2008)
20. Ryan, A., et al.: Automated facial expression recognition system. In: IEEE ICCST, pp. 172–177 (2009)
21. Sariyanidi, E., Gunes, H., Cavallaro, A.: Automatic analysis of facial affect: a survey of registration, representation and recognition. IEEE TPAMI **37**, 1113–1133 (2014)
22. Schuller, B., et .al.: Acoustic emotion recognition: A benchmark comparison of performances. In: IEEE Workshop on Automatic Speech Recognition and Understanding, pp. 552–557 (2009)
23. Sharma, A., Anamika, D.: Facial expression recognition using virtual neutral image synthesis. In: Nat. Conf. Comp. Vi. Patt. Reco. Image Proc. and Graphics (2010)
24. Sloan, D.M., Kring, A.M.: Measuring changes in emotion during psychotherapy: conceptual and methodological issues. Clin. Psychol. Sci. Pract. **14**, 307–322 (2007)
25. Turk, M., Pentland, A.: Eigenfaces for recognition. J. Cogn. Neurosci. **3**(1), 71–86 (1991)
26. Ulukaya, S., Erdem, C.E.: Gaussian mixture model based estimation of the neutral face shape for emotion recognition. Digit. Signal Process. **32**, 11–23 (2014)
27. Wang, Y., et al.: Kernel cross-modal factor analysis for information fusion with application to bimodal emotion recognition. IEEE Trans. Multimed. **14**(3), 597–607 (2012)
28. Yongjin, W., Ling, G.: Recognizing human emotional state from audiovisual signals. IEEE Trans. Multimed. **10**, 936–946 (2008)
29. Zeng, Z.H., Pantic, M., Roisman, G.I., Huang, T.S.: A survey of affect recognition methods: audio, visual, and spontaneous expressions. IEEE TPAMI **31**(1), 38–58 (2009)
30. Zhalehpour, S., Akhtar, Z., Erdem, C.: Multimodal emotion recognition with automatic peak frame selection. In: Proceedings of IEEE INISTA, pp. 116–121 (2014)
31. Zhu, X., Ramanan, D.: Face detection, pose estimation and landmark localization in the wild. In: Computer Vision and Pattern Recognition (CVPR) (2012)