A Multimodal Database for Affect Recognition and Implicit Tagging

Mohammad Soleymani, *Member*, *IEEE*, Jeroen Lichtenauer, Thierry Pun, *Member*, *IEEE*, and Maja Pantic, *Fellow*, *IEEE*

Abstract—MAHNOB-HCI is a multimodal database recorded in response to affective stimuli with the goal of emotion recognition and implicit tagging research. A multimodal setup was arranged for synchronized recording of face videos, audio signals, eye gaze data, and peripheral/central nervous system physiological signals. Twenty-seven participants from both genders and different cultural backgrounds participated in two experiments. In the first experiment, they watched 20 emotional videos and self-reported their felt emotions using arousal, valence, dominance, and predictability as well as emotional keywords. In the second experiment, short videos and images were shown once without any tag and then with correct or incorrect tags. Agreement or disagreement with the displayed tags was assessed by the participants. The recorded videos and bodily responses were segmented and stored in a database. The database is made available to the academic community via a web-based system. The collected data were analyzed and single modality and modality fusion results for both emotion recognition and implicit tagging experiments are reported. These results show the potential uses of the recorded modalities and the significance of the emotion elicitation protocol.

Index Terms—Emotion recognition, EEG, physiological signals, facial expressions, eye gaze, implicit tagging, pattern classification, affective computing.

1 Introduction

Although the human emotional experience plays a central part in our lives, our scientific knowledge about human emotions is still very limited. Progress in the field of affective sciences is crucial for the development of psychology as a scientific discipline or application that has anything to do with humans as emotional beings. More specifically, the application of human-computer interaction relies on knowledge about the human emotional experience, as well as on knowledge about the relation between emotional experience and affective expression.

An area of commerce that could obviously benefit from an automatic understanding of human emotional experience is the multimedia sector. Media items such as movies and songs are often primarily valued for the way in which they stimulate a certain emotional experience. While it might often be the affective experience that a person is looking for, media items are currently primarily tagged by their genre, subject or their factual content. Implicit affective tagging through automatic understanding of an individual's response to media items would make it possible to rapidly tag large quantities of media, on a detailed level and in a way that would be more meaningful to understand how people experience the affective aspects of media content [1]. This allows more effective content retrieval, required to manage the ever-increasing quantity of shared media.

To study human emotional experience and expression in more detail and on a scientific level, and to develop and benchmark methods for automatic recognition, researchers are in need of rich sets of data of repeatable experiments [2]. Such corpora should include high-quality measurements of important cues that relate to the human emotional experience and expression. The richness of the human emotional expressiveness poses both a technological as well as a research challenge. This is recognized and represented by an increasing interest into pattern recognition methods for human behavior analysis that can deal with the fusion of measurements from different sensor modalities [2]. However, obtaining multimodal sensor data is a challenge in itself. Different modalities of measurement require different equipment, developed and manufactured by different companies, and different expertise to set up and operate. The need for interdisciplinary knowledge as well as technological solutions to combine measurement data from a diversity of sensor equipment is probably the main reason for the current lack of multimodal databases of recordings dedicated to human emotional experiences.

To contribute to this need for emotional databases and affective tagging, we have recorded a database of multimodal recordings of participants in their response to affectively stimulating excerpts from movies and images and videos with correct or incorrect tags associated with human actions. The database is freely available to the

- *M. Soleymani and T. Pun are with the Computer Vision and Multimedia Laboratory, Computer Science Department, University of Geneva, Battelle Campus, Building A, Rte. de Drize 7, Carouge (GE) CH-1227, Switzerland. E-mail: {mohammad.soleymani, thierry.pun}@unige.ch.*
- *J. Lichtenauer is with the Department of Computing, Imperial College London, 180 Queen's Gate, London SW7 2AZ, United Kingdom. E-mail: j.lichtenauer@imperial.ac.uk.*
- *M. Pantic is with the Department of Computing, Imperial College London, 180 Queen's Gate, London SW7 2AZ, United Kingdom, and the Faculty of Electrical Engineering, Mathematics and Computer Science (EEMCS), University of Twente, Drienerlolaan 5, Enschede 7522 NB, The Netherlands. E-mail: m.pantic@imperial.ac.uk.*

Manuscript received 12 Nov. 2010; revised 1 July 2011; accepted 6 July 2011; published online 28 July 2011.
Recommended for acceptance by B. Schuller, E. Douglas-Cowie, and A. Batliner. For information on obtaining reprints of this article, please send e-mail to: taffc@computer.org, and reference IEEECS Log Number TAFFCSI-2010-11-0112.
Digital Object Identifier no. 10.1109/T-AFFC.2011.25.

TABLE 1
MAHNOB-HCI Database Content Summary

Participants and modalities	
Nr. of participants	27, 11 male and 16 female
Recorded signals	32-channel EEG (256Hz), Peripheral physiological signals (256Hz), Face and body video using 6 cameras (60f/s), Eye gaze (60Hz) and Audio (44.1kHz)
Emotional responses to videos (Experiment 1)	
Nr. of videos	20
Selection method	Subset of online annotated videos (see section 3.1)
Self-report	Emotional keyword, arousal, valence, dominance, predictability
Rating values	discrete scale of 1 - 9
Implicit tagging (Experiment 2)	
Nr. of videos and images	28 images and 14 videos
Dataset	Pictures from flickr*, and short videos showing human actions [3]
Self-report	Agreement with the displayed tag

* *http://www.flickr.com*

academic community, and is easily accessible through a web-interface.[1] The recordings for all excerpts are annotated through an affective feedback form, filled in by the participants immediately after each excerpt. A summary of the MAHNOB-HCI database characteristics is given in Table 1. The recordings of this database are precisely synchronized and its multimodality permits researchers to study the simultaneous emotional responses using different channels. Two typical sets including responses to emotional videos and implicit tagging or agreement with displayed tags can be used for both emotion recognition as well as multimedia tagging studies. Emotion recognition and implicit tagging baseline results are given for researchers who are going to use the database. The baseline results set a target for the researchers to reach.

In Section 2, we give an overview of existing affective databases, followed by descriptions of the modalities we have recorded in our database in Section 3. Section 4 explains the experimental setup. The first experiment paradigm, some statistics and results of classifications of emotions are presented in Section 5 and for the second experiment in Section 6. A discussion on the use of the database and recommendations for recordings of such databases are given in Section 7, followed by our conclusions in Section 8.

2 Background

Creating affective databases is an important step in emotion recognition studies. Recent advances in emotion recognition have motivated the creation of novel databases containing emotional expressions. These databases mostly include speech, visual, or audio-visual data [5], [6], [7], [8]. The visual modality of the emotional databases includes face and/or body gestures. The audio modality carries acted or genuine emotional speech in different languages. In the last decade, most of the databases consisted only of acted or deliberately expressed emotions. More recently, researchers have begun sharing spontaneous and natural emotional databases such as [6], [7], [9]. We only review the publicly available spontaneous or naturalistic databases and refer the reader to the following review [2] for posed, audio, and audio-visual databases.

Pantic et al. created the MMI web-based emotional database of posed and spontaneous facial expressions with both static images and videos [5], [10]. The MMI database consists of images and videos captured from both frontal and profile view. The MMI database includes data from 61 adults acting different basic emotions and 25 adults reacting to emotional videos. This web-based database gives an option of searching in the corpus and is downloadable.[2]

One notable database with spontaneous reactions is the Belfast database (BE) created by Cowie et al. [11]. The BE database includes spontaneous reactions in TV talk shows. Although the database is very rich in body gestures and facial expressions, the variety in the background makes the data a challenging data set of automated emotion recognition. The BE database was later included in a much larger ensemble of databases in the HUMAINE database [6]. The HUMAINE database consists of three naturalistic and six induced reaction databases. Databases vary in size from 8 to 125 participants and in modalities, from only audio-visual to peripheral physiological signals. These databases were developed independently at different sites and collected under the HUMAINE project.

The "Vera am Mittag" (VAM) audio-visual database [7] is another example of using spontaneous naturalistic reactions during a talk show to develop a database. Twelve hours of audio-visual recordings from a German talk show, "Vera am Mittag," were segmented and annotated. The segments were annotated using valence, activation, and dominance. The audio-visual signals consist of the video and utterances from 104 different speakers.

Compared to audio-visual databases, there are fewer publicly available affective physiological databases. Healey and Picard recorded one of the first affective physiological data sets at MIT, which has reactions of 17 drivers under different levels of stress [4]. Their recordings include electrocardiogram (ECG), galvanic skin response (GSR) recorded from hands and feet, electromyogram (EMG) from the right trapezius, as well as the respiration pattern. The database of stress recognition in drivers is publicly available from Physionet.[3]

The Database for Emotion Analysis using Physiological Signals (DEAP) [9] is a recent database that includes peripheral and central nervous system physiological signals in addition to face videos from 32 participants. The face videos were only recorded from 22 participants. EEG signals were recorded from 32 active electrodes. Peripheral nervous system physiological signals were EMG, electrooculogram (EOG), blood volume pulse (BVP) using plethysmograph, skin temperature, and GSR. The spontaneous

1. http://mahnob-db.eu.

2. http://www.mmifacedb.com/.

3. http://www.physionet.org/pn3/drivedb/.

TABLE 2
The Summary of the Characteristics of the Emotional Databases Reviewed

Database	Nr. Part.	Posed or Spon.	Induced or Natural	Audio	Visual	Peripheral physio.	EEG	Eye gaze
MIT [4]	17	Spon.	Natural	No	No	Yes	No	No
MMI [5]	61,29	Posed & spon.	Induced	No	Yes	No	No	No
HUMAINE [6]	Multiple	Spon.	Both	Yes	Yes	Yes	No	No
VAM [7]	19	Spon.	Natural	Yes	Yes	No	No	No
SEMAINE [8]	20	Spon.	Induced	Yes	Yes	No	No	No
DEAP [9]	32	Spon.	Induced	No	Yes (for 22)	Yes	Yes	No
MAHNOB-HCI	27	Spon.	Induced	Yes	Yes	Yes	Yes	Yes

The last row is our database.

reactions of participants were recorded in response to music video clips. This database is publicly available on the Internet.[4] The characteristics of the reviewed databases are summarized in Table 2.

3 Modalities and Apparatus

3.1 Stimuli and Video Selection

Although the most straightforward way to represent an emotion is to use discrete labels such as fear or joy, label-based representations have some disadvantages. Specifically, labels are not cross-lingual: Emotions do not have exact translations in different languages, e.g., "disgust" does not have an exact translation in Polish [12]. Psychologists therefore often represent emotions or feelings in an n-dimensional space (generally 2 or 3D). The most famous such space, which is used in the present study and originates from cognitive theory, is the 3D valence-arousal-dominance or pleasure-arousal-dominance (PAD) space [13]. The valence scale ranges from unpleasant to pleasant. The arousal scale ranges from passive to active or excited. The dominance scale ranges from submissive (or "without control") to dominant (or "in control, empowered"). Fontaine et al. [14] proposed adding a predictability dimension to PAD dimensions. Predictability level describes to what extent the sequence of events is predictable or surprising for a viewer.

In a preliminary study, 155 video clips containing movie scenes manually selected from 21 commercially produced movies were shown to more than 50 participants; each video clip received 10 annotations on average [15]. The preliminary study was conducted utilizing an online affective annotation system in which the participants reported their emotions in response to the videos played by a web-based video player.

In the preliminary study, the participants were thus asked to self-assess their emotion by reporting the felt arousal (ranging from calm to excited/activated) and valence (ranging from unpleasant to pleasant) on nine points scales. SAM Manikins were shown to facilitate the self-assessments of valence and arousal [16]. Fourteen video clips were chosen based on the preliminary study from the clips which received the highest number of tags in different emotion classes, e.g., the clip with the highest number of sad tags was selected to induce sadness. Three other popular video clips from online resources were added to this set (two for joy and one for disgust). Three past weather forecast reports (retrieved from youtube.com) were also used as neutral emotion clips. The videos from online resources were added to the data set to enable us to distribute some of the emotional video samples with the multimodal database described below. The full list of videos is given in Table 3.

Ultimately, 20 videos were selected to be shown which were between 34.9 and 117 s long ($M = 81.4$ s, $SD = 22.5$ s). Psychologists recommended videos from 1 to 10 minutes long for elicitation of a single emotion [17], [18]. Here, the video clips were kept as short as possible to avoid multiple emotions or habituation to the stimuli while keeping them long enough to observe the effect.

3.2 Facial Expressions and Audio Signals

One of the most well-studied emotional expression channels is facial expressions. A human being uses facial expressions as a natural mean of emotional communication. Emotional expressions are also used in human-human communication to clarify and stress what is said, to signal comprehension, disagreement, and intentions, in brief, to regulate interactions with the environment and other persons in the vicinity [19], [20]. Automatic analysis of facial expression is an interesting topic from both scientific and practical point of view. It has attracted the interest of many researchers since such systems will have numerous applications in behavioral science, medicine, security, and human-computer interaction. To develop and evaluate such applications, large collections of training and test data are needed [21], [22]. In the current database, we are interested in studying the spontaneous responses of participants while

TABLE 3
The Video Clips Listed with Their Sources

Emotion	Excerpt's source
Disgust	Hannibal, The Pianist, Ear worm (blip.tv)
Amusement	Mr. Bean's holiday (2 excerpts), Kill Bill VOL I
Joy	Love actually (2 excerpts), The thin red line, Funny cats (YouTube), Funny (blip.tv)
Fear	The shining (2 excerpts), Silent hill
Sadness	Gangs of New York, The thin red line, American history X
Neutral	AccuWeather New York, Dallas, and Detroit weather report (youtube.com)

The listed emotional keywords were chosen by polling over participants' self-reports in the preliminary study.

4. http://www.eecs.qmul.ac.uk/mmv/data sets/deap/.

(a) frontal view in color (b) close up from the top of the screen

(c) close up from the bottom left (d) close up from the bottom right

(e) profile view (f) wide angle from above the screen

Fig. 1. Snapshots of videos captured from six cameras recording facial expressions and head pose.

watching video clips. This can be used later for emotional implicit tagging of multimedia content.

Fig. 1 shows the synchronized views from the six different cameras. Two types of cameras have been used in the recordings: one Allied Vision Stingray F-046C, color camera (C1), and five Allied Vision Stingray F-046B, monochrome cameras (BW1 to BW5). All cameras recorded with a resolution of 780×580 pixels at 60 frames per second. The two close up cameras above the screen give a near-frontal view of the face in color Fig. 1a or monochrome Fig. 1b. The monochrome views have a better sharpness and less motion blur than the color camera. The two views from the bottom of the screen, Figs. 1c and 1d, give a close up view that may be more useful during down-facing head poses, and make it possible to apply passive stereo imaging. For this, the intrinsic and extrinsic parameters of all cameras have been calibrated. Linear polarizing filters were applied with the two bottom cameras in order to reduce the reflection of the computer screen in eyes and glasses. The profile view Fig. 1e can be used to extract backward-forward head/body movements or to aid the extraction of facial expressions, together with the other cameras. The wide-angle view Fig. 1f captures the upper body, arms and hands, which can also carry important information about a person's affective state.

Although we did not explicitly ask the participants to express or talk during the experiments, we expected some natural utterances and laughter in the recorded audio signals. The audio was recorded for its potential to be used for video tagging, e.g., it has been used to measure the hilarity of videos by analyzing a user's laughter [23]. However, the amount of laughter and audio responses in the database from participants is not enough for such studies and therefore the audio signals were not analyzed. The recorded audio contains two channels. Channel one (or "left" if interpreted as a stereo stream) contains the audio signal from a AKG C 1000 S MkIII room microphone, which includes the room noise as well as the sound of the video stimuli. Channel two contains the audio signal from an AKG HC 577 L head-worn microphone.

3.3 Eye Gaze Data

The Tobii X120[5] eye gaze tracker provides the position of the projected eye gaze on the screen, the pupil diameter, the moments when the eyes were closed, and the instantaneous distance of the participant's eyes to the gaze tracker device. The eye gaze data were sampled at 60 Hz due to instability of the eye gaze tracker system at 120 Hz. The blinking moments are also extractable from eye gaze data by finding the moments in the eye gaze responses where the coordinates are equal to -1. Pupil diameter has been shown to change in different emotional states [24], [25]. Examples of eye gaze responses are shown in Fig. 2.

3.4 Physiological Signals

Physiological responses (ECG, GSR, respiration amplitude, and skin temperature) were recorded with a 1,024 Hz sampling rate and later downsampled to 256 Hz to reduce the memory and processing costs. The trend of the ECG and GSR signals was removed by subtracting the temporal low frequency drift. The low frequency drift was computed by smoothing the signals on each ECG and GSR channels with a 256 points moving average.

GSR provides a measure of the resistance of the skin by positioning two electrodes on the distal phalanges of the middle and index fingers and passing a negligible current through the body. This resistance decreases due to an increase of perspiration, which usually occurs when one is experiencing emotions such as stress or surprise. Moreover, Lang et al. discovered that the mean value of the GSR is related to the level of arousal [26].

ECG signals were recorded using three sensors attached on the participants' body. Two of the electrodes were placed on the chest's upper right and left corners below the clavicle bones and the third electrode was placed on the abdomen below the last rib for setup simplicity. This setup allows precise identification of heart beats and consequently to compute heart rate (HR).

Skin temperature was recorded by a temperature sensor placed participant's little finger. The respiration amplitude was measured by tying a respiration belt around the abdomen of the participant.

Psychological studies regarding the relations between emotions and the brain are uncovering the strong implication of cognitive processes in emotions [27]. As a result, the EEG signals carry valuable information about the participants' felt emotions. EEG signals were recorded using active AgCl electrodes placed according to the international 10-20 system. Examples of peripheral physiological responses are shown in Fig. 2.

4 Experimental Setup

4.1 Experimental Protocol

As explained above, we set up an apparatus to record facial videos, audio and vocal expressions, eye gaze, and physiological signals simultaneously. The experiment was

5. http://www.tobii.com.

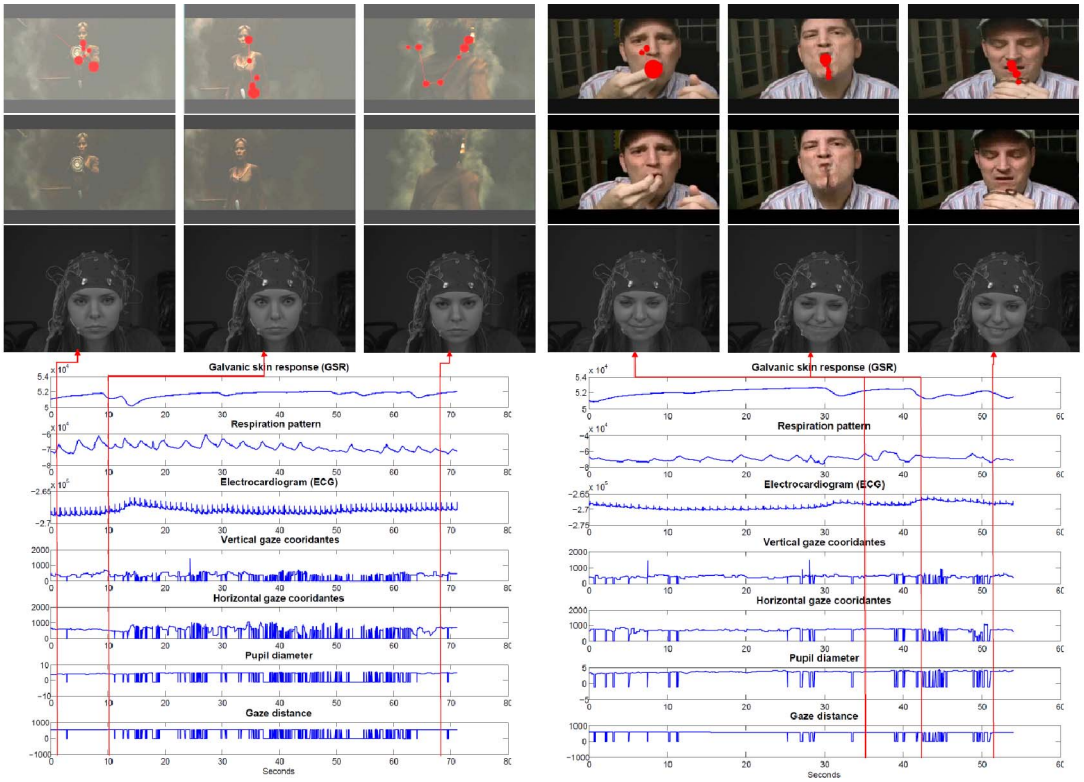


Fig. 2. Natural expressions to a fearful (on the left) and disgusting (on the right) video. The snapshots of the stimuli videos with eye gaze overlaid and without eye gaze overlaid, frontal captured video, raw physiological signals, and raw eye gaze data are shown. In the first row, the red circles show the fixation points and their radius indicates the time spent in each fixation point. The red lines indicate the moments where each of the snapshots was captured.

controlled by the Tobii studio software. The Biosemi active II system[6] with active electrodes was used for physiological signals acquisition. Physiological signals including ECG, EEG (32 channels), respiration amplitude, and skin temperature were recorded while the videos were shown to the participants. In the first experiment, five multiple choice questions were asked during the self-report for each video. For the second experiment, where the feedback was limited to yes and no, two big colored buttons (red and green) were provided.

Thirty participants with different cultural backgrounds volunteered to participate in response to a campus wide call for volunteers at Imperial College, London. Out of the 30 young healthy adult participants, 17 were female and 13 were male; ages varied between 19 to 40 years old ($M = 26.06, SD = 4.39$). Participants had different educational background, from undergraduate students to postdoctoral fellows, with different English proficiency from intermediate to native speakers. The data recorded from three participants (P9, P12, P15) were not analyzed due to technical problems and unfinished data collection. Hence, the analysis results of this paper are only based on the responses recorded from 27 participants.

4.2 Synchronized Setup

An overview of the synchronization in the recording setup is shown in Fig. 3. To synchronize between sensors, we centrally monitored the timings of all sensors, using a MOTU 8pre[7] audio interface ("c" in Fig. 3) that can sample up to eight analog inputs simultaneously. This allowed the derivation of the exact temporal relations between events in each of the eight channels. By recording the external camera trigger pulse signal ("b" in Fig. 3) in a parallel audio track (see the fifth signal in Fig. 4), each recorded video frame could be related to the recorded audio with an uncertainty below 25 μs. More details about the data synchronization can be found in [28].

The gaze tracking data and physiological signals were recorded with separated capture systems. Because neither of

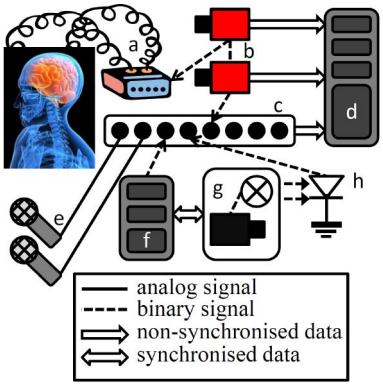


Fig. 3. Overview of our synchronized multisensor data capture system, consisting of (a) a physiological measurement device, (b) video cameras, (c) a multichannel A/D converter, (d) an A/V capture PC, (e) microphones, (f) an eye gaze capture PC, (g) an eye gaze tracker, and (h) a photo diode to capture the pulsed IR-illumination from the eye gaze tracker. Camera trigger was recorded as audio and physiological channels for synchronization.

6. http://www.biosemi.com.
7. http://www.motu.com/products/motuaudio/8pre.

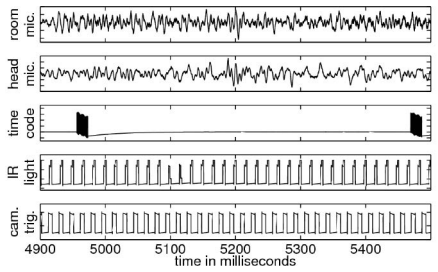


Fig. 4. Five tracks recorded in parallel by MOTU 8pre audio interface. From top to bottom: 1) room microphone, 2) head microphone, 3) serial port time stamp output (transmitted at 9,600 bps), showing 2 time stamp signals, 4) measured infrared light in front of eye tracker, 5) camera trigger.

them allowed the recording of the actual sensor trigger signals, they required alternative synchronization strategies. The physiological data were captured with a multichannel A/D converter ("a" in Fig. 3) that allowed recording one binary input signal alongside the data. This input was used to connect the camera trigger signal. Since the accurate timing of each camera frame is known, this allowed synchronizing the physiological data with all the other modalities.

The eye gaze tracker ("g" in Fig. 3) synchronized with the CPU cycle counter of its dedicated capture PC ("f") with an accuracy of approximately one millisecond. To synchronize the respective CPU cycle counter to the audio interface, we developed an application that periodically (twice per second) outputs binary time stamp signals with the current time, through the serial port output (see the third signal in Fig. 4), with an error below $10\ \mu$s. To get a more accurate timing accuracy than the 1 ms accuracy of the time stamps of the gaze tracking data, the infrared strobe illumination of the gaze tracker was recorded using a photo diode ("h" in Fig. 3 and the fourth signal in Fig. 4). This allowed the correction of the gaze data time stamps with the temporal resolution as high as 10 microseconds.

The start moments of the stimuli data were time stamped using the same synchronized CPU cycle counter as the eye-gaze data. An uncertainty in timing of the stimuli data is introduced by the video player software, as well as by the latency of the audio system, graphics card, and the screen. Furthermore, the accuracy of the time codes of the fragments may introduce further errors in synchronizing the recorded data with the actual stimuli. The room microphone was placed close to the speaker that produced the stimuli sound. Therefore, the recorded ambient sound provides an implicit synchronization, as it includes the sound of the stimuli.

4.3 Practical Considerations

Although the protocol and setup was done carefully, problems arose during recordings. The data recorded from participants 9 and 15 are not complete due to technical problems. The physiological responses of participant 12 are missing due to recording difficulties. The physiological responses to each stimuli were recorded each in a separate file in Biosemi data format (BDF) which is an extension of European data format (EDF) and easily readable in different platforms. For each trial, the response to the 15 seconds neutral video is stored separately. All the files containing physiological signals include the signals recorded 30 s before the start and after the end of their stimuli. In accordance with the Biosemi recording methodology, we did not record a reference electrode with EEG signals. Therefore, EEG signals need rereferencing to a virtual reference, for example, the average reference. The stimuli videos were all encoded in MPEG-4 Xvid format and MPEG layer three format with 44,100 Hz sampling frequency in an audio video interleave container (AVI). The frames were encoded in $1{,}280 * 800$ to match our display resolution.

5 Emotion Recognition Experiment

In this section, we present the emotion recognition experimental paradigm, analysis methods, and experimental results. Three modalities, including peripheral and central nervous system physiological signals and information captured by eye gaze tracker, were used to recognize emotions from participants' responses.

5.1 Emotion Experiment Paradigm

The participants were informed about the experiment, and their rights, in a verbal introduction, by e-mail, and through a consent form. Participants were trained to use the interface before the experiment and during the setup time. The participants were also introduced to the meaning of arousal, valence, dominance, and predictability in the self-assessment procedure, and to the nature of the video content. The five questions which were asked during self-reporting were

1. emotional label/tag,
2. arousal,
3. valence,
4. dominance,
5. predictability [14].

The emotional labels included neutral, anxiety, amusement, sadness, joy, disgust, anger, surprise, and fear. To simplify the interface for the first experiment a keyboard was provided with only nine numerical keys and the participants answered each question by pressing one of the keys. Questions 2 to 5 were on a nine point scale.

In emotional-affective experiments the bias from the emotional state needs to be reduced. For this purpose, a short neutral clip was shown to the participants before each emotional video. The neutral clip was randomly selected from the clips provided by the Stanford psychophysiology laboratory [18]. The 20 emotional video clips were played from the data set in random order. After watching a video clip, the participant filled in the self-assessment form. In total, the time interval between the start of a trial and the end of the self-reporting phase was approximately two and half minutes. This interval included playing the neutral clip, playing the emotional clip, and performing the self-assessment. Running of the whole protocol took, on average, 50 minutes, in addition to 30 minutes set up time.

5.2 Emotional Features

5.2.1 EEG and Physiological Signals

The following peripheral nervous system signals were recorded: GSR, respiration amplitude, skin temperature,

and ECG. Most of the current theories of emotion [29] agree that physiological activity is an important component of an emotion. Heart rate and heart rate variability (HRV) correlate with emotional changes. Pleasantness of stimuli can increase peak heart rate response [26], and HRV decreases with fear, sadness, and happiness [30]. In addition to the HR and HRV features, spectral features derived from HRV were shown to be a useful feature in emotion assessment [31]. Skin temperature was also recorded since it changes in different emotional states [32]. Regarding the respiration amplitude, slow respiration is linked to relaxation, while irregular rhythm, quick variations, and cessation of respiration correspond to more aroused emotions like anger or fear [33], [30]. In total, 102 features were extracted from peripheral physiological responses based on the proposed features in the literature [34], [33], [30].

In addition to the peripheral nervous system responses, electroencephalogram signals were acquired. The power spectral features were extracted from EEG signals. The logarithms of the spectral power from theta ($4\text{ Hz} < f < 8\text{ Hz}$), slow alpha ($8\text{ Hz} < f < 10\text{ Hz}$), alpha ($8\text{ Hz} < f < 12\text{ Hz}$), beta ($12\text{ Hz} < f < 30\text{ Hz}$), and gamma ($30\text{ Hz} < f$) bands were extracted from all 32 electrodes as features. In addition to power spectral features, the difference between the spectral power of all the symmetrical pairs of electrodes on the right and left hemispheres was extracted to measure the possible asymmetry in the brain activities due to the valence of perceived emotion [35], [36]. The asymmetry features were extracted from all mentioned bands except slow alpha. The total number of EEG features of a trial for 32 electrodes is $14 \times 4 + 32 \times 5 = 216$ features. The total number of EEG features of a trial for 32 electrodes is 216 features. (Table 4 lists the features extracted from the physiological signals.)

TABLE 4
This Table Lists All 102 Features Extracted from Physiological Signals

Signal	Extracted features
GSR (20)	average skin resistance, average of absolute derivative, proportion of negative samples in the derivative vs. all samples, 13 spectral power in the bands in [0, 2.4]Hz, zero crossing rate of Skin conductance slow response (SCSR in [0, 0.2]Hz), zero crossing rate of Skin conductance very slow response (SCVSR in [0, 0.08]Hz), SCSR and SCVSR mean of peaks magnitude
ECG (64)	HRV, root mean square of the mean squared difference of successive beats, standard deviation (SD) of beat interval change per respiratory cycle [30], 56 spectral power in the bands from [0, 6]Hz, low frequency [0.01, 0.08]Hz, medium frequency [0.08, 0.15]Hz and high frequency [0.15, 0.5]Hz components of HRV power spectrum, poincaré analysis features (2 features) [33]
Respiration pattern (14)	band energy ratio (difference between the logarithm of energy between the lower ([0.05, 0.25]Hz) and the higher ([0.25, 5]Hz) bands), range, mean of derivative (variation of the respiration signal), breathing rhythm (spectral centroid), breathing rate, average breathe depth (peak to peak), 8 spectral power in the bands in [0, 2.4]Hz
Skin temperature (4)	average, average of its derivative, spectral power in the bands ([0-0.1]Hz,]0.1-0.2]Hz)
EEG (216)	theta, slow alpha, alpha, beta, and gamma Spectral power for each electrode. The spectral power asymmetry between 14 pairs of electrodes in the four bands of alpha, beta, theta and gamma.

Number of features extracted from each channel is given in brackets.

5.2.2 Eye Gaze Data

After removing the linear trend, the power spectrum of the pupil diameter variation was computed. Standard deviation and spectral features were extracted from the pupil diameter. The Hippus effect is the small oscillations of the eye pupil diameter between 0.05 and 0.3 Hz and with amplitude of 1 mm [37], [38]. The Hippus effect has been shown to be present when one is relaxed or passive. In the presence of mental activity the effect will disappear. The Hippus effect is extracted by the first two power spectral features which are covering up to 0.4 Hz.

Eye blinking was extracted by counting the number of times when eye gaze data were not available, i.e., the moments when eyes were closed. The rate of eye blinking is shown to be correlated with anxiety. From the eye blinks, the eye blinking rate, the average and maximum blink duration were extracted as features. In addition to the eye blinking features the amount of time each participant spent with his/her eyes closed was also used as a feature to detect possible eye closing due to unpleasant emotions.

Although participants were asked not to move during experiments, there were small head movements which manifested themselves in the distance between participants' eyes and the eye gaze tracker. The distance of participants to the screen and its changes provide valuable information about participants' posture. The total change in the distance of a participant to the gaze tracker, gaze distance, was calculated to measure the possible approach and avoidance. The amount of time participants spent per trial getting close to or far from the screen was computed as well. These features were named approach and avoidance ratio to represent the amount of time participants spent getting close to or going far from the screen. The frequency of the participants' movement toward the screen during each trial, approach rate, was also extracted. Approach and withdrawal are closely related to emotional experiences [39].

The statistical features were extracted from eye gaze coordinates along the horizontal and vertical axes, namely, the standard deviation, Kurtosis, and skewness of horizontal and vertical projected eye gaze. Moreover, the power spectral density in different bands was extracted to represent different oscillations the eye gaze pattern (see Table 5). These bands were empirically chosen based on the spectral changes in eye gaze movements. Ultimately, 38 features were extracted from the eye gaze data. The features extracted from eye gaze data are listed in Table 5.

5.3 Rating Analysis

Regarding the self-reports, we computed the multirater Cohen's kappa for different annotations. A fair agreement was found on emotional keywords with the average $\kappa = 0.32$. For arousal and valence rating, the cross correlation values were computed which was ($M = 0.40, SD = 0.26$) for

TABLE 5
This Table Lists the Features Extracted from Eye Gaze Data for Emotion Recognition

Eye gaze data	Extracted features
Pupil diameter (6)	average, standard deviation (SD), spectral power in the following bands:]0, 0.2]Hz,]0.2, 0.4]Hz,]0.4, 0,6]Hz and]0.6, 1]Hz
Gaze distance (4)	approach time ratio, avoidance time ratio, approach rate, average approach time
Eye blinking (4)	blink depth, blinking rate, length of the longest blink, time spent with eyes closed
Gaze coordinates (for both horizontal and vertical axis) (24)	SD, skewness, Kurtosis, average fixation time, average scan path length, number of fixation zones (normalized by the video length), spectral power in the following bands]0, 0.2]Hz,]0.2, 0.04]Hz,]0.4, 0.6]Hz,]0.6, 0.8]Hz,]1, 2]Hz, average and SD of the SD of gaze coordinates in each fixation zone

Number of features extracted from each channel is given in brackets.

arousal and ($M = 0.71, SD = 0.12$) for valence. The keyword-based feedbacks were used to generate each participant's ground truth. The histograms of emotional self-reports' keywords and ratings given to all videos are shown in Fig. 5. In Fig. 5, it is visible that the emotions which were not initially targeted (see Table 3) have the least frequencies.

5.4 Emotion Recognition Results

In order to give the reader some baseline classification results, emotion recognition results from three modalities and fusion of ebest modalities are presented. Two classification schemes were defined: first, along the arousal dimension, three classes of calm, medium aroused, and excited, and second along the valence dimension, unpleasant, neutral valence and pleasant. The mapping between emotional keyword and classes which are based on [14] and are given in Table 6.

A participant independent approach was taken to check whether we can estimate a new participant's felt emotion based on others. For each video from the data set, the ground truth was thus defined by the feedback given by each participant individually. The keyword-based feedback was then translated into the defined classes. According to this definition, we can name these classes calm, medium aroused, and excited/activated for arousal and unpleasant, neutral valence, and pleasant for valence (see Table 6).

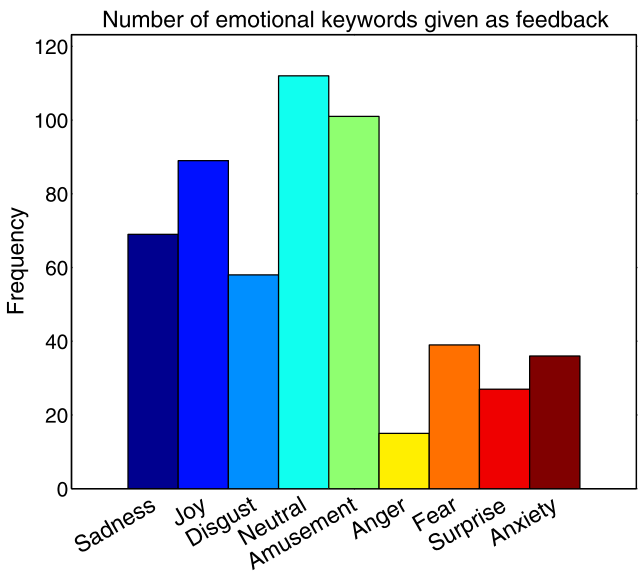


Fig. 5. This bar chart shows the frequency of the emotional keywords assigned to all videos.

TABLE 6
The Emotional Keywords Are Mapped into Three Classes on Arousal and Valence

Arousal classes	Emotional keywords
Calm	sadness, disgust, neutral
Medium arousal	joy and happiness, amusement
Excited/Activated	surprise, fear, anger, anxiety

Valence classes	Emotional keywords
Unpleasant	fear, anger, disgust, sadness, anxiety
Neutral valence	surprise, neutral
Pleasant	joy and happiness, amusement

To reduce the between participant differences, it is necessary to normalize the features. Each feature was separately normalized by mapping to the range $[0, 1]$ on each participant's signals. In this normalization the minimum value for any given feature is subtracted from the same feature of a participant and the results were divided by the difference between the maximum and minimum values.

A leave-one-participant-out cross validation technique was used to validate the user independent classification performance. At each step of cross validation, the samples of one participant were taken out as test set and the classifier was trained on the samples from the rest of the participants. This process was repeated for all participants' data. An implementation of the SVM classifier from libSVM [40] with RBF kernel was employed to classify the samples using features from each of the three modalities. For the SVM classifier, the size of the kernel, γ, was selected between $[0.01, 10]$, based on the average F1 score using a 20-fold cross validation on the training set. The C parameter that regulates the tradeoff between error minimization and margin maximization is empirically set to 1. Prior to classification, a feature selection was used to select discriminative features as follows: First, a one-way ANOVA test was done on the training set for each feature with the class as the independent variable. Then, any feature for which the ANOVA test was not significant ($p > 0.05$) was rejected.

Here, we used three modalities which are peripheral physiological signals, EEG, and eye gaze data. From these three modalities, the results of the classification over the two best modalities were fused to obtain the multimodal fusion results. If the classifiers provide confidence measures on their decisions, combining decisions of classifiers can be done using a summation rule. The confidence measure summation fusion was used due to its simplicity and its proven performance for emotion recognition according to [34].

The data from the 27 participants which had enough completed trials was used. Five hundred thirty-two samples of physiological responses and gaze responses were gathered over a potential data set of $27 \times 20 = 540$ samples;

TABLE 7
The Recognition Rate and F1 Scores of Emotion Recognition for Different Modalities and Fusion of the Two Best Modalities, EEG and Eye Gaze

Modality	Classification rate		Average F1	
dimension	arousal	valence	arousal	valence
Peripheral physiology	46.2%	45.5%	0.38	0.39
EEG	52.4%	57.0%	0.42	0.56
Eye gaze	63.5%	68.8%	0.60	0.68
Fusion (EEG & Gaze)	**67.7%**	**76.1%**	**0.62**	**0.74**
Random (uniform)	33.3%	33.3%	0.36	0.34
Random (weighted)	35.8%	34.4%	0.33	0.33

the eight missing ones were unavailable due to not having enough valid samples in eye gaze data.

The F1 scores and recognition rates for the classification in different modalities are given in the Table 7 and Fig. 6. In Table 7, two random level results are also reported. The first random level results represent a random classifier with uniform distribution, whereas the second random classifier (weighted) uses the training set distribution to randomly choose the class. The confusion matrices for each modality and their fusion show how they performed on each emotion class (Table 8). In these confusion matrices the row represents the classified label and each column represents the ground truth for those samples. For all cases, classification on gaze data performed better than EEG and peripheral signals. This is due to the fact that eye gaze is more correlated with the shown content and similar visual features induce similar emotions. The peripheral physiological responses have a high variance between different participants which makes interparticipant classification difficult to perform. Therefore, the classification using peripheral physiological features gave the worst results among these three modalities. This can be reduced in future studies by using better methods to reduce the between participants' variance. The high arousal, "activated," class was the most challenging class. While EEG and peripheral physiological modalities were completely unable to classify the samples, eye gaze also did not obtain its superior accuracy for this class (see Fig. 5). This might have been caused by the lower number of responses for the emotions assigned to this class. The fusion of the two best modalities, eye gaze and EEG, ultimately outperformed all single modalities.

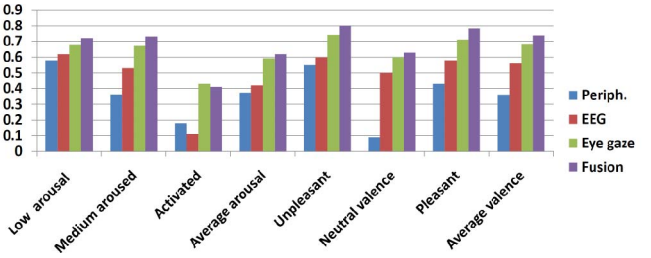


Fig. 6. This bar chart shows the F1 score for classification results of each class from different modalities.

6 Implicit Tagging Experiment

In this section, we describe the implicit tagging experiment paradigm, analysis methods, and experimental results. Two modalities were used to predict the correctness of displayed tags, namely, facial expression (captured by a camera) and the eye gaze location on the screen (captured by an eye gaze tracker). The results presented in this section are limited to the static images only. The sequences with tagged videos were not processed. In total, we analyzed 756 data sequences of 27 participants with the goal of recovering the correctness of the displayed tags. The average sequence length was 5s. The method utilized for facial expression analysis and its results were previously published in [41].

6.1 Implicit Tagging Experiment Paradigm

In this second experiment, 28 images and 14 video fragments were subsequently shown on their own and accompanied by a word tag. Once using a correct tag and once using an incorrect tag. The videos were chosen from the Hollywood human actions database (HOHA) [3] and were between 12 and 22 seconds long ($M = 17.6$ s, $SD = 2.2$ s). For each trial, the following procedure was taken:

1. Untagged Stimulus: The untagged stimulus was displayed (still images were shown during 5 seconds). This allowed the participant to get to know the content of the image/video.
2. Tagged Stimulus: The same stimulus was displayed with tag (still images were shown during 5 seconds). The participants' behavior in this period contained their reaction to the displayed tag.

TABLE 8
Confusion Matrices of Different Classification Schemes (Row: Classified Label; Column: Ground Truth)

Arousal

	1	2	3
1	0.75	0.62	0.65
2	0.22	0.32	0.24
3	0.03	0.06	0.11

(a) Peripheral

	1	2	3
1	0.75	0.45	0.65
2	0.22	0.53	0.29
3	0.03	0.02	0.06

(b) EEG

	1	2	3
1	0.73	0.26	0.43
2	0.18	0.69	0.22
3	0.09	0.05	0.35

(c) Eye Gaze

	1	2	3
1	0.80	0.22	0.53
2	0.16	0.75	0.18
3	0.04	0.03	0.29

(d) DLF

Valence

	1	2	3
1	0.71	0.63	0.54
2	0.03	0.05	0.04
3	0.26	0.32	0.42

(e) Peripheral

	1	2	3
1	0.64	0.35	0.27
2	0.09	0.44	0.13
3	0.27	0.21	0.60

(f) EEG

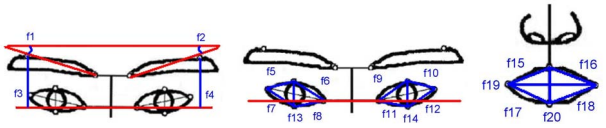
	1	2	3
1	0.72	0.20	0.15
2	0.11	0.56	0.10
3	0.17	0.24	0.75

(g) Eye gaze

	1	2	3
1	0.87	0.26	0.15
2	0.04	0.50	0.03
3	0.09	0.24	0.82

(h) DLF

The numbers on the first row and the first column of tables (a), (b), (c), and (d) represent: 1. calm, 2. medium aroused, 3. activated, and for tables (e), (f), (g), and (h) represent: 1. unpleasant 2. neutral valence 3. pleasant. The confusion matrices relate to classification using (a), (e) peripheral physiological signals, (b), (f) EEG signals, (c), (g) eye gaze data, (d), (h) EEG and eye gaze decision-level fusion.



(a) Points on eyes (b) Points on eyes (c) Points on lip

Fig. 7. The tracked points and features.

3. Question: A question was displayed on the screen to ask whether the participant agreed with the suggested tag. Agreement or disagreement was expressed by pressing a green or a red button, respectively.

Only the color video capturing the frontal view of participants' faces was used in the analysis (see Fig. 1a). The length of each trial was about 11 seconds for images, and slightly more than double the stimulus' length in case of videos. The recording of each participant was segmented in three sets of 28 small clips, according to the order in which the images/videos were presented. Each fragment corresponds to the period between the time point when the stimulus appears and the point when the participant has given his/her feedback. Running of the second experiment's protocol took in average 20 minutes, excluding the setup time.

6.2 Facial Expression Analysis

To extract facial features, the Patras-Pantic particle filter [42] was employed to track 19 facial points. The initial positions of the 19 tracked points were manually labeled for each video and then automatically tracked for the rest of the sequence. After tracking, each frame of the video was represented as a vector of the facial points' 2D coordinates. For each frame, geometric features, f1 to f20, were then extracted based on the positions of the facial points. The extracted features are:

- **Eyebrows**: Angles between the horizontal line connecting the inner corners of the eyes and the line that connects inner and outer eyebrow (f1, f2), the vertical distances from the outer eyebrows to the line that connects the inner corners of the eyes (f3, f4) (see Fig. 7a).
- **Eyes**: Distances between the outer eyes' corner and their upper eyelids (f5, f9), distances between the inner eyes' corner and their upper eyelid (f6, f10), distances between the outer eyes' corner and their lower eyelids (f8, f12), distances between the inner eyes' corner and their lower eyelids (f7, f11), vertical distances between the upper eyelids and the lower eyelids (f13, f14) (see Fig. 7b).
- **Mouth**: Distances between the upper lip and mouth corners (f15, f16), distances between the lower lip and mouth corners (f18, f18), distances between the mouth corners (f19), vertical distance between the upper and the lower lip (f20) (see Fig. 7c).

The line that connects the inner eye corners was used as a reference line since the inner eye corners are stable facial points, i.e., changes in facial expression do not induce any changes in the position of these points. They are also the two most accurately tracked points. For each sequence, the listed 20 features were extracted for all the frames. The difference of these 20 features with their values in a neutral frame was used in further processing.

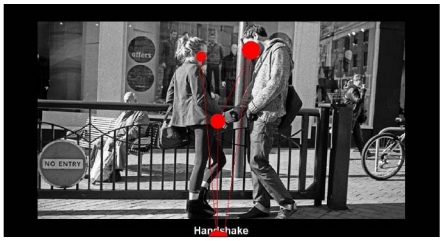


Fig. 8. An example of displayed images is shown with eye gaze fixation and scan path overlaid. The size of the circles represents the time spent staring at each fixation point.

6.3 Analysis of Eye Gaze

Gaze fixations are the coordinates of the points on the display on which the eye gaze stayed fixed for a certain period of time. Each fixation is composed of its duration as well as the two-dimensional coordinates of the projection of the eye gaze on the screen. An example of an eye gaze pattern and fixations points on an image is shown in Fig. 8. The features extracted from eye gaze are listed in Table 9.

6.4 Classification Methods

We have chosen Hidden Markov Models (HMMs) to classify the facial expression sequences according to the correctness of the displayed tags. HMMs have been commonly used for modeling dynamic sequences. For a more detailed description of the utilized HMM framework for facial expression analysis, see the early work on facial-expression-based implicit tagging [41].

As shown in [43], a temporal facial movement consists of four states:

1. neutral—there are no signs of muscular activation;
2. onset—the muscular contraction begins and increases in intensity;
3. apex—a plateau where the intensity reaches a stable level;
4. offset—the relaxation of muscular action.

Based on these states, the number of modeled states in the HMM was set to four. We chose the ergodic topology, in

TABLE 9
This Table Lists the 19 Features Extracted from Eye Gaze Data for Implicit Tagging

Eye gaze	Extracted features
Fixation (15)	average fixations, variance of fixations, total fixations' duration in the tag zone, total fixations' duration in the image, proportion of time spent looking at image and not at the tag, average fixation on image, average fixation on tag, fixation duration ratio between tag and image, maximum fixation duration on image, maximum fixation duration on tag, number of fixations, number of fixations on image, number of fixations on tag, fixation number on image to tag ratio, entropy of fixation durations
Scan path (4)	average scan path length, variance scan path length, total length of the scan path, number of transitions between image and tag zones

which all the states are connected with each other. This means that it is possible to jump from any of the four states to any other. The initial state and transition probabilities were randomly chosen. The emission probabilities were modeled by a mixture of two Gaussians with diagonal covariance matrices. For the implementation of the utilized HMM, the HMM toolbox for MATLAB was used. For each participant, two HMMs were trained: one for the participant's facial responses when the image with a correct tag was displayed and the other for when the incorrect tag was shown. The correctness of the displayed tags was predicted by comparing the likelihood of these two HMMS. The performance of the classification was investigated by a 10-fold cross validation.

Adaboost was employed for the classification of eye gaze data. The general idea of Adaboost is to combine a group of weak classifiers to form a strong classifier. A set of classifiers were trained sequentially and then combined [44]. The later generated classifiers focused more on the mistakes of the earlier classifiers. Let $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ denote the instances we have. For the gaze data, each $\vec{x_i}$ is a 19-dimensional feature vector and $y_i = -1, +1$ is its associated label. The weak classifiers used in this paper are decision stumps in the following form:

$$c(x, i, p, \lambda) = \begin{cases} 1 & \text{if } px^i < p^\lambda \\ -1 & \text{otherwise,} \end{cases} \quad (1)$$

in which i, p, λ are the parameters of a decision stump. i is the feature chosen, p is a polarity flag with value 1 or -1, λ is the decision threshold, x^i is an instance, and x^i is the value of the ith feature of x. The decision stump simply chooses a feature from the instance and then compares its value to a threshold. The algorithm takes two parameters: the data set and the number of iterations T. It generates a sequence of weak classifiers and combines them with weights

$$H(x) = \sum_{t=1} \alpha_t h_t(x). \quad (2)$$

At each iteration, Adaboost chooses the decision stump that minimizes the weighted classification error, which is equivalent to choosing the most discriminative feature. The weight for the weak classifier $\alpha_t = \frac{1}{2}\ln(\frac{1-\epsilon_t}{\epsilon_t})$ decreases when the error increases. The weights are then updated based on the classification result. A larger weight will be assigned to the misclassified samples at each iteration. This increases the importance of the misclassified sample in the next iteration. For the gaze data, 19 features were chosen to model the gaze pattern.

6.5 Facial Expression Results

At the single participant level, the best predication rate was not better than 62 percent. This led to using a strategy to predict the correctness of the displayed tag by the weighted sum of the predictions made from the behavior of different participants. In this combination, the prediction rate on the training set was used as the combining weight of the prediction based on a single participant. This strategy attenuates the negative effect of participants with less consistent behavior in the measured features.

TABLE 10
Prediction Rate of Fusing HMM for Facial Expression Analysis and Fusing Adaboost for Eye Gaze Analysis

N	1	2	3	4	5	6
Facial expression	58%	58%	57%	56%	57%	55%
Eye gaze	64%	64%	66%	73%	66%	61%

N is the number of combined classifiers.

The results of the fusion of the participants' responses are presented in Table 10. N is the number of classifiers, of which each was trained on a single participant's responses. The classifiers whose weights are among the top N are fused using our proposed method.

The best result is achieved by using only one classifier, which means only using one participant's data. Note that the results of $N = 1$ and $N = 2$ should always be the same because the classifier with the second largest weight cannot change a binary decision made by the top 1 classifier. The best result of 58 percent is worse than the best prediction rate on a single participant, which is 62 percent. The prediction rate decreases as the number of participants increases, which implies that the number of effective participants is very limited.

6.6 Eye Gaze Results

Using the gaze responses at the single participant level, the best prediction rate using Adaboost was 66 percent. Here, the same combination strategy was employed to combine multiple participants' responses to detect the correctness of tags.

The result of combining Adaboosts from different participants is presented in Table 10. N is again the number of participants used. When N is equal to 1 or 2, the results are slightly lower than the best result on the single participant. This might be due to the selection of overfitted participants from the training set. The performance of those ineffective participants can be regarded as nearly random guess.

As noted previously, the results for $N = 1$, $N = 2$ are the same. As N goes to 4, the prediction rate increases to 73 percent. Unlike the facial data, combining different participants' gaze responses improves the overall prediction rate. When N is larger than 4, the result starts to deteriorate. This indicates that the number of effective participants is also limited here.

6.7 Modality Fusion

After combining the responses from different participants, the combined classifiers of facial expression and eye gaze were fused at the decision level using a weighted sum of confidence measures. The assigned weights were found based on the performance of fusion on the training set. Since better results were obtained with gaze features, the weight of gaze confidence was set to one and the weight of facial analysis confidence was between $[0, 1]$.

The prediction rate improved from 73.2 to 75 percent by combining the facial and gaze results. The best result was achieved when the facial analysis confidence weight was between 0.1 and 0.35, which gives us an estimate of the relative importance of the two modalities. These results

show that a participant's facial and eye gaze responses convey information about the correctness of tags associated with multimedia content.

7 Discussions and Recommendations

To our knowledge, MAHNOB-HCI is the first database which has five modalities precisely synchronized, namely, eye gaze data, video, audio, and peripheral and central nervous system physiological signals. This database can be of interest to researchers in different fields, from psychology and affective computing to multimedia. In addition to emotion recognition from a single modality or multiple modalities, the relations between simultaneous emotion-related activity and behavior can be studied. The high accuracy of synchronization of this database allows studying the simultaneous effects of emotions on EEG and other modalities, and fusing them in any desired way, from decision-level fusion (DLF) down to combined processing at the signal level. As the results reflect, not all recorded modalities are as correlated with the stimuli as the others. For example, there are not enough audio events and the peripheral physiological signals do not give the same emotion recognition results comparing to the eye gaze and EEG signals.

In any emotional experiment, having more trials gives the opportunity for single participant studies. At the same time, longer sessions make participants tired and unable to feel the stimuli emotions. Considering this tradeoff, we found that an upper limit of 20 video fragments was acceptable. Although 20 videos are enough to compute significant correlations, this number of samples is not sufficient for single-participant emotion recognition.

Inducing emotions and recording affective reactions is a challenging task. Special attention needs to be paid to several crucial factors, such as stimuli that are used, the laboratory environment, as well as the recruitment of participants. Our experience can be distilled into a list of recommendations that will enable the development of additional corpora to proceed smoothly.

The experiment environment in the laboratory should be kept isolated from the outside environment. The participants should not be able to see the examiners or hear noise from the outside. The light and temperature should be controlled to avoid variation in physiological reactions due to uncontrolled parameters.

Choosing the right stimuli material is an important factor in any affective study. They should be long enough to induce emotions and short enough to prevent boredom. Furthermore, to be sure variation in stimuli length does not introduce variance in the measurements between emotional and nonemotional stimuli, we suggest the stimuli durations be equal. The mixture of contradicting emotions can make problems for self-assessments. We recommend using videos which do not induce multiple emotions. A correct participant recruitment can make a big difference in the results. Due to the nature of affective experiments, a motivated participant with the right knowledge for filling the questionnaire in is desirable. The rewards can make the participants more motivated and responsible. However, cash compensation might attract participants who are not motivated or lack desired communication skills. Therefore, rewarded recruitment should be done by carefully considering the desired qualifications. Contact lenses usually cause participants to blink more, which introduces a higher level of artifacts on EEG signals. Therefore, participants with visual correction should avoid using contact lenses as much as possible. Thick glasses affect the eye gaze tracker performance. In the experiments in which both these modalities are recorded, recruiting participants with no visual correction is advised. Properly attending to participants takes an important part of one's attention, which can easily lead to forgetting parts of complicated technical protocols. Therefore, operation of the recording equipment during the experiments should be made as simple as possible (preferably just by pressing a single button). Alternatively, the tasks of controlling and monitoring correct data collection and attending to participants can be divided between multiple laboratory assistants with carefully defined procedures.

8 Conclusions

A multimodal affective database has been recorded and made available to the affective computing community. The large collection of modalities recorded (multicamera video of face, head, speech, eye gaze, pupil size, ECG, GSR, respiration amplitude, and skin temperature) and the high synchronization accuracy between them makes this database a valuable contribution to the ongoing development and benchmarking of emotion-related algorithms that exploit data fusion, as well as to studies on human emotion and emotional expression. Emotion recognition and implicit tagging results from different modalities set a baseline result for researchers who are going to use the database in the future.

Acknowledgments

The work of Soleymani and Pun was supported in part by the Swiss National Science Foundation and in part by the European Community's Seventh Framework Programme (FP7/2007-2011) under grant agreement Petamedia no 216444. The data acquisition part of this work and the work of Pantic and Lichtenauer were supported by the European Research Council under the ERC Starting Grant agreement no. ERC-2007-StG-203143 (MAHNOB). The authors would like to thank J. Doboš, Prof. D. Grandjean, and Dr. G. Chanel for their valuable scientific contributions to the experiments' protocol. They also acknowledge the contributions of J. Jiao for the analysis on the implicit tagging.

References

[1] M. Pantic and A. Vinciarelli, "Implicit Human-Centered Tagging," *IEEE Signal Processing Magazine*, vol. 26, no. 6, pp. 173-180, Nov. 2009.

[2] Z. Zeng, M. Pantic, G.I. Roisman, and T.S. Huang, "A Survey of Affect Recognition Methods: Audio, Visual, and Spontaneous Expressions," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 31, no. 1, pp. 39-58, Mar. 2009.

[3] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld, "Learning Realistic Human Actions from Movies," *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, pp. 1-8, June 2008.

[4] J.A. Healey and R.W. Picard, "Detecting Stress during Real-World Driving Tasks Using Physiological Sensors," *IEEE Trans. Intelligent Transportation Systems,* vol. 6, no. 2, pp. 156-166, June 2005.
[5] M. Pantic, M. Valstar, R. Rademaker, and L. Maat, "Web-Based Database for Facial Expression Analysis," *Proc. IEEE Int'l Conf. Multimedia and Expo,* pp. 317-321, 2005.
[6] E. Douglas-Cowie, R. Cowie, I. Sneddon, C. Cox, O. Lowry, M. McRorie, J.-C. Martin, L. Devillers, S. Abrilian, A. Batliner, N. Amir, and K. Karpouzis, "The HUMAINE Database: Addressing the Collection and Annotation of Naturalistic and Induced Emotional Data," *Proc. Second Int'l Conf. Affective Computing and Intelligent Interaction,* A. Paiva et al., pp. 488-500, 2007.
[7] M. Grimm, K. Kroschel, and S. Narayanan, "The Vera am Mittag German Audio-Visual Emotional Speech Database," *Proc. IEEE Int'l Conf. Multimedia and Expo,* pp. 865-868, Apr. 2008.
[8] G. McKeown, M.F. Valstar, R. Cowie, and M. Pantic, "The SEMAINE Corpus of Emotionally Coloured Character Interactions," *Proc. IEEE Int'l Conf. Multimedia and Expo,* pp. 1079-1084, July 2010.
[9] S. Koelstra, C. Mühl, M. Soleymani, A. Yazdani, J.-S. Lee, T. Ebrahimi, T. Pun, A. Nijholt, and I. Patras, "DEAP: A Database for Emotion Analysis Using Physiological Signals," *IEEE Trans. Affective Computing,* vol. 3, no. 1, pp. 18-31, Jan.-Mar. 2012.
[10] M.F. Valstar and M. Pantic, "Induced Disgust, Happiness and Surprise: An Addition to the MMI Facial Expression Database," *Proc. Int'l Conf. Language Resources and Evaluation, Workshop EMOTION,* pp. 65-70, May 2010.
[11] E. Douglas-cowie, R. Cowie, and M. Schröder, "A New Emotion Database: Considerations, Sources and Scope," *Proc. ISCA Int'l Technical Research Workshop Speech and Emotion,* pp. 39-44, 2000.
[12] J.A. Russell, "Culture and the Categorization of Emotions," *Psychological Bull.,* vol. 110, no. 3, pp. 426-450, 1991.
[13] J.A. Russell and A. Mehrabian, "Evidence for a Three-Factor Theory of Emotions," *J. Research in Personality,* vol. 11, no. 3, pp. 273-294, Sept. 1977.
[14] J.R.J. Fontaine, K.R. Scherer, E.B. Roesch, and P.C. Ellsworth, "The World of Emotions Is Not Two-Dimensional," *Psychological Science,* vol. 18, no. 12, pp. 1050-1057, 2007.
[15] M. Soleymani, J. Davis, and T. Pun, "A Collaborative Personalized Affective Video Retrieval System," *Proc. Third Int'l Conf. Affective Computing and Intelligent Interaction and Workshops,* Sept. 2009.
[16] J.D. Morris, "Observations: Sam: The Self-Assessment Manikin; An Efficient Cross-Cultural Measurement of Emotional Response," *J. Advertising Research,* vol. 35, no. 8, pp. 63-38, 1995.
[17] A. Schaefer, F. Nils, X. Sanchez, and P. Philippot, "Assessing the Effectiveness of a Large Database of Emotion-Eliciting Films: A New Tool for Emotion Researchers," *Cognition and Emotion,* vol. 24, no. 7, pp. 1153-1172, 2010.
[18] J. Rottenberg, R.D. Ray, and J.J. Gross, "Emotion Elicitation Using Films," *Handbook of Emotion Elicitation and Assessment,* series in affective science, pp. 9-28, Oxford Univ. Press, 2007.
[19] *The Psychology of Facial Expression,* J. Russell and J. Fernandez-Dols, eds. Cambridge Univ. Press, 1997.
[20] D. Keltner and P. Ekman, *Facial Expression of Emotion,* second ed., pp. 236-249. Guilford Publications, 2000.
[21] T. Kanade, J.F. Cohn, and T. Yingli, "Comprehensive Database for Facial Expression Analysis," *Proc. IEEE Fourth Int'l Conf. Automatic Face and Gesture Recognition,* pp. 46-53, 2000.
[22] M. Pantic and L.J.M. Rothkrantz, "Toward an Affect-Sensitive Multimodal Human-Computer Interaction," *Proc. IEEE,* vol. 91, no. 9, pp. 1370-1390, Sept. 2003.
[23] S. Petridis and M. Pantic, "Is This Joke Really Funny? Judging the Mirth by Audiovisual Laughter Analysis," *Proc. IEEE Int'l Conf. Multimedia and Expo,* pp. 1444-1447, 2009.
[24] M.M. Bradley, Miccoli, Laura, Escrig, A. Miguel, Lang, and J. Peter, "The Pupil as a Measure of Emotional Arousal and Autonomic Activation," *Psychophysiology,* vol. 45, no. 4, pp. 602-607, July 2008.
[25] T. Partala and V. Surakka, "Pupil Size Variation as an Indication of Affective Processing," *Int'l J. Human-Computer Studies,* vol. 59, nos. 1/2, pp. 185-198, 2003.
[26] P.J. Lang, M.K. Greenwald, M.M. Bradley, and A.O. Hamm, "Looking at Pictures: Affective, Facial, Visceral, and Behavioral Reactions," *Psychophysiology,* vol. 30, no. 3, pp. 261-273, 1993.
[27] R. Adolphs, D. Tranel, and A.R. Damasio, "Dissociable Neural Systems for Recognizing Emotions," *Brain and Cognition,* vol. 52, no. 1, pp. 61-69, June 2003.
[28] J. Lichtenauer, J. Shen, M. Valstar, and M. Pantic, "Cost-Effective Solution to Synchronised Audio-Visual Data Capture Using Multiple Sensors," technical report, Imperial College London, 2010.
[29] D. Sander, D. Grandjean, and K.R. Scherer, "A Systems Approach to Appraisal Mechanisms in Emotion," *Neural Networks,* vol. 18, no. 4, pp. 317-352, 2005.
[30] P. Rainville, A. Bechara, N. Naqvi, and A.R. Damasio, "Basic Emotions Are Associated with Distinct Patterns of Cardiorespiratory Activity," *Int'l J. Psychophysiology,* vol. 61, no. 1, pp. 5-18, July 2006.
[31] R. McCraty, M. Atkinson, W.A. Tiller, G. Rein, and A.D. Watkins, "The Effects of Emotions on Short-Term Power Spectrum Analysis of Heart Rate Variability," *The Am. J. Cardiology,* vol. 76, no. 14, pp. 1089-1093, 1995.
[32] R.A. McFarland, "Relationship of Skin Temperature Changes to the Emotions Accompanying Music," *Applied Psychophysiology and Biofeedback,* vol. 10, pp. 255-267, 1985.
[33] J. Kim and E. André, "Emotion Recognition Based on Physiological Changes in Music Listening," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 30, no. 12, pp. 2067-2083, Dec. 2008.
[34] G. Chanel, J.J.M. Kierkels, M. Soleymani, and T. Pun, "Short-Term Emotion Assessment in a Recall Paradigm," *Int'l J. Human-Computer Studies,* vol. 67, no. 8, pp. 607-627, Aug. 2009.
[35] S.K. Sutton and R.J. Davidson, "Prefrontal Brain Asymmetry: A Biological Substrate of the Behavioral Approach and Inhibition Systems," *Psychological Science,* vol. 8, no. 3, pp. 204-210, 1997.
[36] R.J. Davidson, "Affective Neuroscience and Psychophysiology: Toward a Synthesis," *Psychophysiology,* vol. 40, no. 5, pp. 655-665, Sept. 2003.
[37] V.F. Pamplona, M.M. Oliveira, and G.V.G. Baranoski, "Photorealistic Models for Pupil Light Reflex and Iridal Pattern Deformation," *ACM Trans. Graphics,* vol. 28, no. 4, pp. 1-12, 2009.
[38] H. Bouma and L.C.J. Baghuis, "Hippus of the Pupil: Periods of Slow Oscillations of Unknown Origin," *Vision Research,* vol. 11, no. 11, pp. 1345-1351, 1971.
[39] R.J. Davidson, P. Ekman, C.D. Saron, J.A. Senulis, and W.V. Friesen, "Approach-Withdrawal and Cerebral Asymmetry: Emotional Expression and Brain Physiology I," *J. Personality and Social Psychology,* vol. 58, no. 2, pp. 330-341, 1990.
[40] C.-C. Chang and C.-J. Lin, "LIBSVM: A Library for Support Vector Machines," *Science,* vol. 2, pp. 1-39, 2001.
[41] J. Jiao and M. Pantic, "Implicit Image Tagging via Facial Information," *Proc. Second Int'l Workshop Social Signal Processing,* pp. 59-64, 2010.
[42] I. Patras and M. Pantic, "Particle Filtering with Factorized Likelihoods for Tracking Facial Features," *Proc. IEEE Sixth Int'l Conf. Automatic Face and Gesture Recognition,* pp. 97-102, May 2004.
[43] S. Petridis, H. Gunes, S. Kaltwang, and M. Pantic, "Static vs. Dynamic Modeling of Human Nonverbal Behavior from Multiple Cues and Modalities," *Proc. Int'l Conf. Multimodal Interfaces,* pp. 23-30, 2009.
[44] Y. Freund and R.E. Schapire, "A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting," *Proc. European Conf. Computational Learning Theory,* pp. 23-37, 1995.

Mohammad Soleymani received both the BSc and MSc degrees from the Department of Electrical and Computer Engineering, University of Tehran, Iran. He is now a doctoral student and research assistant in the Computer Vision and Multimedia Laboratory, Computer Science Department, University of Geneva, Switzerland. His research interests include affective computing and multimedia information retrieval. He is a member of the IEEE and the HUMAINE association.

Jeroen Lichtenauer received the MSc and PhD degrees in electrical engineering from Delft University of Technology, Delft, The Netherlands, in 2003 and 2009, respectively. He is a research associate with the Intelligent Behaviour Understanding Group, Department of Computing, Imperial College London, London, United Kingdom. His research interests include real-time signal processing, real-time computer vision.

Thierry Pun received the EE Engineering degree in 1979 and the PhD degree in 1982. He is a full professor and the head of the Computer Vision and Multimedia Laboratory, Computer Science Department, University of Geneva, Switzerland. He has authored or coauthored more than 300 full papers as well as eight patents in various aspects of multimedia processing; his current research interests lie in affective computing and multimodal interaction. He is a member of the IEEE.

Maja Pantic is a professor in affective and behavioural computing at Imperial College London, Department of Computing, United Kingdom, and at the University of Twente, Department of Computer Science, The Netherlands. She has received various awards for her work on automatic analysis of human behavior, including the European Research Council Starting Grant Fellowship and the Roger Needham Award 2011. She currently serves as the editor in chief of *Image and Vision Computing Journal* and as an associate editor for both the *IEEE Transactions on Systems, Man, and Cybernetics Part B* and the *IEEE Transactions on Pattern Analysis and Machine Intelligence*. She is a fellow of the IEEE.