Multimed Tools Appl (2015) 74:7429–7459
DOI 10.1007/s11042-014-1986-2

BAUM-2: a multilingual audio-visual affective face database

Cigdem Eroglu Erdem · Cigdem Turan · Zafer Aydin

Published online: 9 May 2014

Abstract Access to audio-visual databases, which contain enough variety and are richly annotated is essential to assess the performance of algorithms in affective computing applications, which require emotion recognition from face and/or speech data. Most databases available today have been recorded under tightly controlled environments, are mostly acted and do not contain speech data. We first present a semi-automatic method that can extract audio-visual facial video clips from movies and TV programs in any language. The method is based on automatic detection and tracking of faces in a movie until the face is occluded or a scene cut occurs. We also created a video-based database, named as BAUM-2, which consists of annotated audio-visual facial clips in several languages. The collected clips simulate real-world conditions by containing various head poses, illumination conditions, accessories, temporary occlusions and subjects with a wide range of ages. The proposed semi-automatic affective clip extraction method can easily be used to extend the database to contain clips in other languages. We also created an image based facial expression database from the peak frames of the video clips, which is named as BAUM-2i. Baseline image and

This work was supported by the Turkish Scientific and Technical Research Council (TUBITAK) under project 110E056.

C. Eroglu Erdem (✉) · C. Turan · Z. Aydin
Department of Electrical and Electronics Engineering, Bahcesehir University, 34349, Besiktas, Istanbul, Turkey
e-mail: cigdem.eroglu@bahcesehir.edu.tr

C. Turan
e-mail: cigdem.turan@connect.polyu.hk

Z. Aydin
e-mail: zafer.aydin@agu.edu.tr

Present Address:
Z. Aydin
Department of Computer Engineering, Abdullah Gul University, Kocasinan, Kayseri, Turkey

Present Address:
C. Turan
Department of Electronic and Information Engineering, The Hong Kong Polytechnic University, Hung Hom, Kowloon, Hong Kong

video-based facial expression recognition results using state-of-the art features and classifiers indicate that facial expression recognition under tough and close-to-natural conditions is quite challenging.

Keywords Facial expression recognition · Affective database · Audio-visual affective database

1 Introduction

Perception of the emotional and mental state of a person from facial expressions and speech has a central function in successful human-to-human communication. It is envisioned that this capability will also be central in future human-computer interaction and ambient intelligence scenarios [58]. Audio-visual affect recognition has many application areas such as security [43] (lie-detection etc.), education, health-care [4, 30], multi-modal human computer interaction, marketing and advertising.

Automatic affect recognition from facial expressions and/or speech is a challenging task considering the fact that trained human observers can achieve an average correct recognition rate of 87 % [6]. Humans are very good at recognizing happiness and surprise by looking at images even with low resolution [34]. It has been reported that humans are not very good at recognizing anger and sadness and the worst at recognizing fear and disgust.

1.1 Prior work

In the last decade many studies have been published on facial expressions recognition, which are summarized in several survey papers [21, 24, 39, 40, 58]. Most of these methods use two dimensional geometric or appearance based spatial or spatio-temporal facial features, which are given to a pattern recognition algorithm for classification. Facial features extracted from images or video clips can be broadly categorized as *geometrical features* [32, 58] and *appearance based features* [29, 36]. Geometrical features consist of shapes of facial components (eyes, lips etc.) and salient points on the face (nose tip etc.). Appearance based features provide information about the texture of the face as well (e.g. natural wrinkles and creases between the eyes). Static facial expression analysis has been explored in 2D images [29] and 3D data [45, 46]. There are also video-based methods [28].

There is a growing body of research on audio-visual affect recognition, which requires realistic and richly annotated affective databases. The databases that are available to researchers can be broadly classified as acted [5, 23, 26, 32, 33, 35, 45], or naturalistic [22, 35, 53]. Acted databases are generally recorded in laboratory environments, under tightly controlled conditions. The extended Cohn-Kanade database (CK+) [32] is one of the most popular databases, which contains 123 subjects and 327 sequences labeled with the six basic emotions (anger, happiness, sadness, disgust, surprise and fear) as well as contempt. The sequences start with a neutral expression and end at the apex phase of the expression, which were posed in a laboratory environment. The Jaffe database [33], is an acted facial expression database containing 219 images from 10 Japanese females demonstrating the six basic emotions. The MMI database [41] contains laboratory recorded posed videos of facial expressions with full temporal patterns (onset-apex-offset) from 75 subjects. Some of the videos contain an additional profile recording. The Multi-PIE database [23] contains videos and images from 337 subjects from 15 view points and under 19 illumination conditions. Although the database has been collected mainly to serve research on face recognition under

varying conditions, the presence of five facial expressions make the Multi-PIE database also useful for facial expression recognition. The expressions that exist in the database are: smile, surprise, squint, disgust, scream and neutral. The GEMEP corpus [5, 50] contains recordings from 10 actors who were trained by a professional director. The actors were recorded while they are uttering combinations of pseudo-linguistic phoneme sequences. Although the GEMEP database contains over 7000 audio-visual emotion portrayals for 18 emotions, 289 of them were selected for the GEMEP-FERA challenge [50], which contain five emotions (anger, fear, joy, relief, sadness).

Collecting databases which contain spontaneous or naturalistic expressions is very difficult, time consuming, and labor intensive. The SEMAINE [35] database contains naturalistic expressions of emotions from 150 subjects who were interacting with a sensitive artificial listener. The video clips have been recorded in a laboratory, which give a lot of control over illumination and other recording conditions. The UT-Texas database [38] contains dynamic facial expression clips from 284 subjects recorded in lab conditions. The emotions are induced by watching a 10 minute video designed to elicit certain emotions. The FG-NET database [53] also contains induced emotional expressions collected from a limited number of (18) subjects. In IEMOCAP database [10], 12 hours of audio-visual data was recorded from 10 professional actors who were conducting scripted affective dyadic conversations. Their motion capture data was also recorded using a marker-based system, and the utterances were annotated using dimensional attributes (valence, activation, dominance). There are also several recent efforts to collect naturalistic 3D audio-visual [20] or 3D facial expression [59] databases, which require very expensive 3D scanners.

Today, we have access to a huge collection of movies and TV shows that contain many affective face videos, which are acted within certain contexts by mostly professional actors. Therefore, they can be considered to be more naturalistic as compared to the other acted databases in the literature and contain more variety as compared to videos recorded in a laboratory environment. Movies also reflect close-to-real-life conditions in terms of speech data since they contain background noise and sounds that simulate environmental conditions (e.g. wind, daily noise of a city), which does not exist in other lab-recorded databases.

The Vera Am Mittag (VAM) [22] database was collected by manual segmentation of a German TV talk-show and contains unscripted expressions of emotions. The Belfast naturalistic database [17] contains both clips collected manually from TV programs and those recorded during dyadic discussions. A recent work [15, 16], introduced a "facial expressions in the wild" database, which was collected from movies. The method starts by first looking for emotion-related keywords in the subtitles. Then, the video is parsed into clips, the starting and end times of which are extracted from the subtitle information. This method is limited since it depends on the existence of subtitles. Moreover, the synchronization of subtitles and facial expressions may not always be perfect, implying that the extracted clips may contain frames that do not even contain faces.

A comparison of video-based facial expression databases is shown in Table 1. As we can observe from the table, most databases in the literature have been collected manually in laboratory environments. Therefore, there is a need to collect multilingual audio-visual affective databases to aid audio-visual affect recognition research.

1.2 Contribution and scope

As mentioned in the previous section, there is a lack of sufficient close-to-natural affective face databases in the literature to test affect recognition algorithms under challenging conditions. Naturalistic databases are very difficult to collect and most of the databases in the

Table 1 Video-based facial expression databases in the literature

Database	Collect. method	Lab	# Subj.	Subj. age	Audio available	Six basic emotions	Annotation method
BAUM-2	SA	No	286	5–73	Yes	Yes	Mixed
AFEW [16]	SA	No	330	1–70	No	Yes	Cat
FGNET [53]	M	Yes	19	N/A	No	Yes	Cat
Belfast [17]	M	No	N/A	N/A	Yes	No	Con
Semaine [35]	M	Yes	75	N/A	Yes	No	Con
VAM [22]	M	No	20	N/A	Yes	Yes	Both
IEMOCAP [10]	M	Yes	10	N/A	Yes	Yes	Con
CK+ [32]	M	Yes	123	18–50	No	Yes	Cat
GEMEP [50]	M	Yes	10	N/A	No	Yes	Cat
Multi-PIE [23]	M	Yes	337	N/A	No	No	Cat
MMI [41]	M	Yes	75	19–62	No	Yes	Cat
UT-Dallas [38]	M	Yes	284	18–25	No	Yes	Cat

SA: semi-automatic, M: manual, Cat: Categorical, Con: Continuous, Lab: Recorded in a laboratory environment

literature available today are acted. Therefore, in this work, we first present a new approach for semi-automatic extraction of multilingual affective facial audio-visual video clips from movies and TV programs. The method is based on automatic detection and tracking of faces, and extraction of the corresponding audio. Then, the proposed method is used to collect a new multi-lingual database, namely BAUM-2 (see Table 1), which contains facial clips with various head poses, illumination conditions, occlusions, and subjects from various ages, and races recorded under close-to-natural conditions. The BAUM-2 database consists of clips in two languages, which can easily be extended to include other languages. An image based database has also been created, namely BAUM-2i, which contains apex frames of each clip, where the expression is at its peak. The database contains a rich set of annotations such as subject age, approximate head pose, emotion labels and intensity scores of emotions given by at least 5 annotators. Finally, we conducted image and video based facial expression recognition experiments on the BAUM-2 and BAUM-2i databases using state-of-the art feature extraction methods and classifiers, and compared the results with other databases in the literature. The experimental results verify the challenging nature of the BAUM-2 database, which makes it a valuable testbed for researchers to simulate real-life conditions.

The organization of the paper is as follows: In Section 2, an automatic method for extraction of face clips from movies is presented. In Section 3, the collected image and video-based databases are described and the novel features are listed. In Sections 4 and 5, the used feature extraction methods are described and baseline facial expression recognition experiments are given in comparison with other well-known databases in the literature. Finally, discussion and concluding remarks are presented in Section 6.

2 Automatic extraction of audio-visual facial clips from movies

Our goal is to extract audio-visual affective facial clips from a given movie or a TV program. First, we detect a face automatically, and then track it until it is occluded (i.e. can

not be tracked any more) or until a scene cut occurs. The face tracking is done using a state-of-the art method [44] with two improvements that we propose below to enhance its performance while processing a long movie. The automatically extracted candidate clips using this improved Face Tracker are either eliminated or annotated by a human labeler, which makes the overall affective facial clip extraction process semi-automatic. Below we give the details of the automatic face detection, tracking, and clip extraction steps.

2.1 Face detection and tracking

Given a movie, each frame is processed until at least one face is successfully detected on a frame. Then, the detected face(s) and a set of facial landmarks on the face are tracked in consecutive frames using a state-of-the art face tracking method based on Constrained Local Model (CLM) [44], which will be called as "Face Tracker" in the rest of the paper.

Constrained Local Model fitting is formulated as a search for the point distribution model (PDM) parameters [12], where the non-rigid deformations of the face shape is modeled linearly with a set of global rigid transformations [44]:

$$\mathbf{x}_i = s\mathbf{R}(\bar{\mathbf{x}}_i + \Phi_i \mathbf{w}) + \mathbf{t}. \tag{1}$$

In this equation, $\mathbf{x}_i = [x_i, y_i]$ represents the coordinates of the ith landmark on the image, $\bar{\mathbf{x}}_i$ represents the mean location, s denotes a global scale, $\mathbf{R}$ and $\mathbf{t}$ denote global rotation and translation parameters, and $\mathbf{w}$ represents the weights multiplying the columns of the submatrix Φ_i which represent the basis vectors for the variations of landmark i.

In constrained local model fitting, the following cost function is minimized:

$$J(\mathbf{p}) = \lambda(\mathbf{p}) + \sum_{i=1}^{n} D_i(\mathbf{x}_i, f(\mathbf{x})), \tag{2}$$

where $\mathbf{p} = \{s, \mathbf{R}, \mathbf{t}, \mathbf{w}\}$ denotes the set of PDM parameters, $\lambda(.)$ is a regularization function which penalizes complex deformations, and D_i represents the cost induced by the misalignment of the ith landmark in image $f(\mathbf{x})$.

In [44], CLM fitting is formulated using a maximum likelihood probabilistic interpretation, searching for the model parameters that minimizes the above equation. Face Tracker starts by detecting faces on each frame using the Viola-Jones (VJ) face detector [52], and then localizes 66 landmark points on the face using a regularized landmark mean-shift approach and a pre-trained deformable model. An example of the detected landmarks can be seen in Fig. 1c, where the landmarks have been drawn with red circles. The facial landmarks are tracked in consecutive frames using the configuration of the point distribution model in the previous frame as an initial estimate. For more details about Face Tracker the reader is referred to [44].

Although Face Tracker is quite successful in handling partial occlusions and varying head-poses, there are two particular situations that occur frequently in movies, where the landmark placement or tracking performance is not satisfactory. The first type of errors occur when the face detection window is not large enough to cover the whole face. Since the face is first detected using the VJ face detector, which uses Haar-like features on the luminance channel of the image, the detected face window sometimes excludes a portion of the forehead and/or the chin as shown in Fig. 1a. This may lead to incorrect localization of the landmarks as can be seen in Fig. 1b.

The second problem occurs when Face Tracker continues to track the face across a scene-cut. Although Face Tracker can handle partial occlusions by assuming that outlying

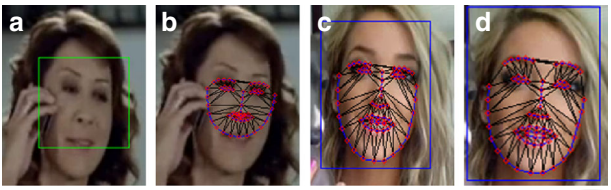

Fig. 1 **a** The detected face using the Viola-Jones (VJ) method [52] **b** The located landmarks (*red circles*) using Face Tracker [44] are not correct. **c** The VJ face detection window (*blue*) does not include the whole forehead and the landmarks (*red circles*) on the eyebrows are located incorrectly. **d** The face detection window is enlarged vertically using skin color information to include the whole face, which leads to correct landmarking

landmark candidates are consistent with the shape model, it is possible that an outlying candidate landmark does not represent a true landmark location although it is consistent with the shape model [44]. An example is given in Fig. 2a, where six consecutive frames from a movie are shown. Face Tracker continues to track the face erroneously although there is a sudden scene cut from the face of the lady to a car.

Below, we propose two improvements to overcome these two problems described above.

2.2 Improvements to Face Tracker using skin color information and SURF feature matching

Two problems of Face Tracker that we faced during automatic facial clip extraction from movies were mentioned in the previous section. Below, we propose two novel improvements

Fig. 2 **a** Six consecutive frames from a movie is shown. Face Tracker continues to track the landmarks although there is a sudden scene-cut. **b** The proposed SURF-features based method helps to stop face tracking at a scene cut. **c** *Green circles* denote matching SURF features between the current frame and the previous frame, *red circles* denote features that have not matched. There are no matching SURF features at the fourth frame from left, indicating a scene cut

to overcome these problems. These improvements enhance the face tracking performance of Face Tracker significantly in facial clip extraction experiments.

In order to solve the first problem of Face Tracker, we propose a method that uses a complementary feature based on skin color detection. Specifically, we adjust the size of the face detection window using skin color information so that the window contains all the facial features from forehead to chin. For example Fig. 1c shows the blue face detection window before color based enlargement and the resulting incorrect facial landmarks around the eyebrows, whereas Fig. 1d shows the blue face detection window enlarged by the skin color information. In order to detect the skin color we used a Bayesian classifier, which has been extensively trained [19]. After the enlargement of the face window, the tracker has a much better performance as seen in Fig. 1d, as the facial landmarks are located correctly around the eyebrows.

In order to solve the second problem of Face Tracker, we propose a scene-cut detection method based on the number of matching Speeded-Up Robust Features (SURF) [7] between two consecutive frames. SURF is a fast and robust method for interest point detection on an image, which is scale and rotation invariant. In order to guarantee invariance to scale changes, the input image is analyzed at different scales. When the number of matching SURF features is more than 40 % of the total number of SURF features located on the face in the previous frame, we accept the landmark tracking result, otherwise we claim that there is a scene-cut and stop facial landmark tracking. If there is no scene-cut or a severe occlusion, most of the SURF features match between two successive frames as shown in Fig. 3. An example of SURF based scene-cut detection is shown in Fig. 2, where the top row shows six consecutive frames of a movie across a scene-cut. The Face Tracker continues to track the facial landmarks across the scene cut, although a red car suddenly appears in the scene. The middle row of the figure shows the face tracking result by using the SURF-based scene-cut estimation method. The bottom row in the figure shows matching (green) and non-matching (red) SURF features between two consecutive frames. We can see that facial landmark tracking correctly stops at the frame where the scene-cut occurs.

2.3 Extraction of the audio-visual video clip

Given a movie, we first extract the audio as raw wav data using Audio Utils library [56]. Then, the faces are detected and tracked with the improved Face Tracker until they are out of the scene or untrackable due to occlusions. Once face tracking is completed, a new video

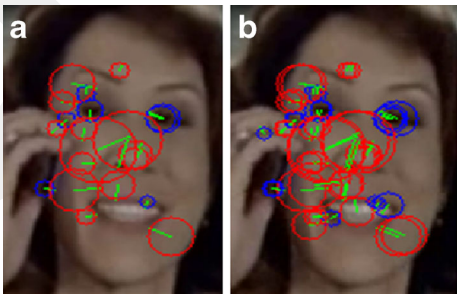

Fig. 3 **a** SURF features detected on frame 353 of the image sequence where *blue circles* indicate dark blobs on light backgrounds and *red circles* indicate light blobs on dark backgrounds. **b** The SURF features of frame 353 and 354 are shown together. The number of matched SURF descriptors are used to estimate whether there is a scene-cut or not

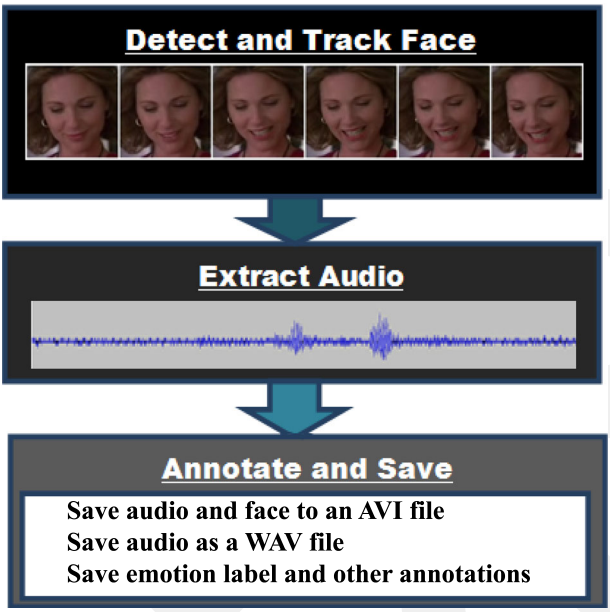

Fig. 4 The flow diagram of the automatic audio-visual clip extraction process, followed by an annotation step

clip is saved, which contains pixels of the face region tracked in each frame, merged with the corresponding audio segment extracted from the whole audio file. The tracked facial landmarks are also saved to a file. The overall flow diagram of the proposed method is summarized in Fig. 4. Several examples of extracted video clips can be seen in Fig. 7.

3 An affective audio-visual face database: BAUM-2

We applied the facial clip extraction method presented in Section 2 to create a multilingual affective face database. In total, 122 movies and TV-series have been processed with the proposed method, 97 of which are in English. The movies and TV series were selected from different genres (comedy, crime, drama, thriller (horror), action etc.) so that the targeted facial expressions for the six basic emotions (anger, happiness, sadness, disgust, surprise, fear and neutral) are represented with sufficient number of samples.

After the automatic processing, most of the extracted facial clips have been discarded during the annotation process since they are either too short, contain multiple and hard-to-recognize facial expressions (e.g. mental states such as confused, thinking etc.), occluded too much or do not have a face with a high enough resolution. Several examples for eliminated video clips are shown in Fig. 5. After this elimination step, the number of video clips that have been retained for annotation is 1047.

When several expressions are conveyed *sequentially* in the same video clip, it is segmented manually into smaller clips so that each segment has a single emotion. If two basic emotions are expressed *simultaneously* (e.g. happily surprised), we keep it. Some annotators may label it as surprise and some as happy. It is eventually given the label that receives the majority of the votes. However, the fact that two emotions may exist at the same time is also a valuable feature of the database.

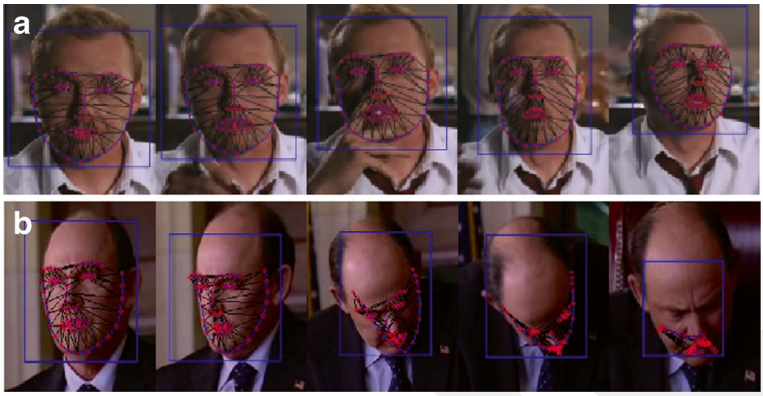

Fig. 5 Two examples of automatically extracted video clips, which are discarded during the annotation step. In (**a**) the man is smoking a cigarette causing an occlusion of the face. He is "thinking" what to say next, which is a mental state rather than an emotion. Since we do not target mental states in BAUM-2, and due to the occlusion, the clip was eliminated. In (**b**) the face is heavily self-occluded and the face tracking is not correct

We give priority to the visual data over the audio data in the sense that if the quality of the facial expression is satisfactory, the clip is included to the database, even if the accompanying speech may have a low quality or even may be totally useless. The audio data is practically useless for audio-visual emotion recognition if the person whose face is being tracked and the person who is talking in the scene at that instant are not the same, or if there is just background noise and no speech in the video clip. We indicate whether the audio is useful or not during the annotation process. We would like to note that the annotation is done by the labelers by listening the audio-visual clip, and therefore the audio data is also annotated.

3.1 Annotation of the audio-visual video clips

There are two main approaches in current literature for the annotation of audio-visual affective data. In the *categorical model* affective data is labeled using discrete categories. A popular choice of categories consist of the six universal emotions [18]: happiness, anger, surprise, sadness, disgust and fear. The categorical model is very intuitive and supported by our experience [25]. However, this model is not sufficient to describe the wide range of emotions that we face in our daily lives. Moreover, more than one emotion can exist in a single image or video. For example, surprise can co-exist with happiness causing a "happily surprised" expression. Similarly, we can talk about "sadly surprised", "angrily surprised", or "fearfully surprised" facial expressions [34].

An alternative annotation model is the *continuous model*, which overcomes some of the problems mentioned above by allowing to represent an emotion by a vector in a 2D space [42, 55]. In this model, the first dimension (evaluation or valence) measures pleasure-displeasure (positive-negativeness) and the second dimension (activation or arousal) measures the likelihood of a person to take an action in the emotional state. The continuous model has the potential to represent the variations of emotions over time allowing to represent the intensity of the emotion and mixed emotions. However, in some works

the problem of emotion recognition in 2D space is simplified as a two-class (positive-negative) or a four-class (quadrants of 2D space) problem causing the potential of the continuous model to be lost [25].

Recently, a model which is in between the above categorical and continuous models has been proposed [34], which consists of C distinct continuous spaces, where an emotion can be represented as a linear combination of C categories. A similar idea has been used in [25] to map a categorical representation to the 2D continuous space by using the confidence values of each emotion, derived from the confusion matrices.

In this work, we follow an approach that is similar to [34] and [25] for the annotation of the audio-visual clips into emotion classes. The automatically extracted video clips are annotated by 5 or 7 labelers into one of the eight emotion classes (happiness, anger, sadness, disgust, surprise, fear, contempt and neutral) after watching the audio-visual clip, as many times as the labeler needs. The labelers also indicate the intensity of the emotion in the clip by giving a score on a scale between 1 and 5, where 5 represents the highest intensity. The annotations of each labeler are fused using the majority voting rule and the scores of the labelers voting for the class winning the majority of the votes are averaged to come up with a single score for the apex frame of the emotion in the clip. If there is a tie between two emotions (e.g. a case of 3-3-1 for seven labelers), we randomly select one of the emotions that received 3 votes. We keep the labels and scores provided by each of the seven labelers, which may be useful for conducting research on mixed emotion recognition and low human agreement cases.

We evaluated the inter-annotator agreement using the Kappa statistic [51, 54], which measures how much the amount of agreement between annotators ("observed" agreement) is different from the agreement that would occur by chance alone ("expected" agreement). Kappa statistic gives a measure of this difference standardized to a scale between -1 and 1, where 1 denotes perfect agreement, 0 denotes agreement by chance, negative values indicate systematic disagreement between annotators. The Kappa statistic gave us a value of 0.55 on BAUM-2 database, which can be interpreted as moderate agreement [51] between the annotators. This is a reasonable value considering the subtlety of some of the facial expressions in the database, reflecting its challenging nature.

We also created an image-based affective face database by manually selecting a peak frame from each clip. The image based facial expression database is named as BAUM-2i. Some examples of peak frames (i.e. the image based database) can be seen in Fig. 6.

We also annotated the video clips in BAUM-2 database using the attributes given below:

1. Duration of the clip in seconds.
2. The category and the score of the emotion determined by majority voting.

Fig. 6 Example images from the BAUM-2i database showing seven different facial expressions. From *left* to *right* neutral, anger, disgust, fear, happiness, sadness, surprise. Note that the intensities of some of the facial expressions are not as high as in acted databases

3. The information whether audio information is useful or not (0: not useful, 1: useful, 2: useful but there is background music/noise to take care of).
4. The language of the speech (if speech exists).
5. The information whether the facial landmarks are tracked successfully or not (0/1), which is assessed qualitatively by visual inspection. We still keep the video clip in the database if the face is tracked correctly by a bounding window, even if the facial landmark tracking is not very successful.
6. The exact or approximate age of the actor. The age is exact if the date of birth of the actor is available (from IMDB or other web sites).
7. The name of the actor (or a unique label if the name is not known).
8. The number of the peak frame selected for the baseline static facial expressions recognition experiments as described below.
9. Head pose of the actor at the peak frame, i.e. frontal or not.
10. All the emotion labels and scores (between 1 and 5) given by 5 or 7 annotators.

3.2 Novel features of the database

The novelties of the BAUM-2 database can be highlighted as follows:

– BAUM-2 database consists of audio-visual facial clips that are more naturalistic as compared to clips recorded in laboratory environments, although they are not completely spontaneous. This is because the clips are expressed within a certain context, and the actor tries to get in the mood of the emotion of the scene to make the facial expression look believable and not exaggerated.
– BAUM-2 is a multilingual database, which to the best of our knowledge, is a feature that does not exist in other audio-visual databases.

Table 2 Basic properties of BAUM-2 database

Property	Description
Number of clips	1047
Number of subjects	286 (118 female , 168 male)
Age Range of Subjects	5–73
Number of clips per language	616 (Turkish)
	431 (English)
Clip length	0.25 - 13.72 sec, 2 sec (average)
Video format	AVI
Number of annotators	5–7
Number of clips per expression (and average scores)	Happiness: 248 (3.05/5)
	Anger: 173 (3.14/5)
	Sadness: 137 (2.81/5)
	Disgust: 51 (3.46/5)
	Surprise: 152 (3.18/5)
	Fear: 68 (3.70/5)
	Contempt: 49 (3.25/5)
	Neutral: 169 (3.2/5)

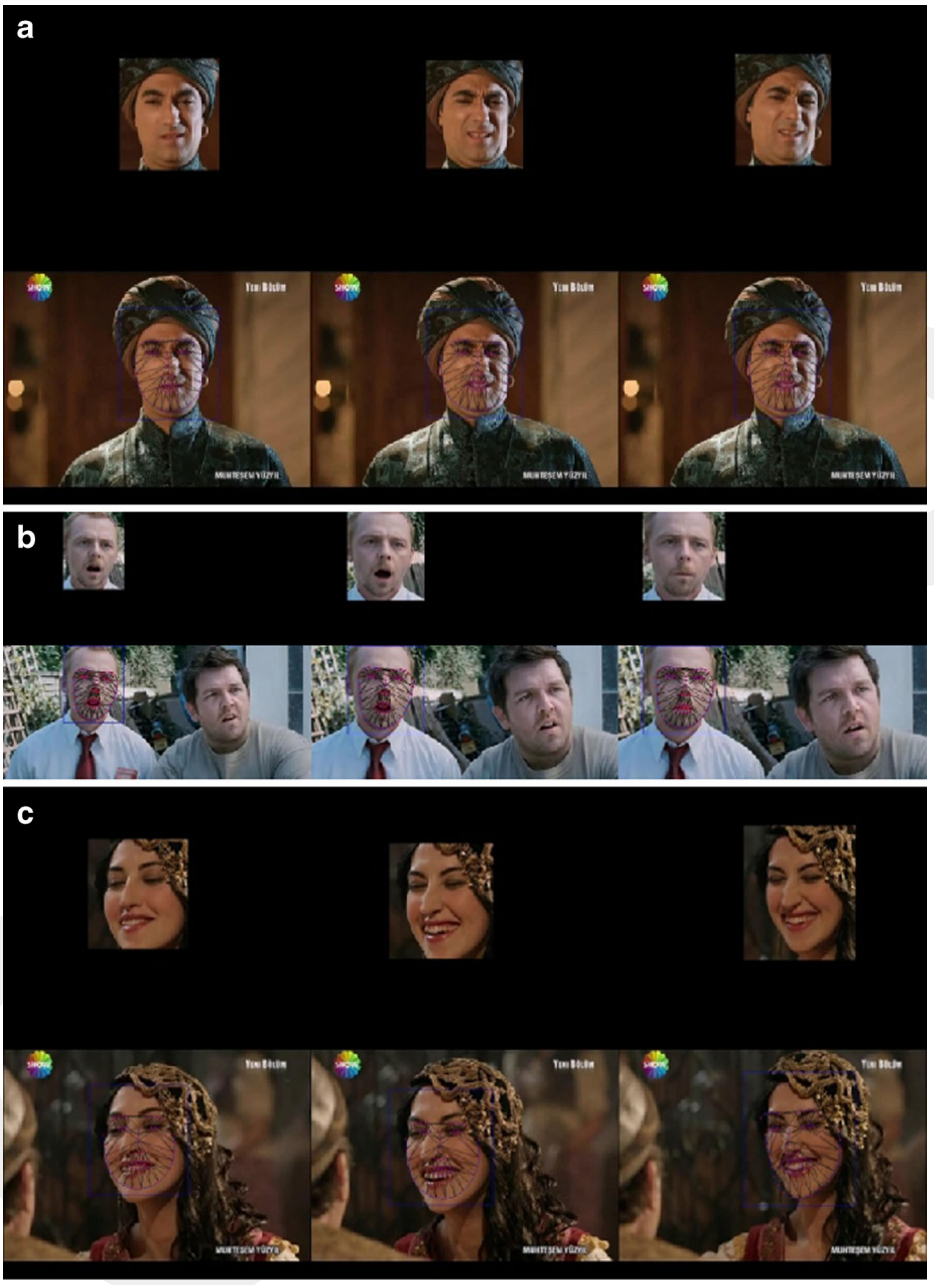

Fig. 7 Examples for annotated facial video clips (*top rows*) from the BAUM-2 database for the six basic emotions and the tracked facial landmarks (*bottom rows*). The numbers in parenthesis denote the average score given by the annotators voting for the winning class. **a** Disgust (4.2) **b** Surprise (3.9) **c** Happiness (4.3) **d** Anger (3.2) **e** Fear (3.6) **f** Sadness (4.3)

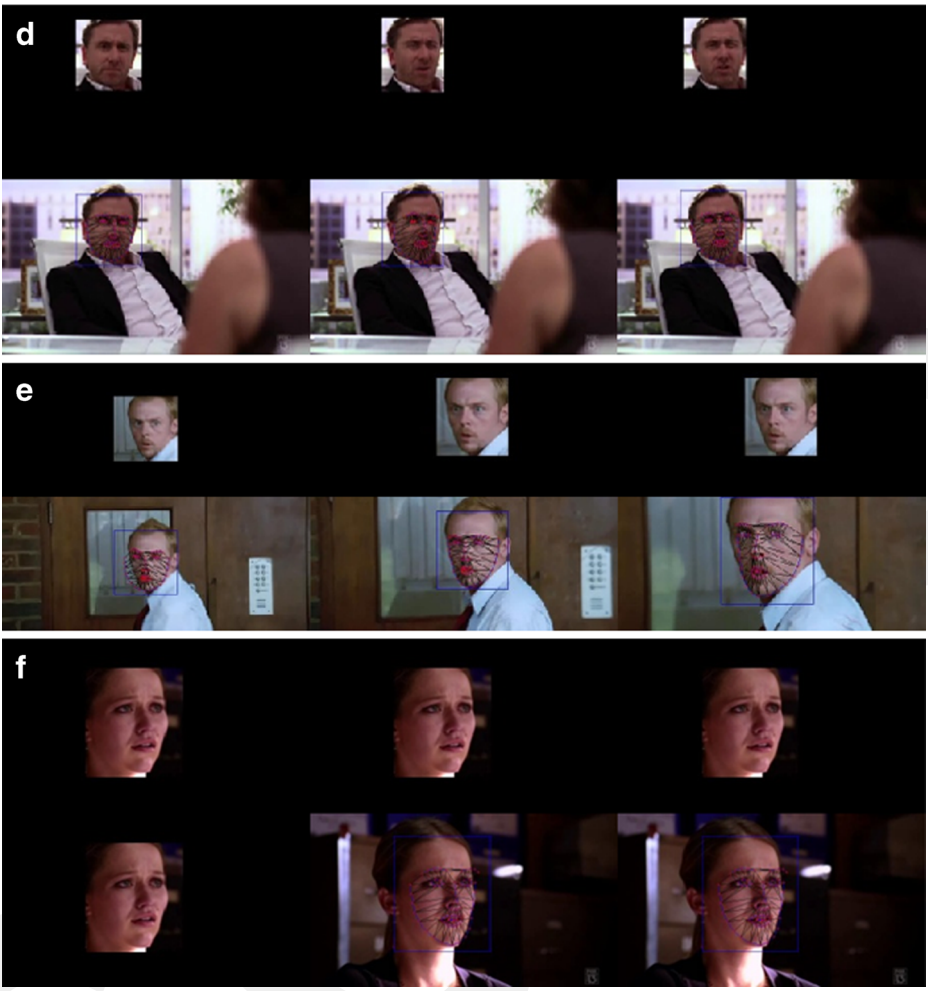

Fig. 7 (continued)

- The database is annotated using eight emotion categories, plus a score between 1 and 5, which is helpful for estimation of the intensity of the facial expression at the apex frame. This representation also has the potential to represent mixed emotions.
- The process of automatic facial clip extraction from movies does not depend on subtitles, and can be used for TV-series and programs that do not have subtitle information, which is different from [16].
- The database is richly annotated, using the attributes listed in the previous section.
- The subjects have a wide age range of 11–75, which makes the database to be useful for research on age and gender estimation.
- The database contains facial clips with uncontrolled variations of head poses, illumination conditions, accessories, makeup and image resolution, all of which impose challenges for facial expression recognition algorithms.

– The audio files are provided separately from video data as well, which will be useful for researchers working on emotion recognition from speech. Speech data generally contains background noise and hence simulate real-life conditions.
– The 2D positions of the tracked facial landmark points on the face at each frame (tracked by Face Tracker [44]) are also provided.

Some properties of the database are summarized in Table 2. Several examples of images and video clips for each emotion from the BAUM-2 database are shown in Figs. 6 and 7. It can be observed that the emotions are (e.g. surprise) are subtle and not as exaggerated as in posed databases.

4 Facial feature extraction

We also carried out baseline facial expression recognition experiments on the BAUM-2 and BAUM-2i databases using various state-of-the art facial feature extraction and classification methods, which are summarized below. We experimented with both appearance and shape-related (geometric) features for facial expression recognition (FER) and also used an approach to represent a video with a single expressive image for video-based FER.

4.1 Local Phase Quantization (LPQ) features

Local Phase Quantization (LPQ) has been originally proposed by Ojansivu et al. [37] for blur-insensitive texture classification. It has recently been successfully applied to facial expression recognition [14]. LPQ can be considered to be closely related to another statistical local feature extraction method, namely Local Binary Patterns (LBP) [36] which has also been extensively used for facial expression recognition [47, 61]. In this work, we utilize LPQ to extract appearance-based features from an expressive face since LPQ has been shown to perform better than Local Binary Patterns (LBP) for facial expression recognition [14, 16].

Let us assume that the image is degraded by centrally symmetric blur. Then, the Fourier transform of the blur function is always real and even has a zero phase at frequencies where it is positive. That means, at frequencies where the Fourier transform of the blur function is positive (e.g. frequencies smaller than the first zero crossing), the phase of the Fourier transform (FT) of the degraded image is exactly equal to the phase of the FT of the ideal image. Based on this fact, the main steps of the LPQ feature extraction method are as follows [37]:

First, 2D short-term DFT of a window W of size $M \times M$ is computed at each pixel position $\mathbf{x}$ of the image $f(\mathbf{x})$:

$$F(\mathbf{u}, \mathbf{x}) = \Sigma_{y \in W} f(\mathbf{x} - \mathbf{y}) \exp^{-j2\pi \mathbf{u}^T y} = \omega_{\mathbf{u}}^T f_{\mathbf{x}}, \tag{3}$$

where $\omega_{\mathbf{u}}$ denotes the basis vector at frequency $\mathbf{u}$ and $\mathbf{x}$ denotes the vector of image pixels.

Then, four samples of $F(\mathbf{u}, \mathbf{x})$ are considered at the frequencies $\mathbf{u}_1 = [a, 0]^T, \mathbf{u}_2 = [0, a]^T, \mathbf{u}_3 = [a, a]^T, \mathbf{u}_4 = [a, -a]^T\}$, where a is a small frequency at which the FT of the blur function is positive (i.e. has zero phase).

The real and imaginary parts of these four samples are concatenated to form a real vector of size 1×8:

$$\mathbf{G}_{\mathbf{x}} = [Re\{F(\mathbf{u}_i, \mathbf{x})\}, Im\{F(\mathbf{u}_i, \mathbf{x})\}], i = 1, 2, 3, 4. \tag{4}$$

The coefficient vectors are decorrelated using a whitening transform assuming that the image is a first order Markov process [37]. This is done since it is known that information is preserved better in quantization if the quantized coefficients are statistically uncorrelated. After decorrelation, each element is quantized to 1 if it is positive, and to 0 if it is negative:

$$q_j(\mathbf{x}) = \begin{cases} 1, & \text{if } g_j(\mathbf{x}) > 0 \\ 0, & \text{otherwise} \end{cases} \tag{5}$$

where $g_j(\mathbf{x})$ is the jth component of $\mathbf{G_x}$ after decorrelation.

The quantized local phase vectors are represented as integer values between 0 and 255:

$$Int = \sum_{j=1}^{8} q_j(\mathbf{x}) 2^{j-1}. \tag{6}$$

The 256 bin histogram of these numbers is used as a feature vector during the classification process. Note that LPQ is not affected by uniform illumination variations since only the phase information is used. For more details about the algorithm, the reader is referred to [37]. We used the implementation available from [2] and used the default parameters (window size is 3×3, DFT is calculated using a uniform window, $a = 0.7$). The dimension of the LPQ feature vector for an image becomes $256 \times 30 = 7680$, where 30 is the number of blocks used on the face to calculate the LPQ features (see Fig. 9).

4.2 Pyramid histogram of oriented gradients (PHOG) features

The histogram of oriented gradients (HOG) descriptors has been originally proposed for human detection, and is based on the idea that local appearance and shape can be represented by distribution of local edge directions with proper contrast normalization [13]. Later, it has been implemented using a spatial pyramid kernel (PHOG) and applied to object classification [8]. PHOG descriptors have also been applied for facial expression recognition [14, 15, 27]. The main steps for calculation of the PHOG descriptors are as follows:

1. Edges of the aligned and cropped faces are extracted using the Canny edge detector.
2. The image is divided into spatial cells depending on the level of the pyramid. As can be seen in Fig. 8, there are $2^l \times 2^l$ cells at level l.
3. The orientation gradients are computed using a 3×3 Sobel mask at the edge contours computed at step 1. The histograms of edge orientations are computed at each cell using K bins and an angle range of [0, 360].
4. The histogram vectors at each pyramid level are concatenated to a get a single vector. That means a vector of length K represents level 0, whereas a vector of length $4K$ represents level 1. Therefore, the length of the PHOG descriptor for the whole image becomes $K \Sigma_{l=0}^{L} 4^l$.

In the experiments, we used the implementation available from [3], with the parameters $K = 8$ and $L = 2$. The dimension of the PHOG feature vector for an image is $30 \times 168 = 5040$ where 168 comes from the concatenation of the histograms of all levels, i.e. $168 = 8 \times \Sigma_{l=0}^{2} 4^l$.

4.3 Geometric features

The geometric features consist of the x and y coordinates of the tracked 66 facial landmark points at the peak frame organized as a feature vector of length 132 as $X =$

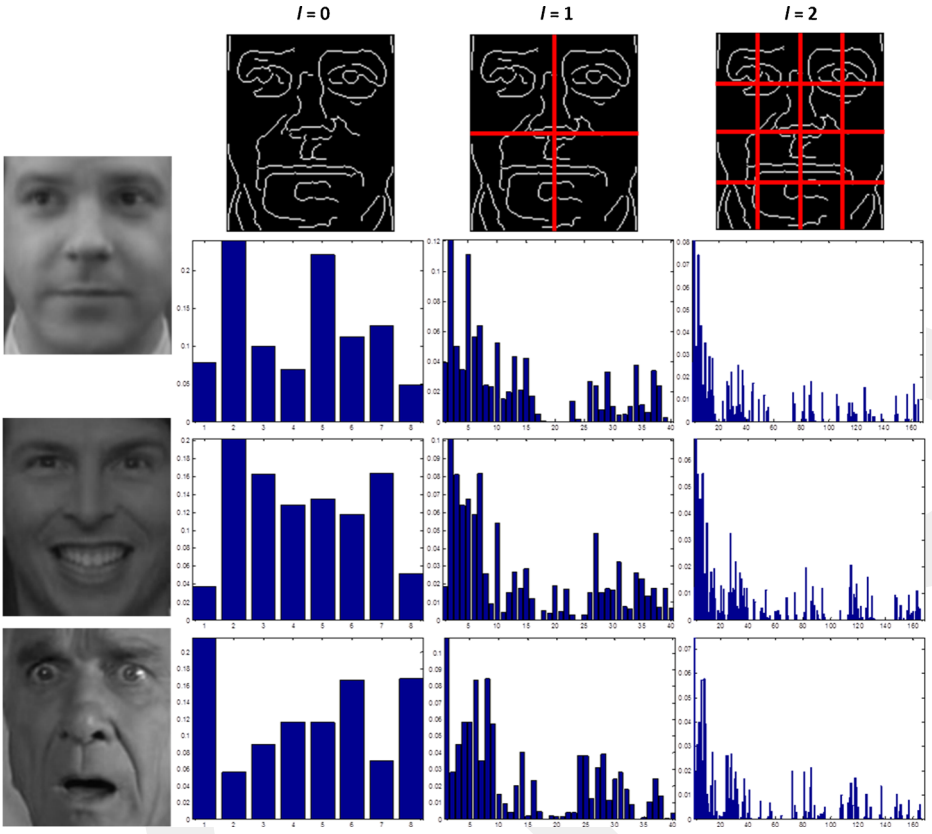

Fig. 8 Implementation of the PHOG descriptor for different number of levels

$[x_1, y_1, x_2, y_2, \ldots, x_N, y_N]$, where $N = 66$. The geometric features are first aligned such that the landmark at the nose tip is at origin, the line connecting the centers of the eyes is scaled to a fixed length, and horizontal to the x-axis [49].

4.4 Emotion Avatar Image (EAI) for video-based facial feature extraction

In order to extract facial features form a whole facial video clip (instead of the peak frame), we used a method introduced in a recent work [57], in which the main idea is to represent a dynamic facial expression with various lengths using a single image called as an "Emotion Avatar Image (EAI)" (see Fig. 11). First, the face at each frame of the video clip is detected and cropped. Then, for each video clip an EAI is estimated, which condenses a video sequence into a single expressive image. This is achieved by iteratively aligning the EAI, with the avatar reference, where the avatar reference is obtained from the aligned average of all video clips. The alignments are done using the SIFT Flow algorithm [31], which can compensate for the global motion of the head and face region while preserving the facial expression related motion. After estimation of the EAI image, the face region is divided into blocks over which LPQ or PHOG features are calculated. The person independent overall accuracy of this EAI based method has been shown to be better than the other methods in the FERA2011 challenge [1].

5 Baseline facial expression recognition experiments

In this section, we provide baseline experiments on the image-based BAUM-2i, and video based BAUM-2 databases and compare the results with other well-known databases in the literature. Below we first describe the experimentation protocols and then give image and video based facial expression recognition results.

5.1 Experimentation protocols

We used two experimentation protocols during the experiments. In the first protocol, we used 7 fold cross-validation, in which we randomly split the feature vectors into 7 folds. This gives us a train-test separation, which is a mix of common and different subjects. We call this protocol as "Partially Person Independent (PPI)" as in [15]. In the second experimentation protocol, we again used 7 fold cross validation, but this time the subjects that appeared in the training set did not appear in the test set. We enforced this condition by randomly splitting the subjects into 7 groups and then used the feature vectors of these subjects to define each fold. We call this second protocol as "Strictly Person Independent (SPI)". Note that in both protocols, we repeated the random sampling procedure if an emotion label is under-represented in any of the folds.

5.2 Image-based facial expression recognition results

We have done experiments on the BAUM-2i and four other databases from the literature, namely CK+ [26, 32], Jaffe [33], MMI [41], and SFEW [16]. The first three (CK+, Jaffe, and MMI) are popular acted databases and the last one SFEW has been recently collected from movies. The total number of images used in the experiments from each database and their distribution among the emotion classes are given in Table 14. Below, we compare the facial expression recognition results on these five databases using the feature extraction methods described in Section 4 and support vector machine based classifiers.

5.2.1 Experimental results on BAUM-2i database

The flowchart of the algorithm we followed for facial expression recognition from images is shown in Fig. 9. As mentioned earlier, the BAUM-2i dataset consists of the hand-annotated peak frames of the video clips in BAUM-2 dataset. The images are first aligned so that the distance between the centers of the eyes is 64 pixels and the line connecting the eye centers is parallel to the x-axis. Then, the face images are cropped to a size of 126×168 and divided into blocks of size 21×21. LPQ [37] or PHOG [8] features are calculated over the 30 blocks shown in Fig. 9, excluding the blocks on the periphery, since they do not carry emotion related information. The feature vectors of the 30 blocks are concatenated and classified using a Support Vector Machine (SVM) classifier.

The average accuracies of PPI and SPI protocols are given in Table 3, where the facial expression recognition results for LPQ, PHOG and Geometric features can be seen. LPQ-all and PHOG-all denote the accuracies that have been obtained using all the images in the database, regardless of the head pose, whereas LPQ-45 and PHOG-45 denote the results that have been obtained using the images with a head pose smaller than approximately 45 degrees, i.e. either frontal or close to frontal. The number of images which have a head pose angle smaller than approximately 45 degrees is 798. The total number of images used in the experiments (excluding contempt) is 998. The geometric features could be tested only on

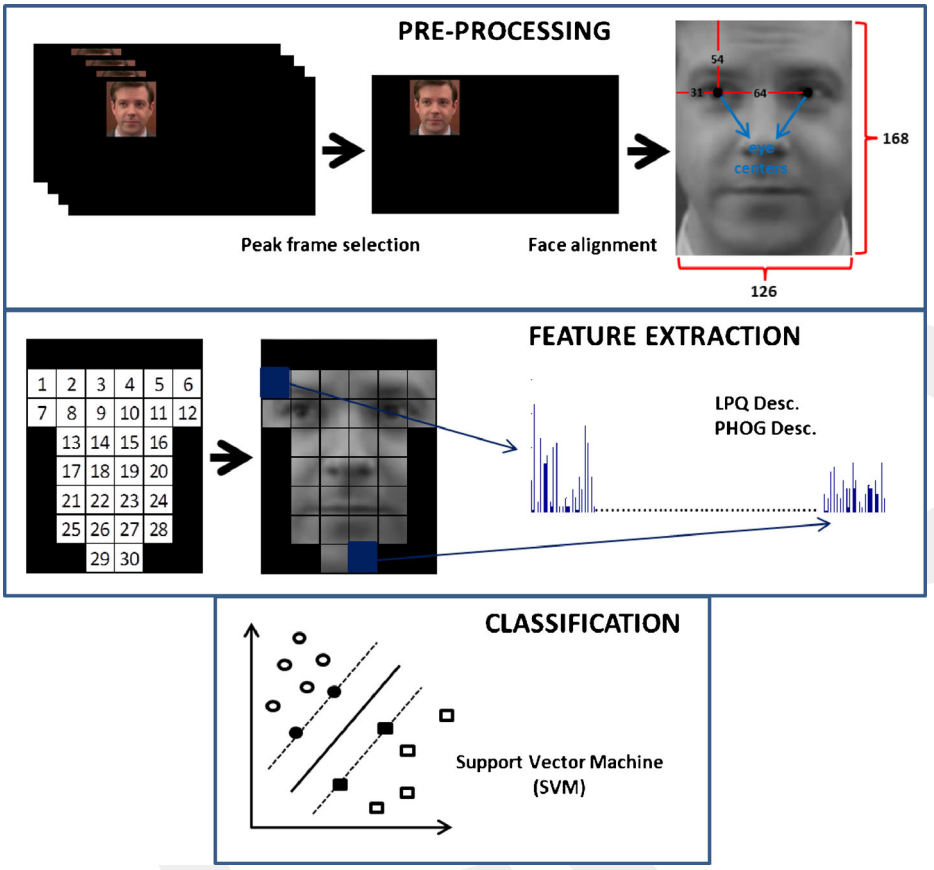

Fig. 9 Main steps of the image-based FER algorithm

536 images, on which Face Tracker gives successful landmarking results. We also applied PCA based dimensionality reduction to the LPQ and PHOG features such that 98 % of the variance in the data is kept. After dimensionality reduction, the length of LPQ features vector decreases from 7680 to 823 and the length of PHOG feature vectors decrease from 5040 to 812.

We can see from Table 3 that the highest facial expression recognition rate for the PPI protocol is achieved for LPQ-45, which is 57.77 % using one-vs-all SVM classifiers with a linear kernel [11]. On the other hand, for the SPI protocol the highest facial expression recognition (FER) rate is achieved for LPQ-45-PCA, which is 51.13 %. The SPI protocol gives lower recognition rates than PPI protocol, as expected. We can observe that using PCA reduces the dimensionality of feature vectors significantly, while keeping the accuracy almost the same (i.e. with a decrease of 1–1.5 % at most). We also tested SVM classifiers with radial basis function kernels, however the accuracies were similar to the ones obtained by a linear kernel.

The confusion matrix for the LPQ-all features under PPI protocol is given in Table 4. We can observe that the recognition rates of happiness, anger and surprise are above 50 %, whereas fear is mostly confused with surprise and disgust is mostly confused with anger.

Table 3 Results on BAUM-2i showing the average image-based facial expression recognition accuracies of compared features and experimentation protocols

	PPI protocol	SPI protocol
LPQ-45	57.77	50.50
LPQ-45-PCA	57.02	51.13
PHOG-45	54.01	47.49
PHOG-45-PCA	52.38	47.99
LPQ-all	57.62	50.10
LPQ-all-PCA	56.81	49.50
PHOG-all	52.51	47.09
PHOG-all-PCA	52.00	47.09
Geometric	49.44	47.76

Accuracies are given in percentages

Fear is mostly confused with surprise since the mouth is open in both expressions. The difference between fear and surprise expressions mainly comes from the eyebrow motion, which is generally subtle in the clips extracted from movies. Some of the clips labeled as fear has been annotated as surprise by a minority of labelers. Hence, such clips are also difficult to annotate for humans. Similarly, clips which have been labeled as disgust have received different annotations from different labelers, indicating that they are also difficult to recognize for humans, and the inter-observer agreement is low.

We would like to emphasize that the low accuracy obtained for the BAUM-2i set is not because of having inadequate number of samples from some of the emotion categories in our training sets. Table 5 shows the distribution of emotion classes at each fold of the 7-fold cross validation procedure. We can observe that the emotion distributions of the folds are similar to each other. Furthermore, there are adequate samples from each category in all training sets.

To analyze the variance of the accuracy estimates, we repeated the cross-validation experiments several times by reshuffling the data samples into 7 folds. We observed that the variance of the classifier accuracy is small (result not shown) and hence our accuracy evaluations are reliable.

Table 4 BAUM-2i confusion matrix for the LPQ-all features and an SVM classifier with the PPI protocol

	Estimated emotion						
	Neutral	Anger	Disgust	Fear	Happiness	Sadness	Surprise
Neutral	93	20	4	2	8	20	22
Anger	28	91	5	3	15	14	17
Disgust	3	11	15	0	10	9	3
Fear	7	12	0	15	0	4	30
Happiness	12	11	0	1	219	1	4
Sadness	35	26	2	3	3	58	10
Surprise	23	22	2	6	7	8	84

Table 5 BAUM-2i 7 fold cross-validation matrix, which shows the distribution of samples to each emotion class at each fold

	Neutral	Anger	Disgust	Fear	Happiness	Sadness	Surprise	Total
Fold 1	25	22	5	12	40	17	21	142
Fold 2	21	27	11	11	26	25	21	142
Fold 3	22	27	7	9	35	19	23	142
Fold 4	31	18	6	10	35	21	21	142
Fold 5	28	25	4	7	40	18	20	142
Fold 6	14	26	12	9	43	15	23	142
Fold 7	28	28	6	10	29	22	23	146

5.2.2 Experimental results on CK+ database

We also carried out experiments on the acted CK+ [32] database, for the six basic emotion classes collected from 123 subjects. The total number of images used and the distribution over the emotion classes are given in Table 14. We used the facial landmarks provided with the CK+ database to align all the images, divide the face into blocks and calculate the LPQ and PHOG features as was done for the the BAUM-2i database described above.

On the CK+ database, using 30 blocks with LPQ features, we obtained an accuracy of 93.17 % using the SPI protocol as can be seen in Table 6. Under the SPI protocol, the highest FER rate is 93.79 % using LPQ-PCA features. These results are much higher than the FER rates on BAUM2i as expected since CK+ is an acted database. The confusion matrix is given in Table 7, where we can see that even fear and disgust have been recognized with a high accuracy, since they are quite exaggerated.

5.2.3 Experiments on Jaffe database

Jaffe is one of the oldest and widely used acted facial expression databases in the literature, which contains images of 10 Japanese females for six basic emotions (see Table 14). We detected the facial features on the images using Face Tracker and used the landmarks around the eyes for image alignment. The results for the Jaffe database are given in Table 8. We can see that the FER accuracies are very high, e.g. we obtained an accuracy of 94.37 % using PHOG-PCA features. The corresponding confusion matrix is shown in Table 9 indicating that all emotions are recognized with high accuracy.

Table 6 Results on CK+ database showing the average facial expression recognition accuracies of compared features and experimentation protocols

	PPI protocol	SPI protocol
LPQ	93.17	93.48
LPQ-PCA	92.86	93.79
PHOG	88.20	90.99
PHOG-PCA	87.58	91.61

Accuracies are given in percentages

Table 7 CK+ confusion matrix using the LPQ features under the PPI protocol

	Estimated emotion					
	Anger	Disgust	Fear	Happiness	Sadness	Surprise
Anger	39	2	3	0	4	0
Disgust	3	56	0	0	2	1
Fear	0	0	22	0	0	1
Happiness	0	1	0	69	0	0
Sadness	3	0	0	0	22	0
Surprise	0	0	0	0	0	76

Table 8 Results on Jaffe database showing the average facial expression recognition accuracies of compared features and experimentation protocols

	PPI protocol	SPI protocol
LPQ	76.06	75.59
LPQ-PCA	76.06	75.59
PHOG	93.90	84.98
PHOG-PCA	94.37	84.98

Accuracies are given in percentages

Table 9 Jaffe confusion matrix using the PHOG-PCA features under the PPI protocol

	Estimated emotion						
	Neutral	Anger	Disgust	Fear	Happiness	Sadness	Surprise
Neutral	28	0	0	0	0	0	0
Anger	2	28	0	0	0	0	0
Disgust	0	0	30	0	0	1	2
Fear	0	0	0	30	0	1	1
Happiness	0	0	0	1	29	0	0
Sadness	0	1	0	0	1	29	0
Surprise	0	0	2	0	0	0	27

Table 10 Results on MMI database showing the average image-based facial expression recognition accuracies of compared features and experimentation protocols

	PPI protocol	SPI protocol
LPQ	68.75	56.25
LPQ-PCA	67.97	57.03
PHOG	71.88	57.03
PHOG-PCA	72.66	58.09

Accuracies are given in percentages

5.2.4 Experiments on MMI database

MMI is a database of acted facial expression sequences, which start and end with a neutral expression. We used all the sequences that contain an emotion label from 25 subjects (see Table 14). We detected the facial landmarks using Face Tracker, which are used for the image alignment and feature extraction process. We used the peak frames from each sequence in the experiments, which were detected by utilizing the total displacement of all facial landmarks with respect to the first (i.e. neutral frame). The automatically detected peak frames were also verified manually and corrected if necessary.

MMI is a more challenging database as compared to CK+ and Jaffe databases, since some of the facial expressions are subtle (e.g. the anger expression of subject 6). Moreover, there are subjects that wear accessories (e.g. eye glasses and head scarf). The FER rates on MMI dataset are given in Table 10, where the highest accuracy of 72.66 % was obtained using the PHOG-PCA features. The corresponding confusion matrix is given in Table 11. We can see that the results are not as high as in CK+ and Jaffe databases but much higher than the BAUM-2i database.

5.2.5 Experiments on SFEW database

We also carried out experiments on the Static Facial Expressions in the Wild (SFEW) database, which has recently been collected from movies [16] starting from the subtitles. We used all the 700 images in this database collected from 95 subjects, and applied the same 7-fold cross validation procedure as in the other databases. The results for the PPI protocol are given in Table 12. We can see that the highest recognition rate of 54.71 % has been

Table 11 MMI confusion matrix using the PHOG-PCA features under the PPI protocol

	Estimated emotion					
	Anger	Disgust	Fear	Happiness	Sadness	Surprise
Anger	15	3	1	0	0	0
Disgust	2	8	1	2	1	0
Fear	1	1	5	0	0	2
Happiness	0	2	0	14	2	0
Sadness	0	3	3	1	19	1
Surprise	0	0	9	0	0	32

Table 12 Results on SFEW database showing the average image-based facial expression recognition accuracies of compared features and experimentation protocols

	PPI protocol
LPQ	54.71
LPQ-PCA	54.29
PHOG	45.43
PHOG-PCA	39.00

Accuracies are given in percentages

obtained for the LPQ features. This is in agreement with the results obtained for BAUM-2i. The corresponding confusion matrix is given in Table 13.

5.2.6 Comparison of image based experiments

We carried out experiments on the new BAUM-2i database and four other databases in the literature. The number of images in each database used in the experiments and their distribution over the emotion classes are given in Table 14. The highest accuracy obtained for each database is shown in Table 15 and a bar plot summarizing the best results in image-based experiments is given in Fig. 10. We can conclude from the image-based experiments that the FER rates on acted databases collected in laboratory environments (CK+, Jaffe, MMI) are much higher than the databases collected from movies (BAUM-2i and SFEW). We could not observe a consistent difference between LPQ and PHOG features, since for BAUM-2i, CK+ and SFEW, LPQ features gave better results whereas for Jaffe and MMI databases PHOG features performed better.

As variety is introduced to the database, such as accessories used by some subjects in the MMI database posing a challenging situation, the recognition rate decreases. Therefore, we can say that it is still quite challenging to recognize the facial expressions using images taken under uncontrolled conditions as in BAUM-2i database.

Table 13 SFEW confusion matrix using the LPQ features under the PPI protocol

	Estimated emotion						
	Neutral	Anger	Disgust	Fear	Happiness	Sadness	Surprise
Neutral	69	7	10	11	8	6	10
Anger	5	46	2	6	14	9	1
Disgust	7	3	54	3	10	5	15
Fear	5	7	8	77	4	7	3
Happiness	11	15	5	2	42	14	15
Sadness	10	4	5	8	14	51	3
Surprise	5	3	15	7	8	7	44

Table 14 The total number of images in each database and their distribution among emotion classes

	BAUM-2i (all)	BAUM-2i (45)	CK+	Jaffe	MMI	SFEW
Anger	173	129	45	30	18	112
Disgust	51	36	59	29	17	17
Fear	68	56	25	32	32	19
Happiness	248	202	69	31	17	114
Sadness	137	108	28	31	22	99
Surprise	152	118	78	30	35	91
Neutral	169	149	–	30	–	100
Total	998	798	304	213	128	700

5.3 Video-based facial expression recognition results

Video-based facial expression recognition on BAUM-2 dataset was done by extracting the Expressive Avatar Images from each video clip, as was described in Section 4.4 (see Fig. 11). First, the face region is resampled to a size of 200×200, then the EAI is estimated and image blocks of size 20×20 are used. We discarded the blocks on the periphery of the face region since they may contain registration errors due to large out-of-plane rotation angles of the head. We used 2 levels to estimate the EAI for each video sequence. The length of the LPQ feature vector is $256 \times 64 = 16384$, and the length of the PHOG feature vector is $168 \times 64 = 10752$. These LPQ and PHOG feature vectors are then classified using an SVM classifier with linear and RBF kernels [11]. The average accuracies obtained are given in Table 16. We can observe that the average facial expression recognition rate is around 55 % for the EAI-LPQ features using a PPI protocol, which drops to 49 % for the person independent protocol. The video-based FER rates are slightly lower than the image-based FER rates, since the EAI approach is based on an averaging of all the images in the video, which diminishes the intensity of the FER in the EAI image as compared to the peak frame alone. If the person is talking, this causes an additional lip motion along with the facial expression, which affects the accuracy of FER in a negative way. The confusion matrix for the EAI-LPQ features and using an SVM classifier with an RBF kernel under the PPI protocol has been given in Table 17. We can observe that happiness (90.32 %) and anger (59.54 %) are the emotions that have the highest two accuracies. On the other hand, fear and disgust have been confused with other emotions. One reason for this is that disgust and fear expressions are not as exaggerated as in acted databases such as CK+ and Jaffe. Moreover, the EAI method performs some averaging over the clip diminishing their intensity.

The average facial expression recognition rate on the acted CK+ database has been reported as 82.6 % [57] using the EAI method (with an SVM classifier with a linear kernel), which is much higher as compared to the BAUM-2 database.

Table 15 Summary of image-based FER experiments

	BAUM-2i	CK+	Jaffe	MMI	SFEW
FER rate	57.62	93.17	94.37	72.66	54.71

Highest recognition rates (in percentages) for the PPI protocol are given for each database

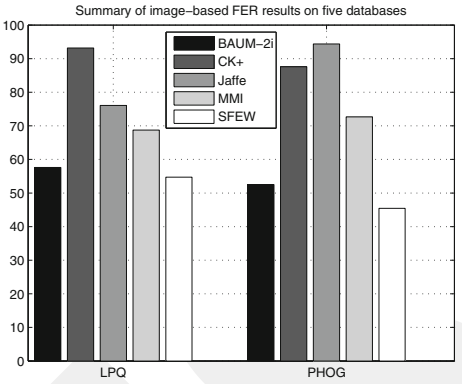

Fig. 10 Comparison of FER results on the five databases for LPQ and PHOG features

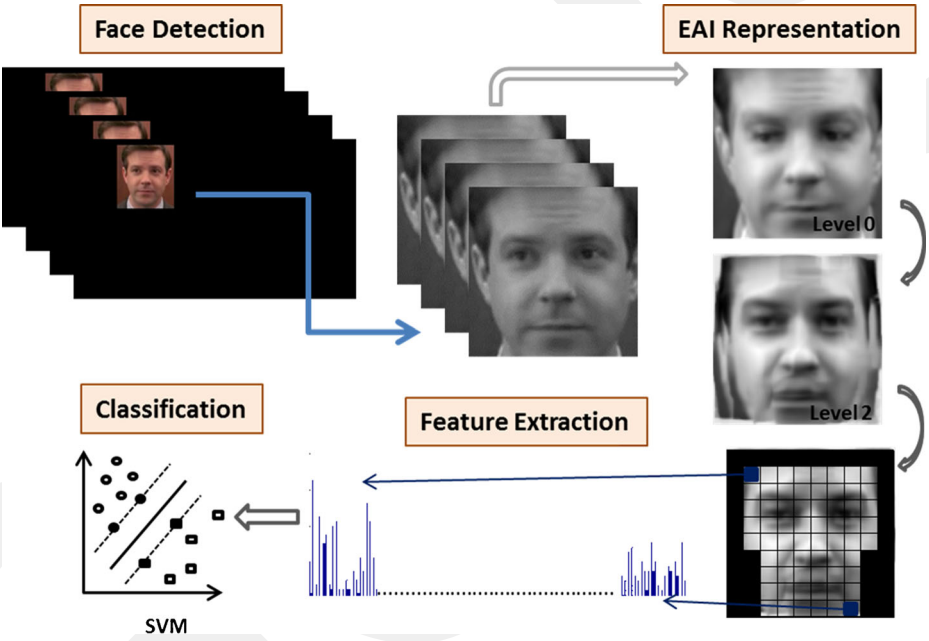

Fig. 11 Main steps of the video-based FER algorithm

Table 16 Average video based facial expression recognition accuracies of compared features and classifiers

	7 fold CV (PPI)		7 fold Subj. Indep. CV (SPI)	
	SVML	SVMR	SVML	SVMR
EAI-LPQ	55.31	55.71	48.89	48.99
EAI-PHOG	50.60	51.70	44.28	45.09

The acronym SVML stands for SVM classifier with linear kernel, SVMR stands for SVM classifier with RBF kernel

Table 17 Confusion matrix for the EAI-LPQ-all features using an SVM classifier with an RBF kernel using the PPI protocol

	Estimated emotion						
	Neutral	Anger	Disgust	Fear	Happiness	Sadness	Surprise
Neutral	97	26	0	0	7	19	20
Anger	26	103	1	0	15	10	18
Disgust	7	15	4	1	13	5	6
Fear	9	17	0	1	4	7	30
Happiness	7	8	0	1	224	5	3
Sadness	35	23	0	3	11	56	9
Surprise	30	29	2	4	9	7	71

5.4 Availability

The database as well as the source code of the method for automatic facial clip extraction from movies are available to researchers upon request through the web site [48] (to be used for research purposes only).

6 Discussion and conclusion

We presented a method for semi-automatic extraction of affective audio-visual facial clips from movies and TV shows and used it to collect two naturalistic affective video and image based face databases. The method, which is based on face detection and facial landmark tracking, can be used on any movie or TV program, independent from the language and the existence of the subtitles.

The collected multilingual database, namely BAUM-2, contains facial expressions with a diverse range of illumination conditions, head poses, occlusions and accessories, which makes it a challenging new data set in this field. The database is richly annotated and contains data from 286 people with a wide range of ages, making it also suitable for conducting research on age and gender estimation.

The video clips in BAUM-2 database have been annotated categorically with an additional score indicating the intensity of the facial expression in that category. The availability of these scores might be useful for conducting research with facial clips containing multiple emotions. For example, if two emotions are expressed simultaneously (e.g. happily surprised, fearfully surprised etc.), we keep it in the database. Some annotators may label it as surprise and some as happy. It is eventually given the label that receives the majority of the votes. This may be viewed as having noisy labels for some clips, which may decrease the facial expression recognition rate. However, the fact that two emotions may exist at the same time is also a valuable feature of the database. The analysis of several simultaneous emotions (e.g. using the labels and scores of all annotators) to design multi-label classifiers has been left as future work.

We evaluated the collected database by performing experiments using state-of-the-art facial feature extraction methods and classifiers. There are many facial expression recognition methods in the literature. We selected four of these methods, as described in Section 4, since they represent the geometric and appearance based features of the face and have been

shown to have good performance in several recent works [15, 16, 32]. The EAI method was chosen since it was the winner of a recent emotion recognition challenge [57]. However, it has its own restrictions such as diminishing the maximum intensity of the emotion in the clip due to averaging. Facial feature extraction is still an active research area.

Baseline facial expression recognition experiments on BAUM-2 indicate that facial expression recognition rates decrease dramatically, when the recording conditions are close to real world conditions with varying illumination, head pose and the existence of lip motion. This is consistent with another database obtained from movies [16]. Hence, further research on facial expression recognition is required to handle the above though conditions. Further research is also required to take into account the deformation of the lip shape when the person is talking. Recognition of emotions from noisy speech data in multiple languages and the fusion of audio and visual modalities is a possible future research direction [9]. In the future, we plan to extend the database by including clips that contain expressions of mental states (e.g. thinking, confused, interested etc.), and some mild occlusions on the face (e.g. hand occlusions).

Another possible future research direction could be to use the automatic audio-visual clip extraction presented in this paper in an active learning setting similar to recent methods proposed for the speech modality [60] to overcome the data scarcity problem in this field. We believe the BAUM-2 database will serve as a valuable resource for researchers working on affect recognition.

Acknowledgments Portions of the research in this paper use the MMI-FacialExpression Database collected by M. Pantic and her group (www.mmifacedb.com).

References

1. FG 2011 facial expression recognition and analysis challenge (FERA 2011), Available [online]. http://sspnet.eu/fera2011/
2. Machine vision group, MATLAB codes for local phase quantization. http://www.cse.oulu.fi/CMV/Downloads/LPQMatlab. Last Accessed: 01/07/2013
3. Phog implementation. http://www.robots.ox.ac.uk/~vgg/research/caltech/phog.html. Last Accessed: 01/07/2013
4. Ashraf AB, Lucey S, Cohn JF, Chen T, Ambadar Z, Prkachin KM, Solomon PE (2009) The painful face—pain expression recognition using active appearance models. Image Vis Comput 27(12):1788–1796
5. Banziger T, Scherer KR (2010) Blueprint for affective computing: a sourcebook, In: Introducing the Geneva multimodal emotion portrayal (GEMEP) corpus. Oxford University Press, pp 271–294
6. Bassili J (1979) Emotion recognition: the role of facial movement and the relative importance of upper and lower areas of the face. J Pers Soc Psychol 37:2049–2058
7. Bay H, Ess A, Tuytelaars T, Gool LV (2008) Surf: speeded up robust features. Comput Vision Image Underst (CVIU) 110(3):346–359
8. Bosch A, Zisserman A, Munoz X (2007) Representing shape with a spatial pyramidal kernel. In: Proceedings of ACM international conference on image and video retrieval, CIVR 2007, pp 401–408
9. Bozkurt E, Erzin E, Erdem CE, Erdem AT (2011) Formant position based weighted spectral features for emotion recognition. Speech Commun 53(9–10):1186–1197
10. Busso C, Bulut M, Lee CC, Kazemzadeh A, Mower E, Kim S, Chang JN, Lee S, Narayanan SS (2008) Iemocap: interactive emotional dyadic motion capture database. J Lang Resour Eval 42(4):335–359
11. Chang CC, Lin CJ (2011) LIBSVM: a library for support vector machines. ACM Trans Intell Syst Technol 2:27:1–27:27
12. Cootes T, Taylor C (1992) Active shape models. In: British machine vision conference (BMVC'92), pp 266–275

13. Dalal N, Triggs B (2005) Histograms of oriented gradients for human detection. In: Proceedings of international conference on computer vision and pattern recognition (CVPR), pp 886–893
14. Dhall A, Goecke R, Gedeon T (2011) Emotion recognition using PHOG and LPQ features. In: Proceedings of the workshop on facial expression recognition and analysis challenge FERA2011, IEEE automatic face and gesture recognition conference FG2011. Santa Barbara
15. Dhall A, Goecke R, Lucey S, Gedeon T (2011) Static facial expression analysis in tough conditions: Data, evaluation protocol and benchmark. In: IEEE international workshop on bechmarking facial image analysis technologies BeFIT, ICCV
16. Dhall A, Goecke R, Lucey S, Gedeon T (2012) Collecting large, richly annotated facial-expression databases from movies. IEEE Multimed 19(3):34–41
17. Douglas-Cowie E, Cowie R, Schoder M (2000) A new emotion database: considerations, sources and scope. In: Proceedings of ISCA ITRW on speech and emotion, pp 39–44
18. Ekman P, Friesen WV (1976) Pictures of facial effect, Consulting Psychologists Press, Palo Alto
19. Erdem CE, Ulukaya S, Karaali A, Erdem AT (2011) Combining haar feature and skin color based classifiers for face detection. In: IEEE 36th international conference on acoustics, speech and signal processing (ICASSP 2011). Prague
20. Fanelli G, Gall J, Romsdorfer H, Weise T, Gool LV (2010) A 3-d audio-visual corpus of affective communication. IEEE Trans Multimed 12(6):591–598
21. Fasel B, Luettin J (2003) Automatic facial expression analysis: a survey. Pattern Recogn 36:259–275
22. Grimm M, Kroschel K, Narayanan S (2008) The Vera am Mittag German audio-visual emotional speech database. In: Proceedings of international conference multimedia and expo (ICME)
23. Gross R, Matthews I, Cohn JF, Kanade T, Baker S (2010) Multi-PIE. Image Vis Comput 28(5):807–813
24. Gunes H, Pantic M (2010) Automatic, dimensional and continuous emotion recognition. Int J Synth Emot 1(1):68–99
25. Hupont I, Baldassarri S, Cerezo E (2013) Facial emotional classification: from a discrete perspective to a continous emotional space. Pattern Anal Appl 16(1):41–54
26. Kanade T, Cohn JF, Tian Y (2000) Comprehensive database for facial expression analysis. In: Proceedings of the Fourth IEEE International Conference on Automatic Face and Gesture Recognition (FG'00). Grenoble, France, pp 46–53
27. Li Z, Imai JI, Kaneko M (2009) Facial component based bag of words and PHOG descriptor for facial expression recognition. In: Proceedings of IEEE international conference on systems, man and cybernetics
28. Littlewort G, Bartlett MS, Fasel I, Susskind J, Movellan J (2006) Dynamics of facial expression extracted automatically from video. Image Vis Comput 24(6):615–625
29. Littlewort G, Whitehill J, Wu T, Fasel I, Frank M, Movellan J, Bartlett M (2011) The computer expression recognition toolbox (CERT). In: IEEE conference on automatic face and gesture recognition (FG 2011)
30. Littlewort GC, Bartlett MS, Lee K (2009) Automatic coding of facial expresssions displayed during posed and genuine pain. Image Vis Comput 27(12):1797–1803
31. Liu C, Yuen J, Torralba A (2011) SIFT flow: dense correspondence across scenes and its applications. IEEE Trans Pattern Anal Mach Intell 33(5):978–994
32. Lucey P, Cohn JF, Kanade T, Saragih J, Ambadar Z, Matthews I (2010) The extended Cohn-Kanade dataset (CK+): a complete dataset for action unit and emotion-specified expression. In: Proceedings of IEEE workshop on CVPR for human communicative behavior analysis. San Francisco
33. Lyons MJ, Akamatsu S, Kamachi M, Gyoba J (1998) Coding facial expressions with gabor wavelets. In: Proceedings of 3rd IEEE international conference on automatic face and gesture recognition, pp 200–205
34. Martinez A, Du S (2012) A model of the perception of facial expressions of emotion by humans: research overview and perspectives. J Mach Learn Res 13:1589–1608
35. Mckeown G, Valstar MF, Cowie R, Pantic M, Schroeder M (2012) The SEMAINE database: annotated multimodal records of emotionally coloured conversations between a person and a limited agent. IEEE Trans Affect Comput 3(1):5–17
36. Ojala T, Pietikainen M, Maenpaa T (2002) Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. IEEE Trans Pattern Anal Mach Intell 24(7):971–987
37. Ojansivu V, Heikkil J (2008) Blur insensitive texture classification using local phase quantization. Lect Notes Comput Sci 5099:236–243
38. O'Toole AJ, Harms J, Snow SL, Hurst DR, Pappas MR, Ayyad JH, Abdi H (2005) A video database of moving faces and people. IEEE Trans Pattern Anal Mach Intell 27(5):812–816
39. Pantic M (2009) Machine analysis of facial behaviour: naturalistic and dynamic behaviour. Philos Trans R Soc B-Biol Sci 364(1535):3505–3513

40. Pantic M, Rothkrantz L (2000) Automatic analysis of facial expressions: the state of the art. IEEE Trans Pattern Anal Mach Intell 22(12):1424–1445
41. Pantic M, Valstar MF, Rademaker R, Maat L (2005) Web-based database for facial expression analysis. In: Proceedings of IEEE international conference on multimedia and expo (ICME'05). Amsterdam http://www.mmifacedb.com/
42. Rusell JA (1980) A circumplex model of affect. J Personal Social Psychol 39:1161–1178
43. Ryan A, Cohn J, Lucey S, Saragih J, Lucey P, la Torre FD, Rossi A (2009) Automated facial expression recognition system. In: Proceedings of the international Carnahan conference on security technology, pp 172–177
44. Saragih JM, Lucey S, Cohn JF (2011) Deformable model fitting by regularized landmark mean-shift. Int J Comput Vis (IJCV) 91:200–215
45. Savran A, Alyuz N, Dibeklioğlu H, Celiktutan O, Gökberk B, Sankur B, Akarun L (2008) Bosphorus database for 3D face analysis. In: First COST 2101 workshop on biometrics and identity management (BIOID 2008)
46. Savran A, Sankur B, Bilge MT (2012) Comparative evaluation of 3D versus 2D modality for automatic detection of facial action units. Pattern Recogn 45(2):767–782
47. Shan C, Gong S, McOwan PW (2009) Facial expression recognition based on local binary patterns: a comprehensive study. Image Vis Comput 27:803–816
48. Turan C, Kansin C, Erdem CE (2013) Bahcesehir University multimodal affective database (BAUM-2). http://baum2.bahcesehir.edu.tr/
49. Ulukaya S, Erdem CE (2012) Estimation of the neutral face shape using gaussian mixture models. In: IEEE international conference on acoustics, speech and signal processing (ICASSP 2012). Kyoto, 1385–1388
50. Valstar MF, Jiang B, Mehu M, Pantic M, Scherer KR (2011) The first facial expression recognition and analysis challenge. In: IEEE international conference face and gesture recognition (FG'2011)
51. Viera AJ, Garrett JM (2005) Understanding interobserver agreement: the kappa statistic. Fam Med 37(5)
52. Viola P, Jones MJ (2004) Robust real-time face detection. Int J Comput Vis 57(2):137–154
53. Wallhoff F (2006) Facial expressions and emotion database [online]. Available: http://www.mmk.ei.tum.de/~waf/fgnet/feedtum.html
54. Watkins MW, Pacheco M (2000) Interobserver agreement in behavioral research. J Behav Educ 10(4):205–212
55. Whissell C Emotion: theory, research and experience. The measurement of emotions, vol. 4, chap. The dictionary of affect in language. Academic, New York
56. Wischik L Avi utils. http://www.wischik.com/lu/programmer/avi_utils.html. Last Accessed: 01/07/2013
57. Yang S, Bhanu B (2012) Understanding discrete facial expressions in video using an emotion avatar image. IEEE Trans Syst Man Cybern - Part B: Cybern 42(4):980–992
58. Zeng ZH, Pantic M, Roisman GI, Huang TS (2009) A survey of affect recognition methods: audio, visual, and spontaneous expressions. IEEE Trans Pattern Anal Mach Intell 31(1):39–58
59. Zhang X, Yin L, Cohn J, Canavan S, Reale M, Horowitz A, Liu P (2013) A highresolution spontaneous 3D dynamic facial expression database. In: International conference on automatic face and gesture recognition (FG'13). Shanghai
60. Zhang Z, Schuller B (2012) Active learning by sparse instance tracking and classifier confidence in acoustic emotion recognition. In: ISCA (ed) Proceedings of INTERSPEECH. Portland
61. Zhao G, Pietikainen M (2007) Dynamic texture recognition using local binary patterns with an application to facial expressions. IEEE Trans Pattern Anal Mach Intell 29(6):915–928

Cigdem Eroglu Erdem received the B.S. and M.S. degrees in Electrical and Electronics Engineering from Bogazici University, Ankara, Turkey in 1995 and 1997, respectively. She received the Ph.D. degree in Electrical and Electronics Engineering from Bogaziçi University, Istanbul, in 2002. From September 2000 to June 2001, she was a visiting researcher in the Department of Electrical and Computer Engineering, University of Rochester, NY, USA. Between 2003–2004, she was a postdoctoral fellow at the Faculty of Electrical Engineering at Delft University of Technology, the Netherlands, where she was also affiliated with the video processing group at Philips Research Laboratories, Eindhoven. She is currently an associate professor in the Department of Electrical and Electronics Engineering at Bahcesehir University, Istanbul, Turkey. Her research interests are in the areas of digital image, video and speech processing, including audio-visual affect recognition, motion estimation, video segmentation and object tracking.

Cigdem Turan received the Bachelors degree in Electrical-Electronics Engineering from Bahcesehir University in 2013. Currently she is pursuing a PhD degree at Hong Kong Polytechnic University, Hong Kong, China. Her research interests are in pattern recognition and image processing with applications to face and facial expression recognition.

Zafer Aydin received the B.Sc and M.Sc. degrees in Electrical Engineering from Bilkent University, Ankara, Turkey in 1999 and 2001, respectively. He then enrolled in the PhD program of the same department in Bilkent University and worked as a teaching assistant for one year. Starting from 2002 he worked as a Graduate Research Assistant and received the PhD degree in Electrical Engineering from Georgia Institute of Technology, Atlanta, GA USA in 2008. He continued his career by working as a post-doctoral fellow in the Department of Genome Sciences at University of Washington, Seattle, WA USA for three years where he did research on bioinformatics and computational biology. From September 2011 to Februray 2014, he worked as an assistant professor in Electrical and Electronics Engineering Department of Bahcesehir University, Istanbul, Turkey. Currently he is an assistant professor in Computer Engineering Department of Abdullah Gul University, Kayseri, Turkey.