BAUM-1: A Spontaneous Audio-Visual Face Database of Affective and Mental States

Article *in* IEEE Transactions on Affective Computing · April 2016

DOI: 10.1109/TAFFC.2016.2553038

4 authors, including:

Zahid Akhtar
State University of New York Polytechnic Institute
98 PUBLICATIONS **945** CITATIONS

SEE PROFILE

Cigdem Erdem
Marmara University
56 PUBLICATIONS **910** CITATIONS

SEE PROFILE

BAUM-1: A Spontaneous Audio-Visual Face Database of Affective and Mental States

Sara Zhalehpour, Onur Onder, Zahid Akhtar, and Cigdem Eroglu Erdem, *Member, IEEE*

Abstract—In affective computing applications, access to labeled spontaneous affective data is essential for testing the designed algorithms under naturalistic and challenging conditions. Most databases available today are acted or do not contain audio data. We present a spontaneous audio-visual affective face database of affective and mental states. The video clips in the database are obtained by recording the subjects from the frontal view using a stereo camera and from the half-profile view using a mono camera. The subjects are first shown a sequence of images and short video clips, which are not only meticulously fashioned but also timed to evoke a set of emotions and mental states. Then, they express their ideas and feelings about the images and video clips they have watched in an unscripted and unguided way in Turkish. The target emotions, include the six basic ones (happiness, anger, sadness, disgust, fear, surprise) as well as boredom and contempt. We also target several mental states, which are unsure (including confused, undecided), thinking, concentrating, and bothered. Baseline experimental results on the BAUM-1 database show that recognition of affective and mental states under naturalistic conditions is quite challenging. The database is expected to enable further research on audio-visual affect and mental state recognition under close-to-real scenarios.

Index Terms—Facial expression recognition, audio-visual affective database, emotional corpora, mental state recognition, affective computing, spontaneous expressions, emotion recognition from speech, dynamic facial expression database

1 Introduction

Non-verbal messages constitute an important part of human-to-human communication, since they enhance or modify our verbal messages and convey information about our emotional and mental states. Major components of non-verbal messages are facial expressions, body/head gestures and the paralinguistic properties of speech. The importance of emotions in our daily interactions motivated researchers to design and implement automatic algorithms to recognize emotional expressions with the ultimate goal of achieving intelligent human-machine interaction [1], [2]. As a result, the field of affective computing has attracted a lot of attention from researchers in the last decade due to its wide range of application areas including multi-modal human computer interaction, security [3] (lie-detection etc.), education, health-care [4], [5], marketing and advertising.

There are six basic facial expressions that have been shown to be universal, which are happiness, surprise, anger, sadness, fear and disgust [6]. It has been reported that humans can recognize emotions such as happiness and surprise easily even from low resolution images [7]. However, their recognition accuracy is very low for anger and sadness, and the worst for fear and disgust. Trained observers are reported to achieve an average facial expression recognition accuracy rate of 87 percent [7].

Access to annotated affective databases is a prerequisite for researchers to train and test the performance of their affect recognition algorithms. Collecting and annotating affective databases is a challenging task, especially if natural (spontaneous) multi-modal expressions are desired. Below we first briefly review the state of the art on affective databases available in the literature and then state the contribution and scope of this work.

1.1 Prior Work

Most of the affective databases that are accessible by researchers are acted and uni-modal [2], [8], [9], [10], [11]. Predominantly, acted datasets are recorded under very constrained conditions and resulting expressions are exaggerated. One of the most popular acted databases is the extended Cohn-Kanade database (CK+) [8], which encompasses 123 subjects with 327 labeled sequences pursuant to six basic emotions (anger, happiness, sadness, disgust, surprise and fear) [6] and contempt. The first frame of each sequence contains a neutral expression and the last frame of the sequence reflects the expression at its apex. Another image-based acted facial expression database is the Jaffe database [12], that contains 219 images of 10 Japanese females exhibiting the six basic emotions. The MMI database [13] consists of mostly posed videos of facial expressions showing complete temporal patterns (i.e., starting from onset, which is followed by apex and offset phases). Some of the videos contain spontaneous

- *S. Zhalehpour is with the INRS-EMT, Montreal, Quebec, Canada. E-mail: sara.zhalehpour@emt.inrs.ca.*
- *O. Onder is with Arcelik Inc., Istanbul, Turkey. E-mail: onurndr@yahoo.com.*
- *Z. Akhtar is with the Department of Mathematics and Computer Science, University of Udine, Via delle Scienze 206, 33100, Udine, Italy. E-mail: zahid.akhtar@uniud.it.*
- *C.E. Erdem is with the Department of Electrical and Electronics Engineering, Bahcesehir University, Ciragan Cad., Besiktas, Istanbul, Turkey. E-mail: cigdem.eroglu@eng.bahcesehir.edu.tr.*

Manuscript received 12 Feb. 2015; revised 17 Mar. 2016; accepted 22 Mar. 2016. Date of publication 11 Apr. 2016; date of current version 12 Sept. 2017. Recommended for acceptance by A. M. Martinez.
For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org, and reference the Digital Object Identifier below.
Digital Object Identifier no. 10.1109/TAFFC.2016.2553038

TABLE 1
Overview of Audio-Visual Affective Databases

Database	Posed versus Non-Posed	Language	Annotation	Number of subjects	Number of video clips
BAUM-1	Non-Posed	Turkish	SBE, 2 NBE, 3 MS	31	1,184
BAUM-2 [24]	Non-Posed	English, Turkish	SBE, 1 NBE	286	1,047
AFEW [22]	Non-Posed	English	SBE	330	1,426
Belfast [21]	Non-Posed	English	SBE	100	239
Semaine [17]	Non-Posed	English	3 BE, 10 NBE	150	959
IEMOCAP [20]	Posed	English	3 BE, 1 NBE	10	1,039
GEMEP [14]	Posed	None	SBE, 12 NBE	10	7,000
Fanelli [27]	Posed, 3D	English	SBE + MS	14	1,109
eNTERFACE [15]	Posed	English	SBE	44	1,166

The acronyms in the table are: SBE: Six Basic Emotions (Happiness, Sadness, Anger, Fear, Disgust, Surprise) and Neutral, NBE: Non-Basic Emotions, MS: Mental States.

laughter. Bosphorus database [10] is a 3D facial expression database incorporating various head poses and occlusions.

There are several audio-visual acted databases in the literature. The GEMEP dataset [9], [14] is comprised of videos of 10 actors coached by a professional director. During the recording session, the actors were asked specifically to utter combinations of meaningless phoneme sequences. It is worth mentioning that, out of 7,000 audio-visual affective video clips reflecting 18 emotions, only 289 samples representing five emotions (anger, fear, joy, relief, sadness) were selected for the GEMEP-FERA challenge [14]. Another audio-visual acted database is eNTERFACE [15], which contains video clips of 44 subjects uttering selected scripts in English while reflecting the six basic emotions.

Assembling datasets that encompass spontaneous or naturalistic expressions is a strenuous and tedious job [16], [17], [18], [19]. The SEMAINE [17] database was collected under constrained lab settings and contains naturalistic expressions from 150 subjects, who are in a conversation with a "sensitive artificial listener". Another induced emotional expressions database is the FEED database [18], which contains 18 subjects. The scripted affective dyadic conversations were induced by 10 professional actors in the IEMOCAP database [20], which resulted in more than 10 hours of audio-visual recordings. In particular, facial motions of the actors were captured using a system utilizing markers, while the utterances were labeled along the three dimensions indicating valence, activation, and dominance. Recently, DISFA [19] has been introduced, which is a spontaneous facial action intensity database. Different from most of the data sets, the Belfast naturalistic database [21] is comprised of clips accumulated manually from TV programs as well as clips recorded through dyadic discussions.

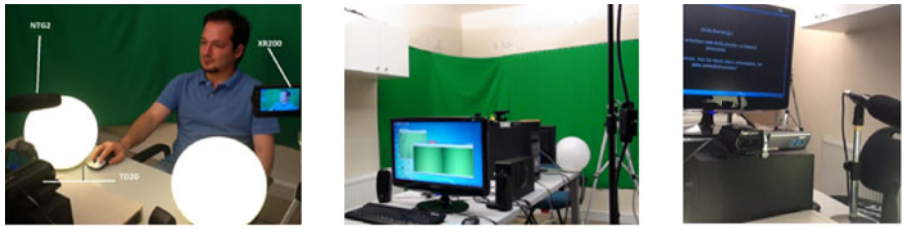
There are recent efforts towards collecting naturalistic databases from movies such as the AFEW [22] and BAUM-2 databases [23], [24]. Although using movie clips to gather affective databases do not replace pure spontaneity, they are more naturalistic and challenging as compared to acted databases. Several other approaches use crowdsourcing to collect facial expressions from the internet [25] in response to short video clips or to score the captured facial expressions [26]. However, they mainly concentrate on positive emotions and do not consider mental states. Also, very few efforts for naturalistic 3D audio-visual [27] or 3D facial expression databases [28] have been rendered using costly 3D scanners.

In Table 1, we give a summary of the audio-visual databases available in the literature. It can be observed that BAUM-1 is the only non-posed database in the literature that contains the six basic emotions as well as several non-basic emotions and mental states in a language different from English.

1.2 Contribution and Scope

As summarized in the previous section, there are a few audio-visual naturalistic affective databases in the literature, which are mostly in English and do not contain any expressions related to the mental states. In this work, we present a re-acted spontaneous audio-visual face database of affective and mental states in Turkish. The subjects watch a sequence of still images and short video clips, which are meticulously devised and timed to evoke a set of emotions and mental states. Emotion elicitation using video clips is a well-validated method [29], [30]. The subjects express their feelings and ideas about the stimuli they have watched on the screen in their own words, without using predetermined scripts. The database contains recordings reflecting the six basic emotions (*happiness*, *anger*, *sadness*, *disgust*, *fear*, *surprise*) as well as *boredom* and *contempt*. The database also contains several mental states, namely *unsure* (confused, undecided), *thinking*, *concentrating*, and *bothered*. The subjects are not guided in any way about how to perform the facial expressions. The database consists of simultaneous recordings of subjects using two cameras, both of which can record in high definition format. The first one is a stereo camera, which is placed in front of the subject at the top of the screen. The second one is a mono camera, which is placed to capture a half profile view of the subject. The categorical annotations of the recordings also include a score indicating the intensity of the emotion on a scale of 0 to 5.

The organization of the paper is as follows. In Section 2, we describe the data acquisition process in detail including emotion elicitation and recording. In Section 3, we give the details of the post-processing steps including segmentation, annotation and organization of the database. In Section 4, we outline the method used for multi-modal recognition of affective and mental states for the purpose of establishing baseline results on the database. In Section 5, experimental results on BAUM-1 database are provided and compared with the results on the eNTERFACE [15] database. Finally in Section 6, we provide conclusions and future directions for research.



(a) (b) (c)

Fig. 1. (a), (b) The setup of the studio showing the location of stereo and mono cameras, microphone and lights. (c) The subjects watch the stimuli shown on the monitor and express their ideas and feelings about the stimuli in their own words.

2 Method for Data Acquisition

The BAUM-1 database has been recorded in a studio, which was designed specifically as described below. In the studio, subjects first watch a stimuli video on a monitor in front of them and then express their feelings in their own words while they are being recorded with multiple cameras.

2.1 Recording Setup

In the designed studio a green curtain has been used as background (see Figs. 1a, 1b). A Sony HDR-TD20 Stereo HD camera is used for recording the frontal view of the face and a Sony HDR-XR200 Mono camera is used for recording from a half profile view with an angle of approximately 45 degrees. Illumination is provided by three 1,000 Watt tungsten (Red Head) lights directed towards the ceiling to provide a smooth lighting. Two spherical lights are located carefully on both sides of the face to minimize shadows on the face of the subject. For recording the audio, a Rode NTG 2 shotgun (directional) microphone has been used. A monitor is located in front of the subject at eye level to watch the stimuli video (see Fig. 1c). A clap-board has been used to assist in the synchronization of audio and video streams during post-processing.

2.2 Recording of Acted Clips

Before starting the spontaneous session, we asked the subjects to utter several sentences in Turkish with certain emotions and imagining specific scenarios, which is a similar procedure as in the eNTERFACE dataset [15]. We call the acted part of the database as BAUM-1a. In the acted recordings, we targeted 8 emotions and mental states, which are listed below together with the imagined scenarios:

- *Happiness*: You just learned that you have won the lottery and you are telling it to a friend of yours.
- *Sadness*: You have to explain your friend that his father has passed away.
- *Fear*: You are kidnapped and the kidnappers are holding a gun towards you. You have to beg for your life.
- *Anger*: You have just caught the thief who has stolen your wallet.
- *Disgust*: You have discovered an insect in your soup.
- *Confusion*: You didn't understand the lecture and asking the lecturer to explain again.
- *Boredom*: You have been waiting for a bus for at least an hour.
- *Interest (Curiosity)*: You want to learn about your friend's secret.

2.3 Elicitation of Emotional and Mental States for Spontaneous Recordings

In BAUM-1 database, our main focus was to obtain spontaneous audio-visual expressions of emotional and mental states. In order to bring a subject into the mood of the emotion, we first asked each subject to watch a carefully designed "stimuli video", which contains a sequence of images and video clips selected to evoke the target emotions and mental states. The target emotions and mental states are: happiness, anger, sadness, disgust, fear, surprise, boredom, contempt, unsure (confused etc.), thinking, concentrating, and bothered.

Emotion elicitation using films is a well-validated procedure in the literature [29], [31]. While watching a certain portion of the stimuli video (or shortly after that), the subjects are asked to explain their feelings and ideas about the video or the image just shown. Subjects are also recorded while they are watching the video to capture spontaneous facial expressions. After the recording process, the whole recording is segmented and annotated, which will be explained in more detail in Section 3.

The total duration of the stimuli video is approximately 35 minutes and there are 30 second intervals in between video clips or images for the subjects to express their own feelings and ideas in their own words. The total recording session lasts about 50 minutes for each subject. The audio-visual stimuli video consist of 29 images and videos that are expected to elicit the desired emotions and mental states. These 29 images and video clips (scenes) have been carefully selected from a larger set of candidates retrieved from the internet and the International Affective Picture System [32] as well as several works of Escher [33] (to elicit confusion). The selection from this larger set has been done by making a demonstration of the whole collection to a jury, which consisted of 19 students from the department of psychology, who were asked to give a score to each video on a scale of 0 to 5. The stimuli, which received an average score below 2 were eliminated and highest scoring stimuli were retained. The ordering for the elicitation of the emotions were designed in such a way that the most negative and intense video clips (such as an autopsy scene) were shown towards the end. There were also neutral clips in between emotion transitions so that the subjects had enough time to recover from one emotion and were ready to enter the mood of the next emotion. Images were also displayed long enough to enable elicitation of the target emotion or mental state. In Table 2, we summarize the contents of each video clip or image in the stimuli video together with the target emotion or mental state.

TABLE 2
Contents and Ordering of the Stimuli Video

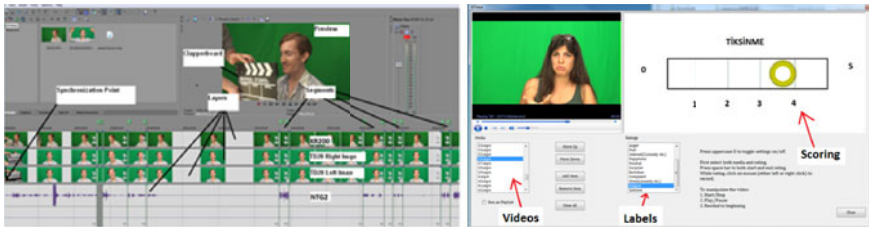
Video/picture number	Content	Target emotional or mental state
1	Horses illusion (image)	Unsure (confusion, undecided)
2	Stair paradox (image)	Unsure, Concentrating, Thinking
3	Wheel illusion (image)	Unsure, Concentrating, Thinking
4	Lines illusion (image)	Unsure, Concentrating, Thinking
5	Puppies (video)	Happiness/Amusement
6	Funny advertisement (video)	Happiness/Amusement
7	Funny stand-up show (video)	Happiness/Amusement
8	Space (image)	Neutral
9	Parking a car (video)	Contempt
10	Children (image)	Happiness
11	Crazy sportsmen (video)	Surprise
12	Bored man (image)	Boredom
13	A video of family fight	Anger/Sadness
14	Child and vulture (image)	Sadness
15	Sick newborn (image)	Sadness
16	Dead child and father (image)	Sadness/Anger
17	Murdering of a cat (video)	Anger/Sadness
18	Fisherman (image)	Neutral/Boredom
19	Thinking person (image)	Neutral/Boredom
20	Man pointing a gun (image)	Anger
21	Shark (image)	Fear
22	A vomiting man (image)	Disgust
23	Waterfalls (image)	Neutral
24	Car accident (video)	Sadness
25	Drunk driver (video)	Sadness/Anger
26	Badly injured hand (image)	Disgust
27	Clips from horror movies	Fear/Bothered
28	Autopsy video	Fear/Bothered
29	Illusionist cutting his wife	Surprise/Fear

Target emotions or mental states are also listed.

2.4 Participants and the Spontaneous Recording Process

The data was collected from 31 subjects, 17 of which are female, which are shown in Fig. 2. All subjects are native speakers of Turkish, and have an age range of 19-65. Prior to each recording session, we explained the whole procedure to the subjects in detail and warned them about possible disturbing scenes (e.g., autopsy scene), indicating that they can quit the session at any point. Each subject also signed a consent form, which states that the subject has understood and accepted the procedure and indicates whether all recordings of the subject can be used and shared for research purposes. All subjects except one of them (subject 5) gave permission for their images to be used in publications. The

Fig. 2. The 31 subjects who volunteered for the recordings of BAUM-1 database. All subjects (except subject 5) gave their consent for their images to be used in publications.



(a) (b)

Fig. 3. (a) Segmentation of the recording into clips using Sony Vegas software. (b) GTrace annotation tool is used for annotating the clips [35].

subjects completed the 50 minute long recording process by watching the stimuli video on their own and expressed their thoughts and feelings with their own words when prompted.

Several limitations of the used emotion elicitation procedure has been observed. First, the subjects may not express completely spontaneous emotions since they are aware of the cameras and they are being recorded. Hence, some subjects may tend to suppress their emotions, whereas some others may tend to exaggerate. We also had a difficulty in elicitation of fear since it is difficult to frighten the subjects in a secure office environment, which is inline with the observations stated in [29]. Another issue is the "discreteness" of the emotions [29], which means inducing a single emotion/mental state, with no traces of others. It was observed that this was especially difficult since several emotions or mental states coexist, which can be observed from the labels given by the annotators. This is explained in more detail in Sections 3.2 and 3.3 below.

3 Post Processing and Annotation

After the video of a subject is recorded, it is divided into smaller segments, which are then annotated. Below we give the details of the segmentation and annotation processes.

3.1 Segmentation and Organization of the Database

During post-processing, first the recorded stereo and mono video and audio streams are synchronized using the clapboard information in the audio and video streams. The precision of the synchronization is about 1 frame, which corresponds to 1/30 seconds. Then the recording is segmented into short video clips using Sony Vegas software (see Fig. 3) so that there is a single emotion or mental state expression in a clip. There might be a few clips which contains simultaneous expression of two emotions (e.g., happily surprised). The segmented clips are then rendered by fusing the audio and video channels and then saved by giving a name indicating the subject and clip number. Original stereo (frontal) and mono (half-profile) videos were recorded in high definition resolution ($1{,}080 \times 1{,}920$) and video clips were rendered in standard definition (576×720) resolution. A mono frontal view version of the database is also prepared, which consists of the right view of the stereo pair downsampled to a resolution of 480×854 [34]. An example recording can be seen in Fig. 4. We also provide subtitles in English using .srt files for each video clip.

3.2 Annotation

BAUM-1a database contains clips containing expressions of five basic emotions (happiness, sadness, anger, disgust, fear) along with expressions of boredom, confusion (unsure) and interest (curiosity). BAUM-1s database contains clips reflecting six basic emotions and also expressions of boredom, contempt, confusion, thinking, concentrating, bothered, and neutral. We used some of the mental state labels indicated by the taxonomy in [36] to annotate the mental states.

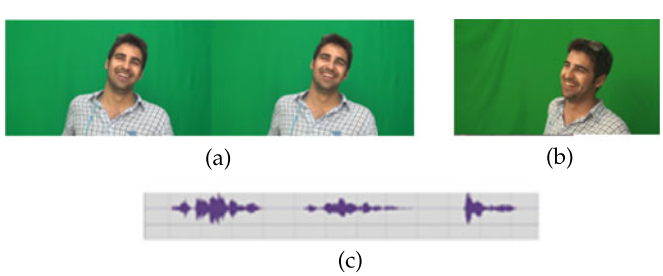
Each clip in the database is annotated by five annotators using the GTrace tool [35] (see Fig. 3b), which has a simple and friendly user interface. Prior to the annotations, the annotators were made familiar with the facial expressions by using the Mind Reading Software [37]. The annotator watches the clip as many times as he/she wants, selects the emotion or mental state that is dominant in the clip and finally gives a score between 0 and 5, which represents the intensity of the emotional or mental state expression in the clip. Finally, each clip is given a label by using majority voting over the five annotators. The scores of the selected label are averaged to determine the final score for the clip.

3.3 Inter-Annotator Agreement

We used the Kappa statistic [38], [39] to estimate the amount of inter-annotator agreement in the BAUM-1 database, which assesses differences between the agreement among annotators (i.e., "observed" agreement) and agreement that would occur by chance alone (i.e., "expected" agreement). The Kappa statistic ranges between -1 to 1, where 1 indicates perfect agreement, while 0 indicates agreement by chance. The negative values specify systematic disagreement between annotators. The expression for calculation of Kappa (κ) is as follows:

$$\kappa = \frac{P_o - P_c}{1 - P_c}, \tag{1}$$

where P_o denotes the relative observed agreement among raters and P_c denotes the hypothetical probability of agreement by chance. The Kappa value was calculated between each pair of annotators and then averaged. The



(a) (b) (c)

Fig. 4. An example recording from the BAUM-1 database. (a) A frontal stereo recording. (b) A mono recording from half profile. (c) The speech channel.

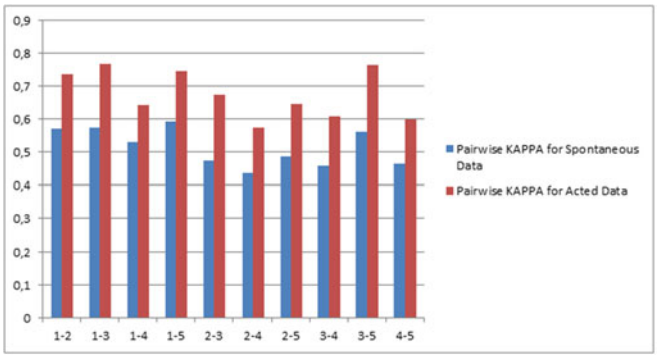


Fig. 5. The Kappa values for each pair of annotators for BAUM-1a database.

average Kappa statistic yielded a value of 0.67 for the BAUM-1a database, which can be considered as substantial agreement [39] between annotators. The highest agreement was between annotator 1 and annotator 3 (see Fig. 5), with a value of 0.76. The average Kappa value for BAUM-1s was 0.51, with a maximum pairwise agreement of 0.57, which can be interpreted as moderate agreement. We noticed that annotator 4 had the least agreement with the other annotators. If we exclude annotator 4 from the averaging process, the Kappa values become 0.72 and 0.54 for BAUM-1a and BAUM-1s, respectively. Moderate agreement is expected for BAUM-1s database since the emotional and mental state expressions are spontaneous and sometimes subtle, hence they are challenging to recognize even for humans. There are also multiple emotions in some sequences (e.g., angrily surprised, happily surprised) and the annotators might have a difficulty in choosing the dominant emotion.

3.4 Facial Feature Point Tracking

In many facial expression recognition algorithms, the first step is to track the location of salient points (i.e., landmarks) on the face. We used three different facial landmark tracking methods and compared them experimentally [40], [41], [42], [43]. In [40], a model based on mixtures of trees is used, which shows a good performance over a wide range of head poses from frontal to profile. CHEHRA tracker [41] is based on a cascade of linear regressors for incremental training of robust discriminative deformable models, which can automatically adapt to person-specific properties and imaging conditions. The IntraFace tracker [42] uses a supervised descent method for non-rigid image alignment to track profile-to-profile faces.

Several landmark tracking examples can be seen in Fig. 6a, which shows that the method in [40] works acceptably well for tracking a total of 68 facial landmarks on the face, which are shown with cyan star signs. The head pose angle in the yaw direction is also estimated in the range $[-90, 90]$ degrees with an increment of 15 degrees, which are shown as written in the forehead in Fig. 6a. The face tracker [40] sometimes fails to estimate the location of the facial landmarks and the head pose as can be seen in Fig. 6b. The facial landmarks are visually inspected, and acceptable tracking results are used in the facial expression recognition experiments. In such difficult cases, CHEHRA and IntraFace trackers have been observed to give better tracking results as shown in Fig. 6c, which track 49 facial points.

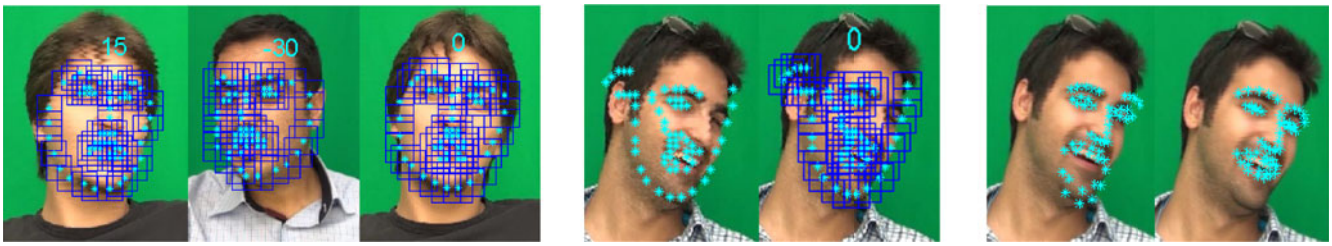


Fig. 6. (a) Facial landmarks (cyan stars) and the head pose angle (written in forehead) have been successfully tracked using the method in [40] for most clips in the database. (b) Facial feature tracking sometimes fails under sudden head movements [40]. (c) CHEHRA [41] (left) and IntraFace [42] (right) face trackers have been observed to give better results in difficult cases.

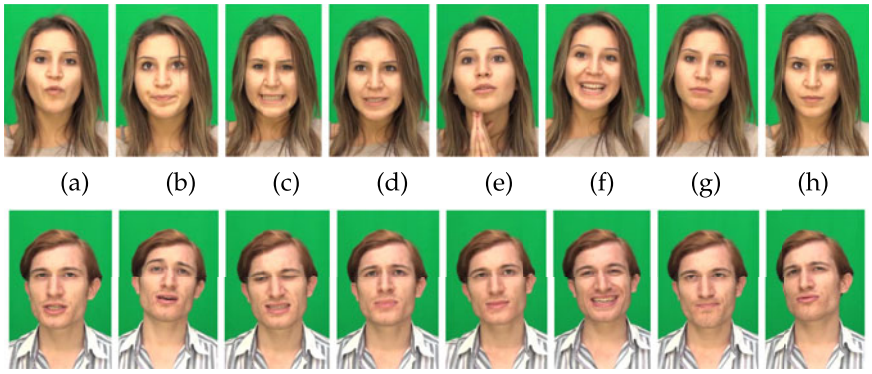


Fig. 7. Images selected from the video recordings in the BAUM-1a acted database are shown for subject 22 (top) and subject 11 (bottom). The figures in parenthesis indicate the sequence and frame numbers of (top / bottom) images. (a) Anger (s6-f75 / s7-f83) (b) Boredom (s9-f118 / s10-f12) (c) Disgust (s7-f10 / s8-f11) (d) Fear (s4-f108 / s6-f53) (e) Interest (s10-f68 / s11-f2) (f) Happiness (s1-f49 / s3-f74) (g) Sadness (s3-f46 / s4-f74) (h) Unsure (s8-f10 / s11-f11).

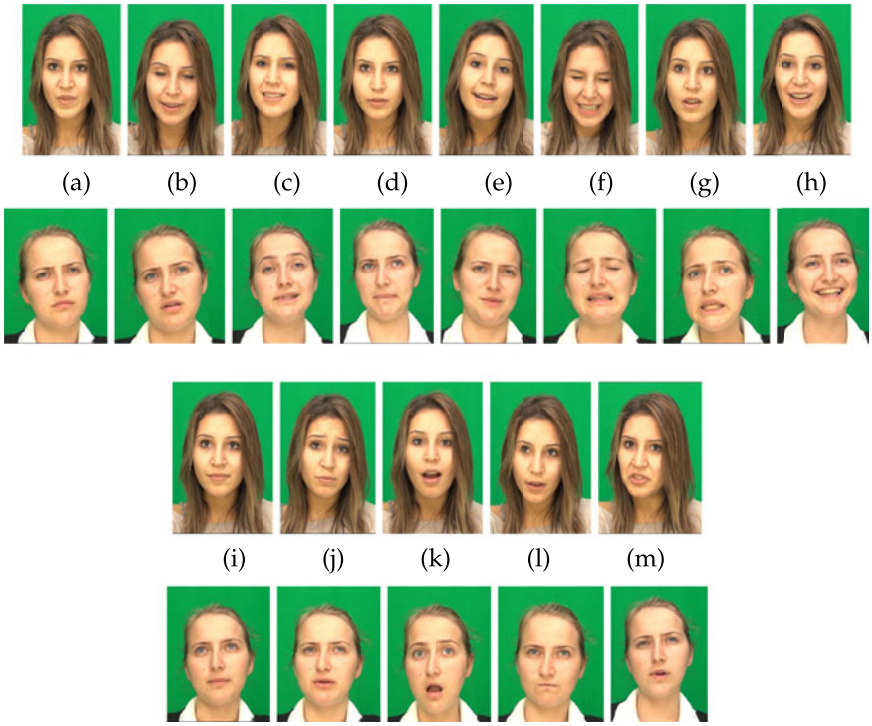


Fig. 8. Example frames from the clips in the BAUM-1s spontaneous database for subject 22 (top) and subject 21 (bottom). The figures in parenthesis indicate the sequence and frame numbers of (top / bottom) images. (a) Anger (s32-f3 / s50-f19) (b) Boredom (s67-f44 / s59-f6) (c) Bothered (s35-f91 / s22-f6) (d) Concentrating (s30-f82 / s21-f15) (e) Contempt (s56-f22 / s36-f130) (f) Disgust (s66-f92 / s78-f14) (g) Fear (s59-f18 / s63-f15) (h) Happiness (s22-f15 / s10-f19) (i) Neutral (s40-f46 / s48-f50) (j) Sadness (s55-f43 / s52-f51) (k) Surprise (s47-f7 / s62-f21) (l) Thinking (s31-f42 / s71-f3) (m) Unsure (s43-f102 / s11-f6).

3.5 Properties and Novelties of the BAUM1 Database

In Figs. 7 and 8 some example frames from the clips in BAUM-1a and BAUM-1s databases are shown, respectively. The properties of the databases are summarized in Table 3. The BAUM-1 database is a novel contribution to the affective computing area since it contains spontaneous recordings of the six basic emotions and several mental states in Turkish.

TABLE 3
Basic Properties of BAUM-1 Database

Property	BAUM-1a	BAUM-1s
	Acted	Spontaneous
Number of clips	273	1,184
Number of subjects		31 (13 female, 18 male)
Age Range of Subjects	19-65	19-65
Language of clips	Turkish	Turkish
Clip Length (min/max/average) in sec.	0.60/16.98/4.07	0.43/9.34/1.82
Video Format	AVI	MP4/AVI
Number of annotators	5	5
Kappa value	0.72	0.54
	Number of Clips Per Expression (min/max/average scores over 5)	
Happiness	27 (2.00/4.36/3.28)	179 (1.00/4.86/3.30)
Anger	43 (2.43/4.83/3.43)	59 (1.00/5.00/2.86)
Sadness	38 (1.97/3.10/3.08)	139 (1.00/4.52/2.83)
Disgust	35 (2.50/4.45/3.51)	86 (1.00/4.90/3.53)
Fear	36 (1.61/4.49/3.30)	38 (1.00/4.70/2.98)
Surprise	-	43 (1.00/4.50/3.38)
Boredom	27 (2.12/4.65/3.16)	22 (1.00/4.40/3.24)
Contempt	-	16 (1.50/4.30/3.20)
Unsure (inc. confused, undecided)	38 (1.78/4.20/2.97)	148 (1.00/5.00/3.02)
Interest (inc. curious)	29 (2.07/3.83/3.00)	-
Neutral	-	187
Thinking	-	112 (1.10/4.33/2.99)
Concentrating	-	64 (1.00/4.30/2.77)
Bothered	-	91 (1.00/4.40/3.05)

3.6 Availability

The database is publicly available for research purposes only and can be obtained officially upon request via the web site [34].

4 Multi-Modal Recognition of Affective and Mental States

In order to demonstrate the usefulness and the challenging nature of the collected BAUM-1 database, we conducted multi-modal affective and mental state recognition experiments on the acted BAUM-1a and spontaneous BAUM-1s databases. In these experiments, we employ a multi-modal affect recognition algorithm based on apex frame selection that was recently developed by the authors [44], [45]. Below, we briefly explain the visual and audio features used in the experiments, and the fusion process. We also briefly summarize the apex frame selection method presented in [44], [45] for the sake of completeness. The experimental results are given in Section 5.

4.1 Extraction of Visual Features from Video

There are many approaches in the literature for facial expression recognition (FER) from images [2], [46]. An affective video contains many frames, where emotions are expressed with varying intensities in each frame. Thus, one of the biggest challenges in emotion recognition from video is determining which frames to use and how to use them in order to attain the maximum possible recognition accuracy. A propitious technique is to use a single frame or a set of selected frames representing the emotional expression with high intensities (i.e., at the apex or peak phase). This approach assumes that there is only a single emotion expressed in the whole video clip. In this work, we use an approach based on apex frame selection [45], which is summarized below. The peak frame selection is preceded by a face detection and alignment process. Then, the procedure followed for computation of visual features is explained.

4.1.1 Face Detection and Alignment

Before selecting the peak frames and extracting the facial features from the video clip, the face is detected or tracked at each frame of the video and aligned so that global transformations of the head are minimized between frames. There are many approaches for face detection and tracking in the literature [40], [47], [48], [49]. In this work, the locations of eyes in all frames are detected utilizing three facial feature trackers, namely the algorithms by Zhu and Ramanan [40], CHEHRA [41] and IntraFace [42]. The face region is rescaled and cropped to a size of 168×126 so that the distance between the eye centers is 64 pixels. The face region is tessellated into sub-blocks of size $8 \times 6 = 48$ and the 18 sub-blocks that are irrelevant to the expression are discarded (e.g., around the hair and background) [24], [45]. In the remaining (relevant) blocks, which are numbered from 1 to 30, we extract the Local Phase Quantization features as described below.

4.1.2 Facial Features

We experimented with two different facial features, namely LPQ [50] and POEM [51] features, which are briefly described below.

Local Phase Quantization (LPQ) features were proposed for blur-insensitive image texture classification [50] and have also has been successfully employed for facial expression recognition [52]. Local Phase Quantization is similar to Local Binary Patterns (LBP) [53] in the sense that they both use local histograms to construct feature vectors. LBP features have has also been popular for facial expression recognition [54], [55]. Since LPQ has demonstrated better performance than LBP for facial expression recognition [22], [24], [50], [52], we adopted the former one to extract texture-based (i.e., appearance) features from facial images.

LPQ features use the phase information of the 2D short-term DFT in local neighborhoods. The resulting DFT is sampled at four frequencies. The samples are decorrelated, quantized and represented using integers between 0-255. Finally, 256-bin histograms in sub-blocks of the face region are concatenated and used as feature vectors to represent the face for classification. More details about extraction of LPQ features can be found in [50].

We exploited the LPQ code available from [56] in our experiments with default parameters. The resulting LPQ feature vector is of length 7,680, since there are 30 blocks, each of which is represented by a 256-bin LPQ histogram.

Patterns of Oriented Edge Magnitudes (POEM) features [51] of a pixel are calculated by replacing the intensity values in the calculation of the traditional LBP features by gradient magnitudes using the accumulated local histogram of gradient directions over a region around that pixel. Since POEM features are calculated at different scales using cells and blocks, it can capture both local and global information. It is also more robust to lighting variations as compared to LBP since gradient magnitudes are used instead of pixel intensities.

4.1.3 Peak Frame Selection

We used the maximum dissimilarity based peak frame selection (MAXDIST) method [45] in the experiments. It is assumed that the potential peak frames are "maximally dissimilar" to the rest of the frames in the video clip. Hence, first the dissimilarity between frames of a video clip are computed using appearance-based features of the face (e.g., the LPQ features). Then, these scores are arranged in a dissimilarity matrix so that the elements in the ith row of the matrix represent the distance scores between the LPQ histogram vectors of frame i and the rest of the frames in the video clip. The scores at each row of the dissimilarity matrix are then averaged and ordered. Finally, the peak frames are determined by selecting rows corresponding to the highest K scores, since those frames are estimated to be the most "dissimilar" frames in the video clip. The reader is referred to [45] for further details of the MAXDIST method.

During the experiments, choosing six peak frames for each video clip have been found to give good results. The final visual feature vector, x_1, of a video clip is calculated by averaging the LPQ feature vectors of the selected peak frames.

4.2 Extraction of Speech Features

The Mel-Frequency Cepstral Coefficients (MFCC) [57] and relative spectral features (RASTA) based on perceptual linear prediction (PLP) [58] were utilized to calculate the audio features for emotion recognition. As a pre-processing step,

TABLE 4
Single and Multi-Modal Affective and Mental State Recognition Accuracies on BAUM-1a Database Using LPQ Features and IntraFace Tracker [42]

	5 Basic Emotions	8 Emotions/ Mental States
Video-Based	47.44%	31.24%
Audio-Based	71.71%	63.53%
Audio-Visual (sum rule)	72.33%	61.09%
Audio-Visual (product rule)	71.56%	61.24%
Audio-Visual (weighted product rule)	75.32%	65.84%

silent intervals of audio including the leading and trailing edges were eliminated by soft-thresholding the energy over short windows. Then, the MFCC and RASTA-PLP features were extracted using 12 and 20 order filters, respectively, utilizing a window of length 25 msec and an overlap ratio of 50 percent. Finally, the 12 MFCC and 13 RASTA-PLP coefficients were merged with the first and second time derivatives and nine statistical functions (such as min, max etc.) were extracted [45]. The above feature extraction process produces an audio feature vector, x_2, of length $75 \times 9 = 675$, which is used for classification.

4.3 Classification and Fusion of Facial and Speech Features

The well-known Support Vector Machine (SVM) [59] classifier was employed for classification of audio and video features. The SVM classifier used for the video features utilizes a linear kernel to surpass the curse of dimensionality problem since the dimension of video features is high. An SVM classifier with a radial basis kernel was used for audio features using one-against-all approach. It is worth noting that audio features were normalized to the interval [0, 1] before classification.

A decision level fusion technique was applied to integrate the decision probabilities of each modality and emotion. After a thorough investigation of several probability fusion approaches [60], [61], it was inferred that the *weighted product rule* was the most successful [62] in our experiments. In particular, in the weighted product rule, the probabilities attained from each modality for a clip under test are multiplied and then the label of the maximum product is chosen [45].

We represent the feature vectors of the audio and visual modalities with x_2 and x_1, respectively. Let λ_2 and λ_1 denote the trained classifiers for the audio and visual modalities, where the probability estimated for the kth emotion/mental state in each modality is denoted by $P(\tilde{\omega}_k|x_i, \lambda_i), i = 1, 2$. These probabilities are merged as given below:

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

$$\omega^* = \max_k P(\omega_k|x_1, x_2), k = 1, 2, \ldots, 6, \quad (3)$$

where $\tilde{\omega}_k$ and ω_k represent the label for the kth emotion without and with fusion, respectively. The parameter ω^* represents the estimated emotion label of the video clip, and W_i denotes the weight used for each modality.

TABLE 5
Confusion Matrix: BAUM-1a Database, 5 Emotions, *Audio* Modality

	Anger	Disgust	Fear	Happiness	Sadness
Anger	**87.88**	6.26	1.82	0.0	4.04
Disgust	10.0	**73.33**	5.0	3.33	8.33
Fear	23.69	8.19	**46.62**	7.00	14.50
Happiness	10.00	6.19	12.38	**60.57**	10.86
Sadness	2.00	2.50	5.33	0.00	**90.17**

Figures represent percentages and the average recognition rate is 71.71 percent.

TABLE 6
Confusion Matrix: BAUM-1a Database, 5 Emotions, *Visual* Modality Using LPQ Features and IntraFace Tracker

	Anger	Disgust	Fear	Happiness	Sadness
Anger	**51.77**	9.87	12.93	17.37	8.06
Disgust	12.50	**59.33**	11.50	0.00	16.67
Fear	21.21	10.19	**32.90**	16.52	19.17
Happiness	13.52	9.52	48.10	**22.67**	6.19
Sadness	13.11	11.33	6.00	6.00	**63.56**

Figures represent percentages and the average recognition rate is 47.44 percent.

5 Baseline Audio-Visual Affective and Mental State Recognition Experiments

In this section, we present the single modality and multimodal affective and mental state recognition results on the collected BAUM-1a and BAUM-1s databases. We also carried out experiments on the eNTERFACE database, which is a well-known acted audio-visual database in the literature, which contains six basic emotions. In all the experiments given below, we employed 5-fold subject independent cross-validation to ensure subject independent results.

5.1 Results on BAUM-1a Database

There are 273 acted video clips in the BAUM-1a database representing eight emotional and mental states. The number of clips in each class are given in Table 3. We conducted two sets of experiments on the BAUM-1a database using LPQ features and IntraFace tracker [42]. In the first set, we used the five basic emotions, namely, anger, disgust, fear, happiness, and sadness. In the second set of experiments we also included boredom, interest and unsure in addition to the five basic emotions. As we can see in Table 4, the audio-based recognition accuracy (71.71 percent) is higher than the video-based accuracy (47.44 percent) for the five emotion case. The audio-visual accuracy is 75.32 percent using the weighted product rule with $W_1 = 1$ and $W_2 = 3$. The confusion matrices are given in Tables 5, 6, and 7 for the five emotion experiments. We can observe from the confusion matrices that anger (87.88 percent) and sadness (90.17 percent) have the highest two recognition rates from the audio channel (see Table 5), and disgust (59.33 percent) and sadness (63.56 percent) have the highest two recognition rates from the video channel (see Table 6). All the emotions except anger benefit from the fusion process and the average recognition rate after fusion rises to 75.32 percent (see Table 7).

TABLE 7
Confusion Matrix: BAUM-1a Database, 5 Emotions, *Audio-Visual* Using LPQ Features and IntraFace Tracker

	Anger	Disgust	Fear	Happiness	Sadness
Anger	**78.71**	12.93	4.04	2.5	1.82
Disgust	10.00	**81.67**	5.00	0.00	3.33
Fear	21.21	2.00	**63.29**	9.00	4.50
Happiness	6.67	9.52	12.86	**64.10**	6.86
Sadness	2.00	2.50	4.00	0.00	**91.50**

Fusion is done with weighted product rule ($W_2 = 3$). Figures represent percentages and the average recognition rate is 75.32 percent.

The gap between audio (63.53 percent) and video accuracies (31.24 percent) becomes larger when we include boredom, interest and unsure to the experiments. The confusion matrices for the 8 class experiments are given in Tables 8, 9, and 10. Similar to the 5 class case, anger (82.75 percent) and sadness (76.78 percent) have the highest two recognition rates from the audio channel (see Table 8). Disgust (74.13 percent) and anger (52.97 percent) have the highest two recognition rates from the video channel (see Table 9). Boredom cannot be recognized from the video channel at all. The average recognition rate rises to 65.84 percent after decision level fusion.

5.2 Results on BAUM-1s Database

We carried experiments on the BAUM-1s database, which has a total of 1,184 clips from 31 subjects. There are 13 emotional and mental states, which are Anger (An), Disgust (Di), Fear (Fe), Happiness (Ha), Sadness(Sa), Surprise (Su), Boredom (Bo), Contempt (Co), Unsure (Un), Neutral (Ne), Thinking (Th), Concentrating (Con), Bothered (Bot).

Below, we present the results of two different sets of experiments using LPQ features and IntraFace tracker [42]. In the first set, we used the six basic emotions (544 clips), and in the second set, we used all the 13 classes listed in the previous paragraph. The single and multi-modal affective and mental state recognition accuracies are given in Table 11. We can see that combining the audio and video based results at the decision level increases the recognition rate by 6 percent for the six emotion case and by 0.5 percent for the 13 emotion case.

TABLE 8
Confusion Matrix: BAUM-1a Database, 8 Emotions/Mental States, *Audio* Modality

	Anger	Boredom	Disgust	Fear	Happiness	Interest	Sadness	Unsure
Anger	**82.75**	4.72	8.48	0.00	0.00	0.00	4.04	0.00
Boredom	18.00	**47.50**	12.50	0.00	8.00	0.00	10.00	4.00
Disgust	10.00	2.50	**76.67**	2.50	5.83	0.00	2.50	0.00
Fear	17.69	4.00	11.19	**56.62**	2.00	0.00	4.50	4.00
Happiness	10.00	0.00	6.19	6.67	**66.95**	0.00	6.86	3.33
Interest	10.00	2.86	6.19	5.00	0.00	**59.29**	10.00	6.67
Sadness	2.00	0.00	11.17	2.00	0.00	3.33	**76.78**	4.72
Unsure	3.33	2.00	7.67	13.67	4.00	5.00	22.67	**41.67**

Figures represent percentages and the average recognition rate is 63.53 percent.

TABLE 9
Confusion Matrix: BAUM-1a Database, 8 Emotions/Mental States, *Visual* Modality

	Anger	Boredom	Disgust	Fear	Happiness	Interest	Sadness	Unsure
Anger	**52.97**	0.00	12.30	3.83	8.94	0.00	14.64	7.32
Boredom	21.48	**0.00**	2.63	16.34	26.35	4.22	13.70	15.28
Disgust	2.50	0.00	**74.13**	6.15	6.85	2.63	3.51	4.22
Fear	21.15	0.00	3.51	**24.24**	2.11	13.25	18.47	17.27
Happiness	23.11	0.00	10.04	0.00	**42.06**	7.03	7.73	10.04
Interest	11.66	0.00	24.59	34.63	6.52	**6.02**	0.00	16.56
Sadness	26.57	0.00	9.13	21.08	6.15	2.34	**25.06**	9.66
Unsure	28.33	0.00	7.73	18.27	7.73	4.22	13.70	**20.03**

Figures represent percentages and the average recognition rate is 31.24 percent.

TABLE 10
Confusion Matrix: BAUM-1a Database, 8 Emotions/Mental States, *Audio-Visual*

	Anger	Boredom	Disgust	Fear	Happiness	Interest	Sadness	Unsure
Anger	**84.29**	4.72	4.44	0.00	2.50	0.00	4.04	0.00
Boredom	14.00	**45.00**	10.00	2.50	8.00	0.00	10.00	10.50
Disgust	7.50	2.50	**76.67**	2.50	8.33	0.00	2.50	0.00
Fear	14.86	0.00	8.19	**61.95**	4.50	0.00	4.50	6.00
Happiness	6.67	0.00	3.33	6.67	**70.29**	3.33	9.71	0.00
Interest	10.00	0.00	5.71	2.86	0.00	**64.76**	13.33	3.33
Sadness	0.00	0.00	6.94	4.22	0.00	3.33	**78.56**	6.94
Unsure	3.33	0.00	5.67	13.67	4.00	7.00	16.67	**49.67**

Fusion is done with weighted product rule ($W_2 = 3$). Figures represent percentages and the average recognition rate is 65.84 percent.

TABLE 11
Single and Multi-Modal Affective and Mental State Recognition Accuracies on BAUM-1s Database Using LPQ Features and IntraFace Tracker [42]

	6 Basic Emotions	13 Emotions / Mental States
Video-Based	45.04%	25.17%
Audio-Based	29.41%	15.29%
Audio-Visual (sum rule)	45.78%	24.91%
Audio-Visual (product rule)	48.57%	25.19%
Audio-Visual (weighted product rule)	51.29%	25.76%

TABLE 12
Single and Multi-Modal Emotion Recognition Accuracies on eNTERFACE Database Using LOSO Cross-Validation (Based on LPQ Features and IntraFace Tracker [42])

	LOSO CV
Video-based Results	42.16%
Audio-based Results	72.95%
Audio-Visual Results (Sum Rule)	74.95%
Audio-Visual Results (Product Rule)	76.48%
Audio-Visual Results (Weighted Product Rule, $W_2 = 3$)	77.02%

TABLE 13
Confusion Matrix: eNTERFACE Database, *Audio* Modality (LOSO)

	Estimated Emotion					
	Anger	Disgust	Fear	Happiness	Sadness	Surprise
Anger	**88.4**	1.4	3.3	2.3	2.3	2.3
Disgust	5.6	**71.2**	7.0	4.2	6.0	6.0
Fear	7.4	9.8	**64.2**	4.7	7.4	6.5
Happiness	7.0	4.7	1.9	**75.3**	7.0	4.2
Sadness	3.3	6.5	5.1	6.0	**72.6**	6.5
Surprise	4.2	5.1	7.4	6.5	10.7	**66.0**

Figures represent percentages and the average recognition rate is 72.95 percent .

5.3 Results on eNTERFACE Database

We also carried out audio-visual emotion recognition experiments on the eNTERFACE'05 dataset [15], which contains clips of 44 subjects from 14 different nationalities. The subjects are asked to act the six basic emotions while uttering selected sentences in English with target emotions.

In the following, we report the emotion recognition results not only for each modality but also after decision level fusion. We use the LPQ features together with the IntraFace tracker [42] to extract the facial features. Since the samples are distributed almost uniformly over the emotions for each subject, we used a leave-one-subject-out cross validation method (LOSO). In Table 12, we present the experimental results of subject independent audio-visual emotion recognition rates. It can be observed that for visual and audio modalities, emotion recognition accuracies are 42.16 and 72.95 percent, respectively. We obtained an accuracy of 77.02 percent after decision level fusion. The reported results clearly indicate that audio based classification is better as compared to the visual-features based classification. The confusion matrices for the audio, visual and audio-visual experiments are given in Tables 13, 14, and 15, respectively. It is also worth noting that multi-modal fusion is beneficial for enhancing the recognition accuracy of all emotions except anger. Anger and happiness have highest emotion recognition accuracies, which are 83.62 and 86.13 percent, respectively, after fusion. Happiness and disgust are the emotions that experience the highest increases in their recognition rates after audio-visual fusion, since they show an increase of 11 and 4 percent in their accuracy, respectively.

TABLE 14
Confusion Matrix: eNTERFACE Database, *Video* Modality (LOSO)

	Estimated Emotion					
	Anger	Disgust	Fear	Happiness	Sadness	Surprise
Anger	**23.43**	20.13	8.00	14.41	11.56	22.48
Disgust	4.89	**59.49**	5.78	21.46	5.27	3.11
Fear	7.94	12.32	**23.05**	14.03	25.40	17.27
Happiness	4.95	14.13	5.52	**59.94**	4.38	11.08
Sadness	6.79	10.92	13.71	7.17	**44.89**	16.51
Surprise	8.25	4.13	10.16	21.33	13.71	**42.41**

Figures represent percentages and the average recognition rate is 42.16 percent.

TABLE 15
Audio-Visual Confusion Matrix for the eNTERFACE Database (LOSO)

	Estimated Emotion					
	Anger	Disgust	Fear	Happiness	Sadness	Surprise
Anger	**83.62**	3.56	4.89	3.37	1.02	3.56
Disgust	3.24	**74.92**	9.90	6.22	3.68	2.03
Fear	7.81	6.60	**65.65**	1.46	7.49	10.98
Happiness	1.33	5.94	0.44	**86.13**	3.33	2.83
Sadness	2.22	4.51	4.89	1.46	**79.56**	7.37
Surprise	3.11	0.89	9.27	7.68	6.67	**72.38**

Figures represent percentages and the average recognition rate is 77.02 percent.

5.4 Summary of the Experimental Results

In order to compare the emotion recognition accuracies on the acted eNTERFACE and BAUM-1 databases, we summarize the uni-modal and multi-modal accuracies in Table 16 using three different face trackers (Zhu, CHEHRA and IntraFace) and two different facial features (LPQ and POEM). We can see that eNTERFACE and BAUM-1a databases, which are both acted in different languages show similar characteristics, in the sense that the emotion recognition accuracy from audio is much higher as compared to the emotion recognition accuracy from video. However, the audio-based accuracies are much lower than the video-based accuracies for the BAUM-1s database both for the 6 class and 13 class cases. BAUM-1s accuracies for the 13 class case are much lower as compared to the 6 class case, which implies that recognition of mental states is quite challenging. If we compare the video-based results, we can observe that IntraFace tracker gives slightly higher accuracies in most of the cases except for BAUM-1a (5 emotions). If we compare the audio-visual accuracies in Table 16, we can

TABLE 16
Single and Multi-Modal Emotion Recognition Accuracies on eNTERFACE, BAUM-1a and BAUM-1s Databases Using Three Different Face Trackers (Zhu [40], CHEHRA [41] and IntraFace [42]) and Two Different Facial Features (LPQ [50], POEM [51]

	eNTERFACE (6 emotions)	BAUM-1a (5 emotions)	BAUM-1a (8 classes)	BAUM-1s (6 emotions)	BAUM-1s (13 classes)
Audio-only	72.95	71.71	63.53	29.41	15.29
Video-only (Zhu+LPQ)	38.22	46.60	26.30	43.75	24.07
Video-only (Zhu+POEM)	32.95	48.69	29.64	46.32	23.65
Video-only (CHEHRA+LPQ)	40.08	43.65	30.16	43.38	25.08
Video-only (CHEHRA+POEM)	33.26	45.13	31.53	44.30	25.00
Video-only (IntraFace+LPQ)	42.16	47.44	31.24	45.04	25.17
Video-only (IntraFace+POEM)	36.12	46.24	32.39	47.06	23.56
Audio-Visual (Zhu+LPQ)	76.79	74.42	65.06	50.00	26.18
Audio-Visual (Zhu+POEM)	75.45	76.38	63.45	50.37	25.51
Audio-Visual (CHEHRA+LPQ)	77.40	73.94	63.53	50.37	25.42
Audio-Visual (CHEHRA+POEM)	74.55	75.24	63.83	49.26	25.25
Audio-Visual (IntraFace+LPQ)	77.02	75.32	65.84	51.29	25.76
Audio-Visual (IntraFace+POEM)	75.78	76.54	64.13	50.18	25.68

Figures are given in percentages.

conclude that the results using three different face trackers and facial features are comparable on all databases with 1-2 percent differences in accuracies.

6 Conclusion

We presented a new spontaneous audio-visual Turkish database, BAUM-1 containing expressions of affective as well as mental states. The challenging nature of the database was demonstrated via baseline audio-visual emotion recognition experiments based on a peak frame selection approach. Although current algorithms can achieve reasonable recognition rates on acted expressions, which are somewhat exaggerated, the accuracies drop dramatically for naturalistic and subtle expressions.

The BAUM-1 database can be useful a resource to the affective computing community as it contains multi-modal, close-to-natural, affective and mental state expressions as opposed to the mostly single-modality and posed databases available in the literature. The database can be combined with other databases in other languages to investigate multi-language cross-corpora aspects of emotion recognition from speech. Investigating the correlation between speech-related and face-related features and emotion-independent face recognition could be other directions for future research.

Acknowledgments

The authors would like to thank Prof. Metehan Irak from the Department of Psychology, Bahcesehir University, for the useful discussions during the design of the stimuli video. The author would also like to acknowledge the support received from the Turkish Scientific and Technical Research Council (TÜBİTAK) under project 110E056.

References

[1] N. Sebe, I. Cohen, and T. S. Huang, "Multimodal approaches for emotion recognition: A survey," in *Internet Imaging VI*. 2005, vol. 5670, no. 5670, pp. 56–67.

[2] Z. H. Zeng, M. Pantic, G. I. Roisman, and T. S. Huang, "A survey of affect recognition methods: Audio, visual, and spontaneous expressions," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 1, pp. 39–58, Jan. 2009.

[3] A. Ryan, J. Cohn, S. Lucey, J. Saragih, P. Lucey, F. D. la Torre, and A. Rossi, "Automated facial expression recognition system," in *Proc. Int. Carnahan Conf. Security Technol.*, 2009, pp. 172–177.

[4] G. C. Littlewort, M. S. Bartlett, and K. Lee, "Automatic coding of facial expresssions displayed during posed and genuine pain," *Image Vision Comput.*, vol. 27, no. 12, pp. 1797–1803, 2009.

[5] A. B. Ashraf, S. Lucey, J. F. Cohn, T. Chen, Z. Ambadar, K. M. Prkachin, and P. E. Solomon, "The painful face — Pain expression recognition using active appearance models," *Image Vision Comput.*, vol. 27, no. 12, pp. 1788–1796, 2009.

[6] P. Ekman and W. V. Friesen, *Pictures of Facial Effect*. Palo Alto, CA, USA: Consulting Psychologists Press, 1976.

[7] J. Bassili, "Emotion recognition: The role of facial movement and the relative importance of upper and lower areas of the face," *J. Pers. Soc. Psychol.*, vol. 37, pp. 2049–2058, 1979.

[8] P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, Z. Ambadar, and I. Matthews, "The extended Cohn-Kanade dataset (CK+): A complete dataset for action unit and emotion-specified expression," presented at the IEEE Workshop on CVPR for Human Communicative Behavior Analysis, San Francisco, CA, USA, 2010.

[9] T. Banziger and K. R. Scherer, "Introducing the Geneva multimodal emotionportrayal (GEMEP) corpus," in *Blueprint for Affective Comput.: A Sourcebook*. London, U.K.: Oxford Univ. Press, 2010, pp. 271–294.

[10] A. Savran, H. D. N. Alyuz, O. Celiktutan, B. Gkberk, B. Sankur, and L. Akarun, "Bosphorus database for 3D face analysis," in *Proc. 1st COST Workshop Biometrics Identity Manag.*, 2008, pp. 47–56.

[11] E. Sariyanidi, H. Gunes, and A. Cavallaro, "Automatic analysis of facial affect: A survey of registration, representation and recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 6, pp. 1113–1133, Jun. 2015.

[12] M. J. Lyons, S. Akamatsu, M. Kamachi, and J. Gyoba, "Coding facial expressions with Gabor wavelets," in *Proc. 3rd IEEE Int. Conf. Autom. Face Gesture Recog.*, Japan, 1998, pp. 200–205.

[13] M. Pantic, M. F. Valstar, R. Rademaker, and L. Maat, "Web-based database for facial expression analysis," in *Proc. IEEE Int. Conf. Multimedia Expo*, 2005, [Online]. Available: http://www.mmifacedb.com/

[14] M. F. Valstar, B. Jiang, M. Mehu, M. Pantic, and K. R. Scherer, "The first facial expression recognition and analysis challenge," in *Proc. IEEE Int. Conf. Face Gesture Recog.*, 2011, pp. 921–926.

[15] O. Martin, I. Kotsia, B. Macq, and I. Pitas, "The eNTERFACE05 audio-visual emotion database," in *Proc. 1st IEEE Workshop Multimedia Database Manag.*, 2006, p. 8.

[16] M. Grimm, K. Kroschel, and S. Narayanan, "The Vera am Mittag German audio-visual emotional speech database," in *Proc. Int. Conf. Multimedia Expo*, 2008, pp. 865–868.

[17] G. Mckeown, M. F. Valstar, R. Cowie, M. Pantic, and M. Schroeder, "The SEMAINE database: Annotated multimodal records of emotionally coloured conversations between a person and a limited agent," *IEEE Trans. Affective Comput.*, vol. 3, no. 1, pp. 5–17, Jan.-Mar. 2012.

[18] F. Wallhoff, "Facial expressions and emotion database," 2006. [online]. Available: http://www.mmk.ei.tum.de/~waf/fgnet/feedtum.html

[19] S. M. Mavadati, M. H. Mahoor, K. Bartlett, P. Trinh, and J. F. Cohn, "DISFA: A spontaneous facial action intensity database," *IEEE Trans. Affective Comput.*, vol. 4, no. 2, pp. 151–160, Apr.-Jun. 2013.

[20] C. Busso, M. Bulut, C. C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, and S. S. Narayanan, "IEMOCAP: Interactive emotional dyadic motion capture database," *J. Language Resources Eval.*, vol. 42, no. 4, pp. 335–359, 2008.

[21] E. Douglas-Cowie, R. Cowie, and M. Schoder, "A new emotion database: Considerations, sources and scope," in *Proc. ISCA ITRW Speech Emotion*, 2000, pp. 39–44.

[22] A. Dhall, R. Goecke, S. Lucey, and T. Gedeon, "Collecting large, richly annotated facial-expression databases from movies," *IEEE Multimedia*, vol. 19, no. 3, pp. 34–41, Jul.-Sep. 2012.

[23] C. Turan, C. Kansin, S. Zhalehpour, Z. Aydin, and C. E. Erdem, "A method for extraction of audio-visual facial clips from movies," in *Proc. IEEE Signal Process. Appl. Conf.*, 2013, pp. 1–4.

[24] C. E. Erdem, C. Turan, and Z. Aydin, "BAUM-2: A multilingual audio-visual affective face database," *Multimedia Tools Appl.*, vol. 74, pp. 7429–7459, 2014.

[25] D. McDuff, R. E. Kaliouby, and R. W. Picard, "Crowdsourcing facial responses to online videos," *IEEE Trans. Affective Comput.*, vol. 3, no. 4, pp. 456–468, Oct.-Dec. 2012.

[26] J.-Y. Zhu, A. Agarwala, A. A. Efros, E. Shechtman, and J. Wang, "Mirror mirror: Crowdsourcing better portraits," *ACM Trans. Graph.*, vol. 33, no. 6, 2014, Art. no. 234.

[27] G. Fanelli, J. Gall, H. Romsdorfer, T. Weise, and L. V. Gool, "A 3-D audio-visual corpus of affective communication," *IEEE Trans. Multimedia*, vol. 12, no. 6, pp. 591–598, Oct. 2010.

[28] X. Zhang, L. Yin, J. Cohn, S. Canavan, M. Reale, A. Horowitz, and P. Liu, "A highresolution spontaneous 3D dynamic facial expression database," presented at the Int. Conf. Automatic Face and Gesture Recognition, Shanghai, China, Apr. 2013.

[29] J. J. Gross and R. W. Levenson, "Emotion elicitation using films," *Cognition Emotion*, vol. 9, no. 1, pp. 87–108, 1995.

[30] R. Westermann, K. Spies, G. Stahl, and F. W. Hesse, "Relative effectiveness and validity of mood induction procedures: A meta-analysis," *Eur. J. Social Psychology*, vol. 26, no. 4, pp. 557–580, 1996.

[31] X. Zhang, L. Yin, J. F. Cohn, S. J. Canavan, M. Reale, A. Horowitz, and P. Liu, "A high-resolution spontaneous 3D dynamic facial expression database," in *Proc. IEEE Int. Conf. Autom. Face Gesture Recog.*, 2013, pp. 1–6.

[32] P. J. Lang, M. M. Bradley, and B. N. Cuthbert, "International affectivepicture system (IAPS): Technical manual and affective ratings," NIMH Center for the Study of Emotion and Attention, Gainesville, FL, USA, Tech. Rep. no. A-4, 1997.

[33] M. C. Escher. (2016, Apr.) [Online]. Available: http://www.mcescher.com/

[34] O. Onder, S. Zhalehpour, and C. E. Erdem, "Bahcesehir University multimodal face database of spontaneous affective and mental states (BAUM-1)," 2014. [Online]. Available: http://baum1.bahcesehir.edu.tr/

[35] R. Cowie, C. Cox, J.-C. Martin, A. Batliner, D. Heylen, and K. Karpouzis, "Issues in data labelling," in *Emotion-Oriented Systems: The Humaine Handbook*. Berlin, Germany: Springer-Verlag, 2011, pp. 215–244.

[36] R. A. Kaliouby, "Mind-reading machines: Automated inference of complex mental states," Univ. Cambridge, Cambridge, U.K., Tech, Rep. ISSN 1476-2986 UCAM-CL-TR-636, 2005.

[37] S. Baron-Cohen, "Mind Reading: The Interactive Guide to Emotions," London, Philadelphia, U.K.: Jessica Kingsley Publishers, 2004.

[38] M. W. Watkins and M. Pacheco, "Interobserver agreement in behavioral research," *J. Behavioral Edu.*, vol. 10, no. 4, pp. 205–212, 2000.

[39] A. J. Viera and J. M. Garrett, "Understanding interobserver agreement: The Kappa statistic," *Family Med.*, vol. 37, no. 5, 2005.

[40] X. Zhu and D. Ramanan, "Face detection, pose estimation and landmark localization in the wild," in *Comput. Vis. Pattern Recog.*, 2012, pp. 2879–2886.

[41] A. Asthana, S. Zafeiriou, S. Cheng, and M. Pantic, "Incremental face alignment in the wild," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2014, pp. 1859–1866.

[42] X. Xuehan and F. D. la Torre, "Supervised descent method and its applications to face alignment," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2013, pp. 532–539.

[43] X. Xuehan and F. D. la Torre, "Global supervised descent method," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2015, pp. 2664–2673.

[44] S. Zhalehpour, Z. Akhtar, and C. E. Erdem, "Multimodal emotion recognition with automatic peak frame selection," in *Proc. IEEE Int. Symp. Innovations Intell. Syst. Appl.*, 2014, pp. 116–121.

[45] S. Zhalehpour, Z. Akhtar, and C. E. Erdem, "Multimodal emotion recognition based on peak frame selection from video," *Signal, Image Video Process.*, 2015, DOI: 10.1007/s11760-015-0822-0

[46] S. Ulukaya and C. E. Erdem, "Gaussian mixture model based estimation of the neutral face shape for emotion recognition," *Digit. Signal Process.*, vol. 32, pp. 11–23, 2014.

[47] P. Viola and M. J. Jones, "Robust real-time face detection," *Int. J. Comput. Vision*, vol. 57, no. 2, pp. 137–154, 2004.

[48] C. E. Erdem, S. Ulukaya, A. Karaali, and A. T. Erdem, "Combining haar feature and skin color based classifiers for face detection," in *Proc. IEEE 36th Int. Conf. Acoust., Speech Signal Process.*, 2011, pp. 1497–1500.

[49] J. M. Saragih, S. Lucey, and J. F. Cohn, "Deformable model fitting by regularized landmark mean-shift," *Int. J. Comput. Vis.*, vol. 91, pp. 200–215, 2011.

[50] V. Ojansivu and J. Heikkil, "Blur insensitive texture classification using local phase quantization," *Lecture Notes Comput. Sci.*, vol. 5099, pp. 236–243, 2008.

[51] N. Vu and A. Caplier, "Face recognition with patterns of oriented edge magnitudes," in *Proc. Eur. Conf. Comput. Vis.*, 2010, vol. 6311, pp. 313–326.

[52] A. Dhall, R. Goecke, and T. Gedeon, "Emotion recognition using PHOG and LPQ features," in *Proc. Workshop Facial Expression Recog. Anal. Challenge, IEEE Autom. Face Gesture Recog. Conf.*, 2011, pp. 878–883.

[53] T. Ojala, M. Pietikainen, and T. Maenpaa, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," *IEEE Trans. Pattern Anlaysis Mach. Intell.*, vol. 24, no. 7, pp. 971–987, Jul. 2002.

[54] C. Shan, S. Gong, and P. W. McOwan, "Facial expression recognition based on local binary patterns: A comprehensive study," *Image Vision Comput.*, vol. 27, pp. 803–816, 2009.

[55] G. Zhao and M. Pietikainen, "Dynamic texture recognition using local binary patterns with an application to facial expressions," *IEEE Trans. Pattern Anlaysis Mach. Intell.*, vol. 29, no. 6, pp. 915–928, Jun. 2007.

[56] "Machine vision group, matlab codes for local phase quantization," 2013. [Online]. Available: http://www.cse.oulu.fi/CMV/Downloads/LPQMatlab

[57] X. Huang, A. Acero, and H. Hon, *Spoken Language Processing: A Guide to Theory, Algorithm, and System Development*. Englewood Cliffs, NJ, USA: Prentice Hall, 2001.

[58] H. Hermansky and N. Morgan, "RASTA processing of speech," *IEEE Trans. Speech Audio Process.*, vol. 2, no. 4, pp. 578–589, Oct. 1994.

[59] C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," *ACM Trans. Intell. Syst. Technol.*, vol. 2, pp. 27:1–27:27, 2011.

[60] P. K. Atrey, M. A. Hossain, A. E. Saddik, and M. S. Kankanhalli, "Multimodal fusion for multimedia analysis: A survey," *Multimedia Syst.*, vol. 16, pp. 345–379, 2010.

[61] J. Kittler, M. H. R. P. Duin, and J. Matas, "On combining classifiers," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 20, no. 3, p. 226–239, Mar. 1998.

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

Sara Zhalehpour received the BSc and MSc degrees in telecommunications engineering from the University of Tabriz, Iran, in 2009 and 2012, respectively. She received her second MSc degree in electrical and electronics engineering from Bahcesehir University, Turkey, in 2014. Since 2014, she has been working toward the PhD degree at INRS-EMT, Montreal, Canada. Her main research areas of interest are audio-visual emotion recognition and human-computer interaction.

Onur Onder received the BSc degree in electrical and electronics engineering from Ege University, Izmir, in 2010. He received the MSc degree in electrical and electronics engineering from Bahcesehir University, Istanbul, in 2014. He is currently working toward the PhD degree at Dokuz Eylul University, Izmir. His research interests include digital image and video processing, human-computer interaction, machine learning, and remote sensing.

Zahid Akhtar received the PhD degree in electronic and computer engineering from the University of Cagliari, Italy, in 2012. He is currently working as a research associate in the Department of Mathematics and Computer Science, University of Udine, Italy. His research interests include computer vision, pattern recognition and image processing with applications to biometrics, affective computing and security systems.

Cigdem Eroglu Erdem received the BS and MS degrees in electrical and electronics engineering from Bilkent University, Ankara, Turkey, in 1995 and 1997, respectively. She received the PhD degree in electrical and electronics engineering from Bogazici 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. Between 2003-2004, she was a postdoctoral fellow at the Faculty of Electrical Engineering, Delft University of Technology, The Netherlands, where she was also affiliated with the video processing group at Philips Research Laboratories, Eindhoven. She is currently a professor in the Department of Electrical and Electronics Engineering, Bahcesehir University, Istanbul, Turkey. Her research interests are in the areas of digital image and video processing, including affective computing, vision and scene understanding and human computer. She is a member of the IEEE.