DEAP: A Database for Emotion Analysis Using Physiological Signals

Sander Koelstra, *Student Member, IEEE*, Christian Mühl,
Mohammad Soleymani, *Student Member, IEEE*, Jong-Seok Lee, *Member, IEEE*,
Ashkan Yazdani, Touradj Ebrahimi, *Member, IEEE*, Thierry Pun, *Member, IEEE*,
Anton Nijholt, *Member, IEEE*, and Ioannis (Yiannis) Patras, *Senior Member, IEEE*

Abstract—We present a multimodal data set for the analysis of human affective states. The electroencephalogram (EEG) and peripheral physiological signals of 32 participants were recorded as each watched 40 one-minute long excerpts of music videos. Participants rated each video in terms of the levels of arousal, valence, like/dislike, dominance, and familiarity. For 22 of the 32 participants, frontal face video was also recorded. A novel method for stimuli selection is proposed using retrieval by affective tags from the last.fm website, video highlight detection, and an online assessment tool. An extensive analysis of the participants' ratings during the experiment is presented. Correlates between the EEG signal frequencies and the participants' ratings are investigated. Methods and results are presented for single-trial classification of arousal, valence, and like/dislike ratings using the modalities of EEG, peripheral physiological signals, and multimedia content analysis. Finally, decision fusion of the classification results from different modalities is performed. The data set is made publicly available and we encourage other researchers to use it for testing their own affective state estimation methods.

Index Terms—Emotion classification, EEG, physiological signals, signal processing, pattern classification, affective computing.

1 Introduction

Emotion is a psycho-physiological process triggered by conscious and/or unconscious perception of an object or situation and is often associated with mood, temperament, personality and disposition, and motivation. Emotions play an important role in human communication and can be expressed either verbally through emotional vocabulary or by expressing nonverbal cues such as intonation of voice, facial expressions, and gestures. Most of the contemporary human-computer interaction (HCI) systems are deficient in interpreting this information and suffer from the lack of emotional intelligence. In other words, they are unable to identify human emotional states and use this information in deciding upon proper actions to execute. The goal of affective computing is to fill this gap by detecting emotional cues occurring during human-computer interaction and synthesizing emotional responses.

Characterizing multimedia content with relevant, reliable, and discriminating tags is vital for multimedia information retrieval. Affective characteristics of multimedia are important features for describing multimedia content and can be presented by such emotional tags. Implicit affective tagging refers to the effortless generation of subjective and/or emotional tags. Implicit tagging of videos using affective information can help recommendation and retrieval systems to improve their performance [1], [2], [3]. The current data set is recorded with the goal of creating an adaptive music video recommendation system. In our proposed music video recommendation system, a user's bodily responses will be translated to emotions. The emotions of a user while watching music video clips will help the recommender system to first understand the user's taste and then to recommend a music clip which matches the user's current emotion.

The database presented explores the possibility of classifying emotion dimensions induced by showing music videos to different users. To the best of our knowledge, the responses to this stimuli (music video clips) have never been explored before, and the research in this field was mainly focused on images, music, or nonmusic video segments [4], [5]. In an adaptive music video recommender, an emotion recognizer trained by physiological responses to content of a similar nature, music videos are better able to fulfill its goal.

Various discrete categorizations of emotions have been proposed, such as the six basic emotions proposed by Ekman et al. [6] and the tree structure of emotions proposed

- *S. Koelstra and I. Patras are with the School of Electronic Engineering and Computer Science, Queen Mary University of London, Mile End Road, London E1 4NS, United Kingdom.
E-mail: {sander.koelstra, I.Patras}@eecs.qmul.ac.uk.*
- *C. Mühl and A. Nijholt are with the Human Media Interaction Group, University of Twente, Postbus 217, 7500 AE Enschede, The Netherlands. E-mail: muehlc@ewi.utwente.nl, a.nijholt@utwente.nl.*
- *M. Soleymani and T. Pun are with the Computer Science Department, University of Geneva, Battelle Campus, Building A 7, route de Drize, Carouge (Geneva) CH-1227, Switzerland.
E-mail: {mohammad.soleymani, Thierry.Pun}@unige.ch.*
- *J.-S. Lee is with the School of Integrated Technology, Yonsei University, Korea. E-mail: jong-seok.lee@yonsei.ac.kr.*
- *A. Yazdani and T. Ebrahimi are with the Multimedia Signal Processing Group, Institute of Electrical Engineering (IEL), Ecole Polytechnique Fédérale de Lausanne (EPFL), Station 11, Lausanne CH-1015, Switzerland. E-mail: {ashkan.yazdani, touradj.ebrahimi}@epfl.ch.*

*Manuscript received 1 Nov. 2010; revised 1 Apr. 2011; accepted 18 Apr. 2011; published online 6 June 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-2010-11-0104.
Digital Object Identifier no. 10.1109/T-AFFC.2011.15.*

1949-3045/12/$31.00 © 2012 IEEE Published by the IEEE Computer Society

by Parrott [7]. Dimensional scales of emotion have also been proposed, such as Plutchik's emotion wheel [8] and the valence-arousal scale by Russell [9]. In this work, we use Russell's valence-arousal scale, widely used in research on affect, to quantitatively describe emotions. In this scale, each emotional state can be placed on a 2D plane with arousal and valence as the horizontal and vertical axes. While arousal and valence explain most of the variation in emotional states, a third dimension of dominance can also be included in the model [9]. Arousal can range from inactive (e.g., uninterested, bored) to active (e.g., alert, excited), whereas valence ranges from unpleasant (e.g., sad, stressed) to pleasant (e.g., happy, elated). Dominance ranges from a helpless and weak feeling (without control) to an empowered feeling (in control of everything). For self-assessment along these scales, we use the well-known self-assessment manikins (SAM) [10].

Emotion assessment is often carried out through analysis of users' emotional expressions and/or physiological signals. Emotional expressions refer to any observable verbal and nonverbal behavior that communicates emotion. So far, most of the studies on emotion assessment have focused on the analysis of facial expressions and speech to determine a person's emotional state. Physiological signals are also known to include emotional information that can be used for emotion assessment, but they have received less attention. They comprise the signals originating from the central nervous system (CNS) and the peripheral nervous system (PNS).

Recent advances in emotion recognition have motivated the creation of novel databases containing emotional expressions in different modalities. These databases mostly cover speech, visual data, or audiovisual data (e.g., [11], [12], [13], [14], [15]). The visual modality includes facial expressions and/or body gestures. The audio modality covers posed or genuine emotional speech in different languages. Many of the existing visual databases include only posed or deliberately expressed emotions.

Healey [16], [17] recorded one of the first affective physiological data sets. She recorded 24 participants driving around the Boston area and annotated the data set by the drivers' stress level. Responses of 17 of the 24 participants are publicly available.[1] Her recordings include electrocardiogram (ECG), galvanic skin response (GSR) recorded from hands and feet, and electromyogram (EMG) from the right trapezius muscle and respiration patterns.

To the best of our knowledge, the only publicly available multimodal emotional databases which include both physiological responses and facial expressions are the enterface 2005 emotional database and MAHNOB HCI [4], [5]. The first one was recorded by Savran et al. [5]. This database includes two sets. The first set has electroencephalogram (EEG), peripheral physiological signals, functional near infrared spectroscopy (fNIRS), and facial videos from five male participants. The second data set only has fNIRS and facial videos from 16 participants of both genders. Both databases recorded spontaneous responses to emotional images from the international affective picture system (IAPS) [18]. An extensive review of affective audiovisual databases can be found in [13] and [19]. The MAHNOB HCI database [4] consists of two experiments. The responses including EEG, physiological signals, eye gaze, audio, and facial expressions of 30 people were recorded. In the first experiment, participants watched 20 emotional videos extracted from movies and online repositories. The second experiment was a tag agreement experiment in which images and short videos with human actions were shown the participants first without a tag and then with a displayed tag. The tags were either correct or incorrect and participants' agreement with the displayed tag was assessed.

There has been a large number of published works in the domain of emotion recognition from physiological signals [16], [20], [21], [22], [23], [24]. Of these studies, only a few achieved notable results using video stimuli. Lisetti and Nasoz [23] used physiological responses to recognize emotions in response to movie scenes. The movie scenes were selected to elicit six emotions, namely, sadness, amusement, fear, anger, frustration, and surprise. They achieved a high recognition rate of 84 percent for the recognition of these six emotions. However, the classification was based on the analysis of the signals in response to preselected segments in the shown video known to be related to highly emotional events.

Some efforts have been made toward implicit affective tagging of multimedia content. Kierkels et al. [25] proposed a method for personalized affective tagging of multimedia using peripheral physiological signals. Valence and arousal levels of participants' emotions when watching videos were computed from physiological responses using linear regression [26]. Quantized arousal and valence levels for a clip were then mapped to emotion labels. This mapping enabled the retrieval of video clips based on keyword queries. So far, this novel method achieved low precision.

Yazdani et al. [27] proposed using a brain-computer interface (BCI) based on P300 evoked potentials to emotionally tag videos with one of the six Ekman basic emotions [28]. Their system was trained with eight participants and then tested on four others. They achieved a high accuracy on selecting tags. However, in their proposed system, a BCI only replaces the interface for explicit expression of emotional tags, i.e., the method does not implicitly tag a multimedia item using the participant's behavioral and psycho-physiological responses.

In addition to implicit tagging using behavioral cues, multiple studies used multimedia content analysis (MCA) for automated affective tagging of videos. Hanjalic and Xu [29] introduced "personalized content delivery" as a valuable tool in affective indexing and retrieval systems. In order to represent affect in video, they first selected video—and audio—content features based on their relation to the valence-arousal space. Then, arising emotions were estimated in this space by combining these features. While valence arousal could be used separately for indexing, they combined these values by following their temporal pattern. This allowed for determining an affect curve, shown to be useful for extracting video highlights in a movie or sports video.

Wang and Cheong [30] used audio and video features to classify basic emotions elicited by movie scenes. Audio was classified into music, speech, and environment signals and these were treated separately to shape an aural affective

[1] http://www.physionet.org/pn3/drivedb/.

feature vector. The aural affective vector of each scene was fused with video-based features such as key lighting and visual excitement to form a scene feature vector. Finally, using the scene feature vectors, movie scenes were classified and labeled with emotions.

Soleymani et al. [31] proposed a scene affective characterization using a Bayesian framework. Arousal and valence of each shot were first determined using linear regression. Then, arousal and valence values in addition to content features of each scene were used to classify every scene into three classes, namely, calm, excited positive, and excited negative. The Bayesian framework was able to incorporate the movie genre and the predicted emotion from the last scene or temporal information to improve the classification accuracy.

There are also various studies on music affective characterization from acoustic features [32], [33], [34]. Rhythm, tempo, Mel-frequency cepstral coefficients (MFCC), pitch, and zero crossing rate are among common features which have been used to characterize affect in music.

A pilot study for the current work was presented in [35]. In that study, six participants' EEG and physiological signals were recorded as each watched 20 music videos. The participants rated arousal and valence levels and the EEG and physiological signals for each video were classified into low/high arousal/valence classes.

In the current work, music video clips are used as the visual stimuli to elicit different emotions. To this end, a relatively large set of music video clips was gathered using a novel stimuli selection method. A subjective test was then performed to select the most appropriate test material. For each video, a one-minute highlight was selected automatically. Thirty-two participants took part in the experiment and their EEG and peripheral physiological signals were recorded as they watched the 40 selected music videos. Participants rated each video in terms of arousal, valence, like/dislike, dominance, and familiarity. For 22 participants, frontal face video was also recorded.

This paper aims at introducing this publicly available[2] database. The database contains all recorded signal data, frontal face video for a subset of the participants, and subjective ratings from the participants. Also included is the subjective ratings from the initial online subjective annotation and the list of 120 videos used. Due to licensing issues, we are not able to include the actual videos, but YouTube links are included. Table 1 gives an overview of the database contents.

To the best of our knowledge, this database has the highest number of participants in publicly available databases for analysis of spontaneous emotions from physiological signals. In addition, it is the only database that uses music videos as emotional stimuli.

We present an extensive statistical analysis of the participant's ratings and of the correlates between the EEG signals and the ratings. Preliminary single trial classification results of EEG, peripheral physiological signals, and MCA are presented and compared. Finally, a fusion algorithm is utilized to combine the results of each modality and arrive at a more robust decision.

TABLE 1
Database Content Summary

Online subjective annotation	
Number of videos	120
Video duration	1 minute affective highlight (section 2.2)
Selection method	60 via last.fm affective tags, 60 manually selected
No. of ratings per video	14 - 16
Rating scales	Arousal Valence Dominance
Rating values	Discrete scale of 1 - 9
Physiological Experiment	
Number of participants	32
Number of videos	40
Selection method	Subset of online annotated videos with clearest responses (see section 2.3)
Rating scales	Arousal Valence Dominance Liking *(how much do you like the video?)* Familiarity *(how well do you know the video?)*
Rating values	Familiarity: discrete scale of 1 - 5 Others: continuous scale of 1 - 9
Recorded signals	32-channel 512Hz EEG Peripheral physiological signals Face video (for 22 participants)

The layout of the paper is as follows. In Section 2, the stimuli selection procedure is described in detail. The experiment setup is covered in Section 3. Section 4 provides a statistical analysis of the ratings given by participants during the experiment and a validation of our stimuli selection method. In Section 5, correlates between the EEG frequencies and the participants' ratings are presented. The method and results of single-trial classification are given in Section 6. The conclusion of this work follows in Section 7.

2 Stimuli Selection

The stimuli used in the experiment were selected in several steps. First, we selected 120 initial stimuli, half of which were chosen semi-automatically and the rest manually. Then, a one-minute highlight part was determined for each stimulus. Finally, through a web-based subjective assessment experiment, 40 final stimuli were selected. Each of these steps is explained below.

2.1 Initial Stimuli Selection

Eliciting emotional reactions from test participants is a difficult task and selecting the most effective stimulus materials is crucial. We propose here a semi-automated method for stimulus selection, with the goal of minimizing the bias arising from the manual stimuli selection.

Sixty of the 120 initially selected stimuli were selected using the Last.fm[3] music enthusiast website. Last.fm allows users to track their music listening habits and receive recommendations for new music and events. Additionally, it allows the users to assign tags to individual songs, thus

[2] http://www.eecs.qmul.ac.uk/mmv/data sets/deap/.

[3] http://www.last.fm.

creating a folksonomy of tags. Many of the tags carry emotional meanings, such as "depressing" or "aggressive." Last.fm offers an API, allowing one to retrieve tags and tagged songs.

A list of emotional keywords was taken from [7] and expanded to include inflections and synonyms, yielding 304 keywords. Next, for each keyword, corresponding tags were found in the Last.fm database. For each found affective tag, the 10 songs most often labeled with this tag were selected. This resulted in a total of 1,084 songs.

The valence-arousal space can be subdivided into four quadrants, namely, low arousal/low valence (LALV), low arousal/high valence (LAHV), high arousal/low valence (HALV), and high arousal/high valence (HAHV). In order to ensure diversity of induced emotions, from the 1,084 songs, 15 were selected manually for each quadrant according to the following criteria.

Does the tag accurately reflect the emotional content? Examples of songs subjectively rejected according to this criterion include songs that are tagged merely because the song title or artist name corresponds to the tag. Also, in some cases the lyrics may correspond to the tag, but the actual emotional content of the song is entirely different (e.g., happy songs about sad topics).

Is a music video available for the song? Music videos for the songs were automatically retrieved from YouTube, corrected manually where necessary. However, many songs do not have a music video.

Is the song appropriate for use in the experiment? Since our test participants were mostly European students, we selected those songs most likely to elicit emotions for this target demographic. Therefore, mainly, European or North American artists were selected.

In addition to the songs selected using the method described above, 60 stimulus videos were selected manually, with 15 videos selected for each of the quadrants in the arousal/valence space. The goal here was to select those videos expected to induce the most clear emotional reactions for each of the quadrants. The combination of manual selection and selection using affective tags produced a list of 120 candidate stimulus videos.

2.2 Detection of One-Minute Highlights

For each of the 120 initially selected music videos, a one-minute segment for use in the experiment was extracted. In order to extract a segment with maximum emotional content, an affective highlighting algorithm is proposed.

Soleymani et al. [31] used a linear regression method to calculate arousal for each shot of in movies. In their method, the arousal and valence of shots were computed using a linear regression on the content-based features. Informative features for arousal estimation include loudness and energy of the audio signals, motion component, visual excitement, and shot duration. The same approach was used to compute valence. There are other content features such as color variance and key lighting that have been shown to be correlated with valence [30]. The detailed description of the content features used in this work is given in Section 6.2.

In order to find the best weights for arousal and valence estimation using regression, the regressors were trained on all shots in 21 annotated movies in the data set presented in [31]. The linear weights were computed by means of a relevance vector machine (RVM) from the RVM toolbox provided by Tipping [36]. The RVM is able to reject uninformative features during its training; hence, no further feature selection was used for arousal and valence determination.

The music videos were then segmented into one-minute segments with 55 seconds overlap between segments. Content features were extracted and provided the input for the regressors. The emotional highlight score of the ith segment e_i was computed using the following equation:

$$e_i = \sqrt{a_i^2 + v_i^2}. \tag{1}$$

The arousal, a_i, and valence, v_i, were centered. Therefore, a smaller emotional highlight score (e_i) is closer to the neutral state. For each video, the one-minute-long segment with the highest emotional highlight score was chosen to be extracted for the experiment. For a few clips, the automatic affective highlight detection was manually overridden. This was done only for songs with segments that are particularly characteristic of the song, well known to the public, and most likely to elicit emotional reactions. In these cases, the one-minute highlight was selected so that these segments were included.

Given the 120 one-minute music video segments, the final selection of 40 videos used in the experiment was made on the basis of subjective ratings by volunteers, as described in the next section.

2.3 Online Subjective Annotation

From the initial collection of 120 stimulus videos, the final 40 test video clips were chosen by using a web-based subjective emotion assessment interface. Participants watched music videos and rated them on a discrete 9-point scale for valence, arousal, and dominance. A screenshot of the interface is shown in Fig. 1. Each participant watched as many videos as he/she wanted and was able to end the rating at any time. The order of the clips was randomized, but preference was given to the clips rated by the least number of participants. This ensured a similar number of ratings for each video (14-16 assessments per video were collected). It was ensured that participants never saw the same video twice.

After all of the 120 videos were rated by at least 14 volunteers each, the final 40 videos for use in the experiment were selected. To maximize the strength of elicited emotions, we selected those videos that had the strongest volunteer ratings and at the same time a small variation. To this end, for each video x we calculated a normalized arousal and valence score by taking the mean rating divided by the standard deviation (μ_x/σ_x).

Then, for each quadrant in the normalized valence-arousal space, we selected the 10 videos that lie closest to the extreme corner of the quadrant. Fig. 2 shows the score for the ratings of each video and the selected videos highlighted in green. The video whose rating was closest to the extreme corner of each quadrant is mentioned explicitly. Of the 40 selected videos, 17 were selected via Last.fm affective tags, indicating that useful stimuli can be selected via this method.

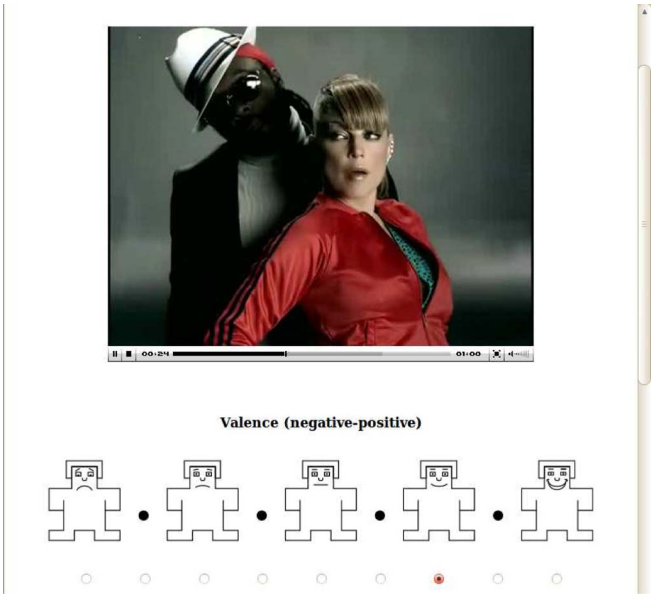


Fig. 1. Screenshot of the web interface for subjective emotion assessment.

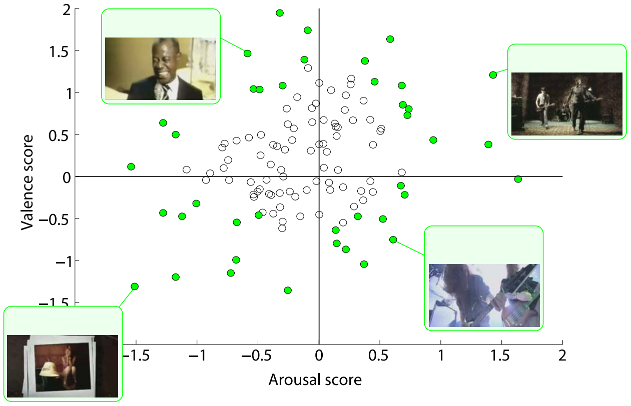


Fig. 2. μ_x/σ_x value for the ratings of each video in the online assessment. Videos selected for use in the experiment are highlighted in green. For each quadrant, the most extreme video is detailed with the song title and a screenshot from the video.

3 Experiment Setup

3.1 Materials and Setup

The experiments were performed in two laboratory environments with controlled illumination. EEG and peripheral physiological signals were recorded using a Biosemi ActiveTwo system[4] on a dedicated recording PC (Pentium 4, 3.2 GHz). Stimuli were presented using a dedicated stimulus PC (Pentium 4, 3.2 GHz) that sent synchronization markers directly to the recording PC. For presentation of the stimuli and recording the users' ratings, the "Presentation" software by Neurobehavioral systems[5] was used. The music videos were presented on a 17-inch screen (1,280 × 1,024, 60 Hz) and, in order to minimize eye movements, all video stimuli were displayed at 800 × 600 resolution, filling approximately 2/3 of the screen. Subjects were seated approximately 1 meter from the screen. Stereo Philips speakers were used and the music volume was set at a relatively loud level; however, each participant was asked before the experiment whether the volume was comfortable and it was adjusted when necessary.

EEG was recorded at a sampling rate of 512 Hz using 32 active AgCl electrodes (placed according to the international 10-20 system). Thirteen peripheral physiological signals (which will be further discussed in section 6.1) were also recorded. Additionally, for the first 22 of the 32 participants, frontal face video was recorded in DV quality using a Sony DCR-HC27E consumer-grade camcorder. The face video was not used in the experiments in this paper, but is made publicly available along with the rest of the data. Fig. 3 illustrates the electrode placement for acquisition of peripheral physiological signals.

3.2 Experiment Protocol

Thirty-two healthy participants (50 percent females), aged between 19 and 37 (mean age 26.9), participated in the experiment. Prior to the experiment, each participant signed a consent form and filled out a questionnaire. Next, they were given a set of instructions to read informing them of the experiment protocol and the meaning of different scales used for self-assessment. An experimenter was also present there to answer any questions. When the instructions were clear to the participant, he/she was led into the experiment room. After the sensors were placed and their signals checked, the participants performed a practice trial to familiarize themselves with the system. In this unrecorded trial, a short video was shown, followed by a self-assessment by the participant. Next, the experimenter started the physiological signals recording and left the room, after which the participant started the experiment by pressing a key on the keyboard.

The experiment started with a two-minute baseline recording, during which a fixation cross was displayed to the participant (who was asked to relax during this period). Then, the 40 videos were presented in 40 trials, each consisting of the following steps:

1. A 2-second screen displaying the current trial number to inform the participants of their progress.
2. A 5-second baseline recording (fixation cross).
3. The 1-minute display of the music video.

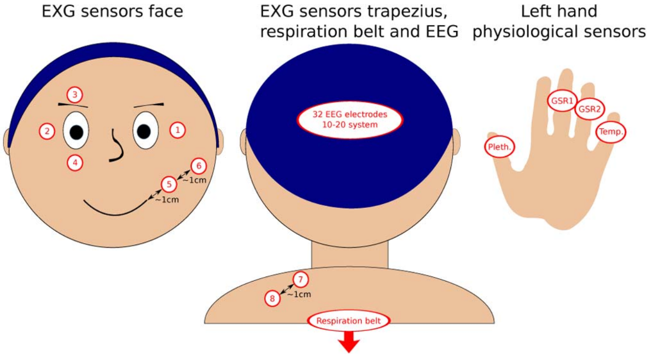

EXG sensors face
EXG sensors trapezius, respiration belt and EEG
Left hand physiological sensors
1
2
3
4
5
6
~1cm
~1cm
32 EEG electrodes 10-20 system
7
8
~1cm
Respiration belt
GSR1
GSR2
Temp.
Pleth.

Fig. 3. Placement of peripheral physiological sensors. Four electrodes were used to record EOG and four for EMG (zygomaticus major and trapezius muscles). In addition, GSR, blood volume pressure (BVP), temperature, and respiration were measured.

[4] http://www.biosemi.com.
[5] http://www.neurobs.com.

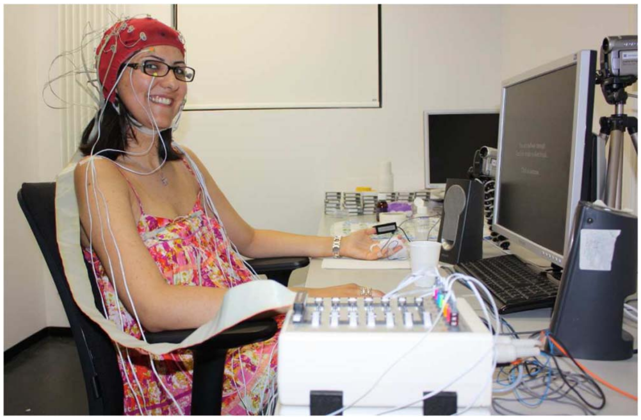


Fig. 4. A participant shortly before the experiment.

4. Self-assessment for arousal, valence, liking, and dominance.

After 20 trials, the participants took a short break. During the break, they were offered some cookies and noncaffeinated, nonalcoholic beverages. The experimenter then checked the quality of the signals and the electrodes placement and the participants were asked to continue the second half of the test. Fig. 4 shows a participant shortly before the start of the experiment.

3.3 Participant Self-Assessment

At the end of each trial, participants performed a self-assessment of their levels of arousal, valence, liking, and dominance. Self-assessment manikins [37] were used to visualize the scales (see Fig. 5). For the liking scale, thumbs down/thumbs up symbols were used. The manikins were displayed in the middle of the screen with the numbers 1-9 printed below. Participants moved the mouse strictly horizontally just below the numbers and clicked to indicate their self-assessment level. Participants were informed they could click anywhere directly below or in-between the numbers, making the self-assessment a continuous scale.

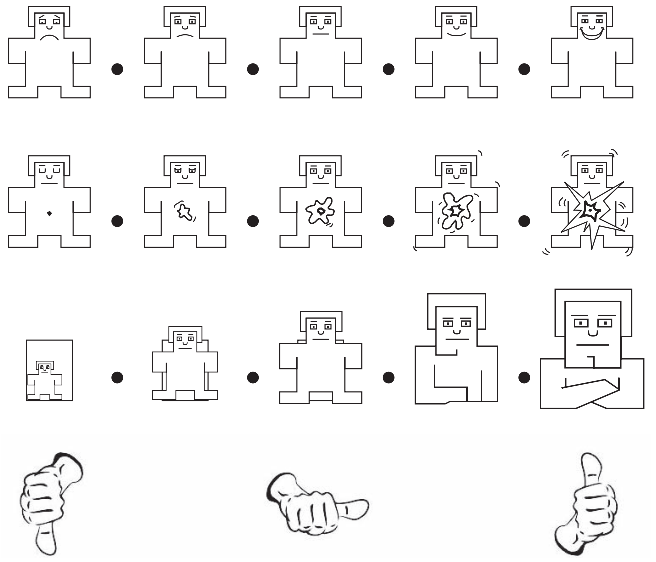


Fig. 5. Images used for self-assessment. From the top: Valence SAM, arousal SAM, dominance SAM, and liking.

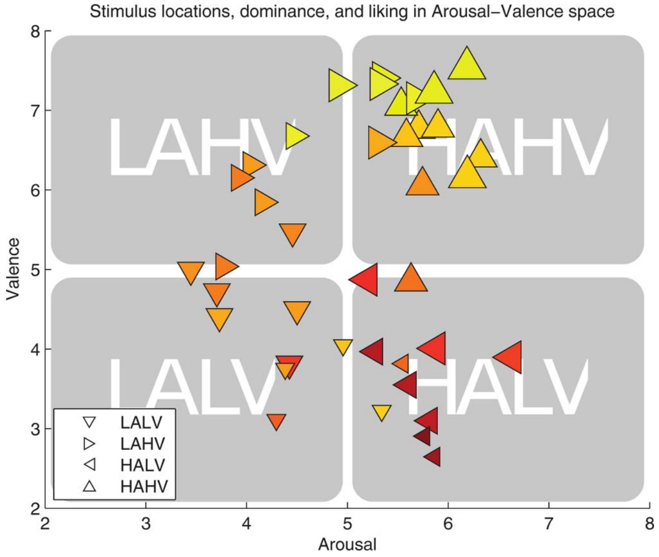


Fig. 6. The mean locations of the stimuli on the arousal-valence plane (AV plane) for the four conditions (LALV, HALV, LAHV, and HAHV). Liking is encoded by color: Dark red is low liking and bright yellow is high liking. Dominance is encoded by symbol size: Small symbols stand for low dominance and big for high dominance.

The valence scale ranges from unhappy or sad to happy or joyful. The arousal scale ranges from calm or bored to stimulated or excited. The dominance scale ranges from submissive (or "without control") to dominant (or "in control, empowered"). A fourth scale asks for participants' personal liking of the video. This last scale should not be confused with the valence scale. This measure inquires about the participants' tastes, not their feelings. For example, it is possible to like videos that make one feel sad or angry. Finally, after the experiment, participants were asked to rate their familiarity with each of the songs on a scale of 1 ("Never heard it before the experiment") to 5 ("Knew the song very well").

4 Analysis of Subjective Ratings

In this section, we describe the effect the affective stimulation had on the subjective ratings obtained from the participants. First, we will provide descriptive statistics for the recorded ratings of liking, valence, arousal, dominance, and familiarity. Second, we will discuss the covariation of the different ratings with each other.

Stimuli were selected to induce emotions in the four quadrants of the valence-arousal space (LALV, HALV, LAHV, and HAHV). The stimuli from these four affect elicitation conditions generally resulted in the elicitation of the target emotion aimed for when the stimuli were selected, ensuring that large parts of the arousal-valence plane are covered (see Fig. 6). Wilcoxon signed-rank tests showed that low and high arousal stimuli induced different valence ratings ($p < 0.0001$ and $p < 0.00001$). Similarly, low and high valenced stimuli induced different arousal ratings ($p < 0.001$ and $p < 0.0001$).

The emotion elicitation specifically worked well for the high arousing conditions, yielding relative extreme valence ratings for the respective stimuli. The stimuli in the low arousing conditions were less successful in the elicitation of strong valence responses. Furthermore, some stimuli of the

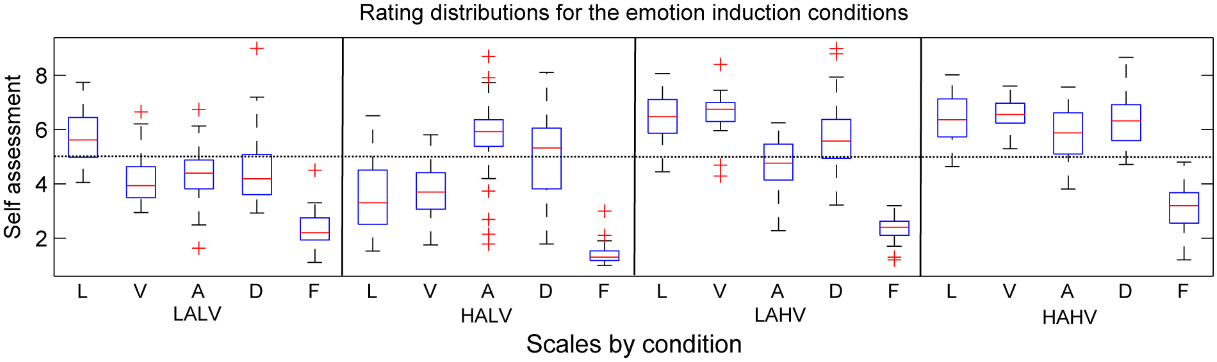


Fig. 7. The distribution of the participants' subjective ratings per scale (L—general rating, V—valence, A—arousal, D—dominance, and F—familiarity) for the four affect elicitation conditions (LALV, HALV, LAHV, and HAHV).

LAHV condition induced higher arousal than expected on the basis of the online study. Interestingly, this results in a C-shape of the stimuli on the valence-arousal plane also observed in the well-validated ratings for the IAPS [18] and the international affective digital sounds system (IADS) [38], indicating the general difficulty in inducing emotions with strong valence, but low arousal. The distribution of the individual ratings per conditions (see Fig. 7) shows a large variance within conditions, resulting from between-stimulus and participant variations, possibly associated with stimulus characteristics or interindividual differences in music taste, general mood, or scale interpretation. However, the significant differences between the conditions in terms of the ratings of valence and arousal reflect the successful elicitation of the targeted affective states (see Table 2).

The distribution of ratings for the different scales and conditions suggests a complex relationship between ratings. We explored the mean intercorrelation of the different scales over participants (see Table 3), as they might be indicative of possible confounds or unwanted effects of habituation or fatigue. We observed high positive correlations between liking and valence, and between dominance and valence. Seemingly, without implying any causality, people liked music which gave them a positive feeling and/or a feeling of empowerment. Medium positive correlations were observed between arousal and dominance, and between arousal and liking. Familiarity correlated moderately positive with liking and valence. As already observed above, the scales of valence and arousal are not independent, but their positive correlation is rather low, suggesting that participants were able to differentiate between these two important concepts. Stimulus order had only a small effect on liking and dominance ratings and no significant relationship with the other ratings, suggesting that effects of habituation and fatigue were kept to an acceptable minimum.

In summary, the affect elicitation was in general successful, though the low valence conditions were partially biased by moderate valence responses and higher arousal. High scale intercorrelations observed are limited to the scale of valence with those of liking and dominance, and might be expected in the context of musical emotions. The rest of the scale intercorrelations are small or medium in strength, indicating that the scale concepts were well distinguished by the participants.

TABLE 2
The Mean Values (and Standard Deviations) of the Different Ratings of Liking (1-9), Valence (1-9), Arousal (1-9), Dominance (1-9), and Familiarity (1-5) for Each Affect Elicitation Condition

Cond.	Liking	Valence	Arousal	Dom.	Fam.
LALV	5.7 (1.0)	4.2 (0.9)	4.3 (1.1)	4.5 (1.4)	2.4 (0.4)
HALV	3.6 (1.3)	3.7 (1.0)	5.7 (1.5)	5.0 (1.6)	1.4 (0.6)
LAHV	6.4 (0.9)	6.6 (0.8)	4.7 (1.0)	5.7 (1.3)	2.4 (0.4)
HAHV	6.4 (0.9)	6.6 (0.6)	5.9 (0.9)	6.3 (1.0)	3.1 (0.4)

TABLE 3
The Means of the Subject-Wise Intercorrelations between the Scales of Valence, Arousal, Liking, Dominance, Familiarity and the Order of the Presentation (i.e., Time) for All 40 Stimuli

Scale	Liking	Valence	Arousal	Dom.	Fam.	Order
Liking	1	0.62*	0.29*	0.31*	0.30*	0.03*
Valence		1	0.18*	0.51*	0.25*	0.02
Arousal			1	0.28*	0.06*	0.00
Dom.				1	0.09*	0.04*
Fam.					1	-
Order						1

Significant correlations ($p < 0.05$) according to Fisher's method are indicated by stars.

5 Correlates of EEG and Ratings

For the investigation of the correlates of the subjective ratings with the EEG signals, the EEG data were common average referenced, downsampled to 256 Hz, and high-pass filtered with a 2 Hz cutoff frequency using the EEGlab[6] toolbox. We removed eye artifacts with a blind source separation technique.[7] Then, the signals from the last 30 seconds of each trial (video) were extracted for further analysis. To correct for stimulus-unrelated variations in power over time, the EEG signal from the 5 seconds before each video was extracted as baseline.

The frequency power of trials and baselines between 3 and 47 Hz was extracted with Welch's method with windows of 256 samples. The baseline power was then subtracted from the trial power, yielding the change of power relative to the pre-stimulus period. These changes of power were averaged over the frequency bands of theta (3-7 Hz), alpha (8-13 Hz), beta (14-29 Hz), and gamma (30-47 Hz). For the correlation statistic, we computed the Spearman correlated coefficients

[6] http://sccn.ucsd.edu/eeglab/.
[7] http://www.cs.tut.fi/~gomezher/projects/eeg/aar.htm.

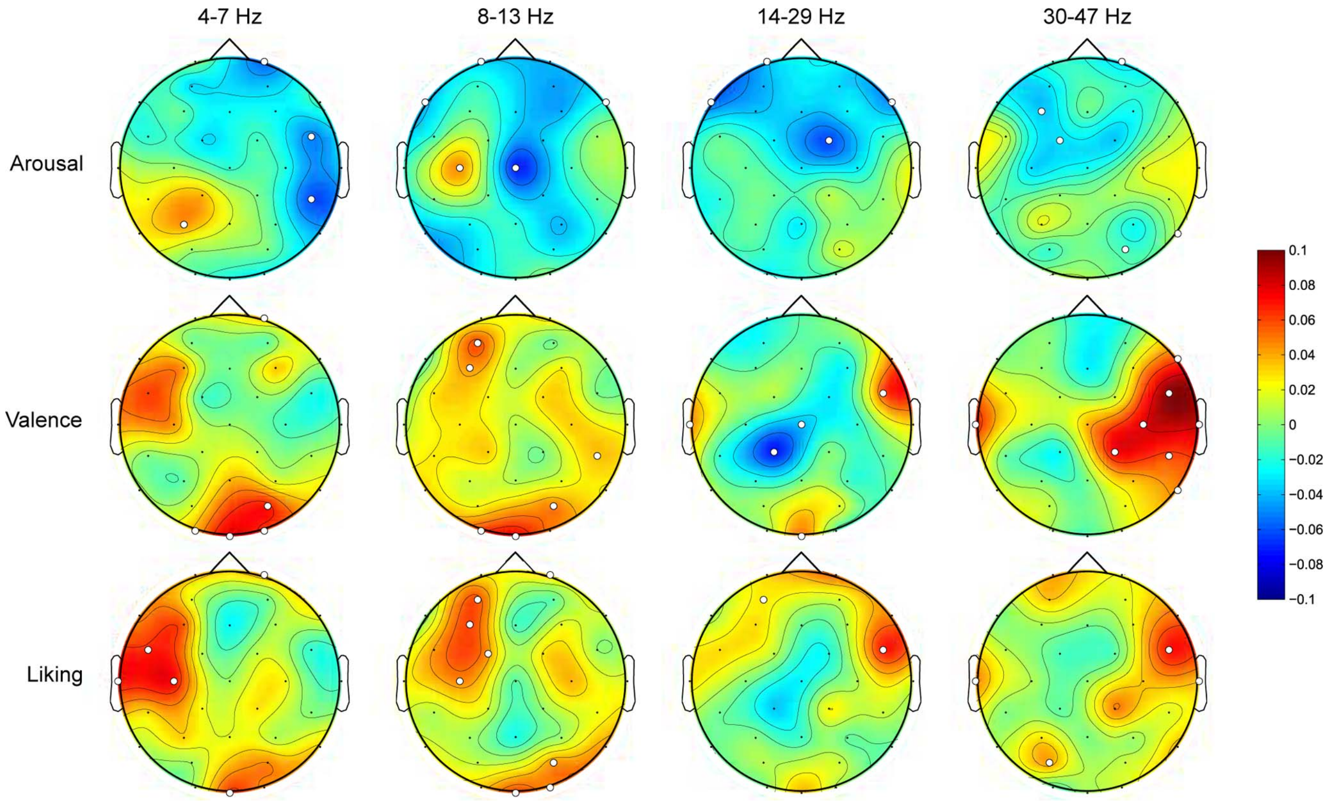


Fig. 8. The mean correlations (over all participants) of the valence, arousal, and general ratings with the power in the broad frequency bands of theta (4-7 Hz), alpha (8-13 Hz), beta (14-29 Hz), and gamma (30-47 Hz). The highlighted sensors correlate significantly ($p < 0.05$) with the ratings.

between the power changes and the subjective ratings, and computed the p-values for the left- (positive) and right-tailed (negative) correlation tests. This was done for each participant separately and, assuming independence [39], the 32 resulting p-values per correlation direction (positive/negative), frequency band, and electrode were then combined to one p-value via Fisher's method [40].

Fig. 8 shows the (average) correlations with significantly ($p < 0.05$) correlating electrodes highlighted. Below, we will report and discuss only those effects that were significant with $p < 0.01$. A comprehensive list of the effects can be found in Table 4.

For arousal we found negative correlations in the theta, alpha, and gamma bands. The central alpha power decrease for higher arousal matches the findings from our earlier pilot study [35] and an inverse relationship between alpha power and the general level of arousal has been reported before [41], [42].

Valence showed the strongest correlations with EEG signals and correlates were found in all analyzed frequency bands. In the low frequencies, theta and alpha, an increase of valence led to an increase of power. This is consistent with the findings in the pilot study. The location of these effects over occipital regions, thus over visual cortices, might indicate a relative deactivation, or top-down inhibition, of these due to participants focusing on the pleasurable sound [43]. For the beta frequency band we found a central decrease, also observed in the pilot, and an occipital and

TABLE 4
The Electrodes for Which the Correlations with the Scale Were Significant ($* = p < 0.01, ** = p < 0.001$)

	Theta				Alpha				Beta				Gamma			
	Elec.	$\bar{R}$	R^-	R^+	Elec.	$\bar{R}$	R^-	R^+	Elec.	$\bar{R}$	R^-	R^+	Elec.	$\bar{R}$	R^-	R^+
Arousal	CP6*	-0.06	-0.47	0.25	Cz*	-0.07	-0.45	0.23	FC2*	-0.06	-0.40	0.28				
Valence	Oz**	0.08	-0.23	0.39	PO4*	0.05	-0.26	0.49	CP1**	-0.07	-0.49	0.24	T7**	0.07	-0.33	0.51
	PO4*	0.05	-0.26	0.49					Oz*	0.05	-0.24	0.48	CP6*	0.06	-0.26	0.43
									FC6*	0.06	-0.52	0.49	CP2*	0.08	-0.21	0.49
									Cz*	-0.04	-0.64	0.30	C4**	0.08	-0.31	0.51
													T8**	0.08	-0.26	0.50
													FC6**	0.10	-0.29	0.52
													F8*	0.06	-0.35	0.52
Liking	C3*	0.08	-0.35	0.31	AF3*	0.06	-0.27	0.42	FC6*	0.07	-0.40	0.48	T8*	0.04	-0.33	0.49
					F3*	0.06	-0.42	0.45								

Also shown is the mean of the subject-wise correlations ($\bar{R}$), the most negative (R^-), and the most positive correlation (R^+).

right temporal increase of power. Increased beta power over right temporal sites was associated with positive emotional self-induction and external stimulation in [44]. Similarly, Onton and Makeig [45] have reported a positive correlation of valence and high-frequency power, including beta and gamma bands, emanating from anterior temporal cerebral sources. Correspondingly, we observed a highly significant increase of left and especially right temporal gamma power. However, it should be mentioned that the EMG (muscle) activity is also prominent in the high frequencies, especially over anterior and temporal electrodes [46].

The liking correlates were found in all analyzed frequency bands. For theta and alpha power, we observed increases over left fronto-central cortices. Liking might be associated with an approach motivation. However, the observation of an increase of left alpha power for a higher liking conflicts with findings of a left frontal activation, leading to lower alpha over this region, often reported for emotions associated with approach motivations [47]. This contradiction might be reconciled when taking into account that it is quite possible that some disliked pieces induced an angry feeling (due to having to listen to them or simply due to the content of the lyrics), which is also related to an approach motivation, and might hence result in a left-ward decrease of alpha. The right temporal increases found in the beta and gamma bands are similar to those observed for valence, and the same caution should be applied. In general, the distribution of valence and liking correlations shown in Fig. 8 seem very similar, which might be a result of the high intercorrelations of the scales discussed above.

Summarizing, we can state that the correlations observed partially concur with observations made in the pilot study and in other studies exploring the neuro-physiological correlates of affective states. They might, therefore, be taken as valid indicators of emotional states in the context of multimodal musical stimulation. However, the mean correlations are seldom larger than ± 0.1, which might be due to high interparticipant variability in terms of brain activations, as individual correlations between ± 0.5 were observed for a given scale correlation at the same electrode/frequency combination. The presence of this high interparticipant variability justifies a participant-specific classification approach, as we employ it, rather than a single classifier for all participants.

6 Single Trial Classification

In this section, we present the methodology and results of single-trial classification of the videos. Three different modalities were used for classification, namely, EEG signals, peripheral physiological signals, and MCA. Conditions for all modalities were kept equal and only the feature extraction step varies.

Three different binary classification problems were posed: the classification of low/high arousal, low/high valence, and low/high liking. To this end, the participants' ratings during the experiment are used as the ground truth. The ratings for each of these scales are thresholded into two classes (low and high). On the 9-point rating scales, the threshold was simply placed in the middle. Note that for some subjects and scales, this leads to unbalanced classes. To give an indication of how unbalanced the classes are, the mean and standard deviation (over participants) of the percentage of videos belonging to the high class per rating scale are as follows: arousal 59 percent (15 percent), valence 57 percent (9 percent), and liking 67 percent (12 percent).

In light of this issue, in order to reliably report results, we report the F1-core, which is commonly employed in information retrieval and takes the class balance into account, contrary to the mere classification rate. In addition, we use a naive Bayes classifier, a simple and generalizable classifier which is able to deal with unbalanced classes in small training sets.

First, the features for the given modality are extracted for each trial (video). Then, for each participant, the F1 measure was used to evaluate the performance of emotion classification in a leave-one-out cross validation scheme. At each step of the cross validation, one video was used as the test set and the rest were used as training set. We use Fisher's linear discriminant J for feature selection:

$$J(f) = \frac{|\mu_1 - \mu_2|}{\sigma_1^2 + \sigma_2^2}, \tag{2}$$

where μ and σ are the mean and standard deviation for feature f. We calculate this criterion for each feature and then apply a threshold to select the maximally discriminating ones. This threshold was empirically determined at 0.3.

A Gaussian naive Bayes classifier was used to classify the test set as low/high arousal, valence, or liking.

The naive Bayes classifier G assumes independence of the features and is given by

$$G(f_1, .., f_n) = \operatorname*{argmax}_c p(C = c) \prod_{i=1}^{n} p(F_i = f_i | C = c), \tag{3}$$

where F is the set of features and C the classes. $p(F_i = f_i | C = c)$ is estimated by assuming Gaussian distributions of the features and modeling these from the training set.

The following section explains the feature extraction steps for the EEG and peripheral physiological signals. Section 6.2 presents the features used in MCA classification. In Section 6.3, we explain the method used for decision fusion of the results. Finally, Section 6.4 presents the classification results.

6.1 EEG and Peripheral Physiological Features

Most of the current theories of emotion [48], [49] agree that physiological activity is an important component of an emotion. For instance, several studies have demonstrated the existence of specific physiological patterns associated with basic emotions [6].

The following peripheral nervous system signals were recorded: GSR, respiration amplitude, skin temperature, electrocardiogram, blood volume by plethysmograph, electromyograms of Zygomaticus and Trapezius muscles, and electrooculogram (EOG). GSR provides a measure of the resistance of the skin by positioning two electrodes on the distal phalanges of the middle and index fingers. 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. [20] discovered that the mean value of the GSR is related to the level of arousal.

A plethysmograph measures blood volume in the participant's thumb. This measurement can also be used to compute the heart rate (HR) by identification of local maxima (i.e., heart beats), interbeat periods, and heart rate variability (HRV). Blood pressure and HRV correlate with emotions since stress can increase blood pressure. Pleasantness of stimuli can increase peak heart rate response [20]. In addition to the HR and HRV features, spectral features derived from HRV were shown to be a useful feature in emotion assessment [50].

Skin temperature and respiration were recorded since they vary with different emotional states. Slow respiration is linked to relaxation while irregular rhythm, quick variations, and cessation of respiration correspond to more aroused emotions like anger or fear.

Regarding the EMG signals, the Trapezius muscle (neck) activity was recorded to investigate possible head movements during music listening. The activity of the Zygomaticus major was also monitored since this muscle is activated when the participant laughs or smiles. Most of the power in the spectrum of an EMG during muscle contraction is in the frequency range between 4 and 40 Hz. Thus, the muscle activity features were obtained from the energy of EMG signals in this frequency range for different muscles. The rate of eye blinking is another feature which is correlated with anxiety. Eye blinking affects the EOG signal and results in easily detectable peaks in that signal. For further reading on psychophysiology of emotion, we refer the reader to [51].

All the physiological responses were recorded at a 512 Hz sampling rate and later downsampled to 256 Hz to reduce prcoessing time. 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.

In total, 106 features were extracted from peripheral physiological responses based on the proposed features in the literature [22], [26], [52], [53], [54] (see also Table 5).

From the EEG signals, power spectral features were extracted. The logarithms of the spectral power from theta (4-8 Hz), slow alpha (8-10 Hz), alpha (8-12 Hz), beta (12-30 Hz), and gamma (30+ Hz) 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 hemisphere was extracted to measure the possible asymmetry in the brain activities due to emotional stimuli. The total number of EEG features of a trial for 32 electrodes is 216. Table 5 summarizes the list of features extracted from the physiological signals.

6.2 MCA Features

Music videos were encoded into the MPEG-1 format to extract motion vectors and I-frames for further feature extraction. The video stream has been segmented at the shot level using the method proposed in [55].

From a movie director's point of view, lighting key [30] [56] and color variance [30] are important tools to evoke emotions. We therefore extracted lighting key from frames in the HSV space by multiplying the average value V (in HSV) by the standard deviation of the values V (in HSV). Color variance was obtained in the CIE LUV color space by computing the determinant of the covariance matrix of L, U, and V.

TABLE 5
Features Extracted from EEG and Physiological Signals

Signal	Extracted features
GSR	average skin resistance, average of derivative, average of derivative for negative values only (average decrease rate during decay time), proportion of negative samples in the derivative vs. all samples, number of local minima in the GSR signal, average rising time of the GSR signal, 10 spectral power in the [0-2.4]Hz bands, zero crossing rate of Skin conductance slow response (SCSR) [0-0.2]Hz, zero crossing rate of Skin conductance very slow response (SCVSR) [0-0.08]Hz, SCSR and SCVSR mean of peaks magnitude
Blood volume pressure	Average and standard deviation of HR, HRV, and inter beat intervals, energy ratio between the frequency bands [0.04-0.15]Hz and [0.15-0.5]Hz, spectral power in the bands ([0.1-0.2]Hz, [0.2-0.3]Hz, [0.3-0.4]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.
Respiration pattern	band energy ratio (difference between the logarithm of energy between the lower (0.05-0.25Hz) and the higher (0.25-5Hz) bands), average respiration signal, mean of derivative (variation of the respiration signal), standard deviation, range or greatest breath, breathing rhythm (spectral centroid), breathing rate, 10 spectral power in the bands from 0 to 2.4Hz, average peak to peak time, median peak to peak time
Skin temperature	average, average of its derivative, spectral power in the bands ([0-0.1]Hz, [0.1-0.2]Hz)
EMG and EOG	eye blinking rate, energy of the signal, mean and variance of the signal
EEG	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.

Hanjalic and Xu [29] showed the relationship between video rhythm and affect. The average shot change rate and shot length variance were extracted to characterize video rhythm. Fast moving scenes or objects' movements in consecutive frames are also an effective factor for evoking excitement. To measure this factor, the motion component was defined as the amount of motion in consecutive frames computed by accumulating magnitudes of motion vectors for all B and P-frames.

Colors and their proportions are important parameters to elicit emotions [57]. A 20-bin color histogram of hue and lightness values in the HSV space was computed for each I-frame and subsequently averaged over all frames. The resulting bin averages were used as video content-based features. The median of the L value in HSL space was computed to obtain the median lightness of a frame.

Finally, visual cues representing shadow proportion, visual excitement, grayness, and details were also determined according to the definition given in [30].

Sound also has an important impact on affect. For example, loudness of speech (energy) is related to evoked arousal, while rhythm and average pitch in speech signals are related to valence [58]. The audio channels of the videos were extracted and encoded into mono MPEG-3 format at a sampling rate of 44.1 kHz. All audio signals were normalized

TABLE 6
Low-Level Features Extracted from Audio Signals

Feature category	Extracted features
MFCC	MFCC coefficients (13 features) [59], Derivative of MFCC (13 features), Autocorrelation of MFCC (13 features)
Energy	Average energy of audio signal [59]
Formants	Formants up to 5500Hz (female voice) (five features)
Time frequency	MSpectrum flux, Spectral centroid, Delta spectrum magnitude, Band energy ratio [59], [60]
Pitch	First pitch frequency
Zero crossing rate	Average, Standard deviation [59]
Silence ratio	Proportion of silence in a time window [62]

TABLE 7
Average Accuracies (ACC) and F1-Scores (F1, Average of Score for Each Class) over Participants

	Arousal		Valence		Liking	
Modality	ACC	F1	ACC	F1	ACC	F1
EEG	0.620	0.583**	0.576	0.563**	0.554	0.502
Peripheral	0.570	0.533*	0.627	0.608**	0.591	0.538**
MCA	0.651	0.618**	0.618	0.605**	0.677	0.634**
Random	0.500	0.483	0.500	0.494	0.500	0.476
Majority class	0.644	0.389	0.586	0.368	0.670	0.398
Class ratio	0.562	0.500	0.525	0.500	0.586	0.500

Stars indicate whether the F1-score distribution over subjects is significantly higher than 0.5 according to an independent one-sample t-test ($** = p < 0.01$, $* = p < 0.05$). For comparison, expected results are given for classification based on random voting, voting according to the majority class, and voting with the ratio of the classes.

to the same amplitude range before further processing. A total of 53 low-level audio features were determined for each of the audio signals. These features, listed in Table 6, are commonly used in audio and speech processing and audio classification [59], [60]. MFCC, formants, and the pitch of audio signals were extracted using the PRAAT software package [61].

6.3 Fusion of Single-Modality Results

Fusion of the multiple modalities explained above aims at improving classification results by exploiting the complementary nature of different modalities. In general, approaches for modality fusion can be classified into two broad categories, namely, feature fusion (or early integration) and decision fusion (or late integration) [63]. In feature fusion, the features extracted from signals of different modalities are concatenated to form a composite feature vector and then input to a recognizer. In decision fusion, on the other hand, each modality is processed independently by the corresponding classifier and the outputs of the classifiers are combined to yield the final result. Each approach has its own advantages. For example, implementing a feature fusion-based system is straightforward, while a decision fusion-based system can be constructed by using existing unimodal classification systems. Moreover, feature fusion can consider synchronous characteristics of the involved modalities, whereas decision fusion allows us to model asynchronous characteristics of the modalities flexibly.

An important advantage of decision fusion over feature fusion is that since each of the signals is processed and classified independently in decision fusion, it is relatively easy to employ an optimal weighting scheme to adjust the relative amount of the contribution of each modality to the final decision according to the reliability of the modality. The weighting scheme used in our work can be formalized as follows: For a given test datum X, the classification result of the fusion system is

$$c^* = \arg\max_i \left\{ \prod_{m=1}^{M} P_i(X|\lambda_m)^{\alpha_m} \right\}, \quad (4)$$

where M is the number of modalities considered for fusion, λ_m is the classifier for the mth modality, and $P_i(X|\lambda_m)$ is its output for the ith class. The weighting factors α_m, which satisfy $0 \leq \alpha_m \leq 1$ and $\sum_{m=1}^{M} \alpha_m = 1$, determine how much each modality contributes to the final decision and represents the modality's reliability.

We adopt a simple method where the weighting factors are fixed once their optimal values are determined from the training data. The optimal weight values are estimated by exhaustively searching the regular grid space, where each weight is incremented from 0 to 1 by 0.01 and the weighting values producing the best classification results for the training data are selected.

6.4 Results and Discussion

Table 7 shows the average accuracies and F1-scores (average F1-score for both classes) over participants for each modality and each rating scale. We compare the results to the expected values (analytically determined) of voting randomly, voting according to the majority class in the training data, and voting for each class with the probability of its occurrence in the training data. For determining the expected values of majority voting and class ratio voting, we used the class ratio of each participant's feedback during the experiment. These results are slightly too high as, in reality, the class ratio would have to be estimated from the training set in each fold of the leave-one-out cross validation.

Voting according to the class ratio gives an expected F1-score of 0.5 for each participant. To test for significance, an independent one-sample t-test was performed, comparing the F1-distribution over participants to the 0.5 baseline. As can be seen from the table, eight out of the nine F1-scores obtained are significantly better than the class ratio baseline. The exception is the classification of liking using EEG signals ($p = 0.068$). When voting according to the majority class, relatively high accuracies are achieved, due to the imbalanced classes. However, this voting scheme also has the lowest F1-scores.

Overall, classification using the MCA features fares significantly better than EEG and peripheral ($p < 0.0001$ for both), while EEG and peripheral scores are not significantly different ($p = 0.41$) (tested using a two-sided repeated samples t-test over the concatenated results from each rating scale and participant).

The modalities can be seen to perform moderately complementarily, where EEG scores best for arousal, peripheral for valence, and MCA for liking. Of the different

TABLE 8
F1-Scores for Fusion of the Best Two Modalities and All Three Modalities Using the Equal Weights and Optimal Weights Scheme

	Arousal			Valence			Liking		
	Modality	Optimal w.	Equal w.	Modality	Optimal w.	Equal w.	Modality	Optimal w.	Equal w.
Best single modality	MCA	0.618	—	PER	0.608	—	MCA	0.634	—
Best two modalities	EEG,MCA	0.631	0.629	MCA,PER	0.638	0.652	MCA,PER	0.622	0.642
All three modalities	All	0.616	0.618	All	0.647	0.640	All	0.618	0.607

For comparison, the F1-score for the best single modality is also given.

rating scales, valence classification performed best, followed by liking, and, last, arousal.

Table 8 gives the results of multimodal fusion. Two fusion methods were employed: the method described in Section 6.3 and the basic method where each modality is weighed equally. The best results were obtained when only the two best-performing modalities were considered. Though fusion generally outperforms the single modalities, it is only significant for the case of MCA, PER weighted equally in the valence scale ($p = 0.025$).

While the results presented are significantly higher than random classification, there remains much room for improvement. Signal noise, individual physiological differences, and limited quality of self-assessments make single-trial classification challenging.

7 Conclusion

In this work, we have presented a database for the analysis of spontaneous emotions. The database contains physiological signals of 32 participants (and frontal face video of 22 participants), where each participant watched and rated their emotional response to 40 music videos along the scales of arousal, valence, and dominance as well as their liking of and familiarity with the videos. We presented a novel semi-automatic stimuli selection method using affective tags, which was validated by an analysis of the ratings participants gave during the experiment. Significant correlates were found between the participant ratings and EEG frequencies. Single-trial classification was performed for the scales of arousal, valence, and liking using features extracted from the EEG, peripheral, and MCA modalities. The results were shown to be significantly better than random classification. Finally, decision fusion of these results yielded a modest increase in the performance, indicating at least some complementarity to the modalities.

The database is made publicly available and it is our hope that other researchers will try their methods and algorithms on this highly challenging database.

Acknowledgments

The research leading to these results has been performed in the frameworks of European Community's 17th Framework Program (FP7/2007-2011) under grant agreement no. 216444 (PetaMedia). Furthermore, the authors gratefully acknowledge the support of the BrainGain Smart Mix Programme of the Netherlands Ministry of Economic Affairs, the Netherlands Ministry of Education, Culture, and Science and the Swiss National Foundation for Scientific Research and the NCCR Interactive Multimodal Information Management (IM2). The authors also thank Sebastian Schmiedeke and Pascal Kelm at the Technische Universität Berlin for performing the shot boundary detection on this data set.

References

[1] M.K. Shan, F.F. Kuo, M.F. Chiang, and S.Y. Lee, "Emotion-Based Music Recommendation by Affinity Discovery from Film Music," *Expert Systems with Applications,* vol. 36, no. 4, pp. 7666-7674, May 2009.

[2] M. Tkalčič, U. Burnik, and A. Košir, "Using Affective Parameters in a Content-Based Recommender System for Images," *User Modeling and User-Adapted Interaction,* vol. 20, pp. 1-33-33, Sept. 2010.

[3] J.J.M. Kierkels, M. Soleymani, and T. Pun, "Queries and Tags in Affect-Based Multimedia Retrieval," *Proc. IEEE Int'l Conf. Multimedia and Expo,* pp. 1436-1439, 2009.

[4] M. Soleymani, J. Lichtenauer, T. Pun, and M. Pantic, "A Multi-Modal Database for Affect Recognition and Implicit Tagging," *IEEE Trans. Affective Computing,* vol. 3, no. 1, pp. 42-55, Jan.-Mar. 2012.

[5] A. Savran, K. Ciftci, G. Chanel, J.C. Mota, L.H. Viet, B. Sankur, L. Akarun, A. Caplier, and M. Rombaut, "Emotion Detection in the Loop from Brain Signals and Facial Images," *Proc. eNTERFACE,* July 2006.

[6] P. Ekman, W.V. Friesen, M. O'Sullivan, A. Chan, I. Diacoyanni-Tarlatzis, K. Heider, R. Krause, W.A. LeCompte, T. Pitcairn, and P.E. Ricci-Bitti, "Universals and Cultural Differences in the Judgments of Facial Expressions of Emotion," *J. Personality and Social Psychology,* vol. 53, no. 4, pp. 712-717, Oct. 1987.

[7] W.G. Parrott, *Emotions in Social Psychology: Essential Readings.* Psychology Press, 2001.

[8] R. Plutchik, "The Nature of Emotions," *Am. Scientist,* vol. 89, p. 344, 2001.

[9] J.A. Russell, "A Circumplex Model of Affect," *J. Personality and Social Psychology,* vol. 39, no. 6, pp. 1161-1178, 1980.

[10] M.M. Bradley and P.J. Lang, "Measuring Emotion: The Self-Assessment Manikin and the Semantic Differential," *J. Behavior Therapy Experimental Psychiatry,* vol. 25, no. 1, pp. 49-59, Mar. 1994.

[11] M. Pantic, M. Valstar, R. Rademaker, and L. Maat, "Web-Based Database for Facial Expression Analysis," *Proc. Int'l Conf. Multimedia and Expo,* pp. 317-321, 2005.

[12] E. Douglas-Cowie, R. Cowie, and M. Schröder, "A New Emotion Database: Considerations, Sources and Scope," *Proc. Int'l Symp. Computer Architecture,* pp. 39-44, 2000.

[13] H. Gunes and M. Piccardi, "A Bimodal Face and Body Gesture Database for Automatic Analysis of Human Nonverbal Affective Behavior," *Proc. 18th Int'l Conf. Pattern Recognition,* vol. 1, pp. 1148-1153, 2006.

[14] G. Fanelli, J. Gall, H. Romsdorfer, T. Weise, and L. Van Gool, "A 3-D Audio-Visual Corpus of Affective Communication," *IEEE Trans. Multimedia,* vol. 12, no. 6, pp. 591-598, Oct. 2010.

[15] M. Grimm, K. Kroschel, and S. Narayanan, "The Vera am Mittag German Audio-Visual Emotional Speech Database," *Proc. Int'l Conf. Multimedia and Expo,* pp. 865-868, 2008.

[16] J.A. Healey, "Wearable and Automotive Systems for Affect Recognition from Physiology," PhD dissertation, MIT, 2000.

[17] 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.

[18] P. Lang, M. Bradley, and B. Cuthbert, "International Affective Picture System (IAPS): Affective Ratings of Pictures and Instruction Manual," Technical Report A-8, Univ. of Florida, 2008.

[19] 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, Jan. 2009.

[20] P. Lang, M. Greenwald, M. Bradely, and A. Hamm, "Looking at Pictures—Affective, Facial, Visceral, and Behavioral Reactions," *Psychophysiology,* vol. 30, no. 3, pp. 261-273, May 1993.

[21] 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.

[22] J. Wang and Y. Gong, "Recognition of Multiple Drivers' Emotional State," *Proc. Int'l Conf. Pattern Recognition,* pp. 1-4, 2008.

[23] C.L. Lisetti and F. Nasoz, "Using Noninvasive Wearable Computers to Recognize Human Emotions from Physiological Signals," *EURASIP J. Applied Signal Processing,* vol. 2004, no. 1, pp. 1672-1687, Jan. 2004.

[24] G. Chanel, J. 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.

[25] J. Kierkels, M. Soleymani, and T. Pun, "Queries and Tags in Affect-Based Multimedia Retrieval," *Proc. Int'l Conf. Multimedia and Expo,* pp. 1436-1439, June 2009.

[26] M. Soleymani, G. Chanel, J.J.M. Kierkels, and T. Pun, "Affective Characterization of Movie Scenes Based on Content Analysis and Physiological Changes," *Int'l J. Semantic Computing,* vol. 3, no. 2, pp. 235-254, June 2009.

[27] A. Yazdani, J.-S. Lee, and T. Ebrahimi, "Implicit Emotional Tagging of Multimedia Using EEG Signals and Brain Computer Interface," *Proc. SIGMM Workshop Social Media,* pp. 81-88, 2009.

[28] P. Ekman, W. Friesen, M. Osullivan, A. Chan, I. Diacoyanni-tarlatzis, K. Heider, R. Krause, W. Lecompte, T. Pitcairn, P. Riccibitti, K. Scherer, M. Tomita, and A. Tzavaras, "Universals and Cultural-Differences in the Judgments of Facial Expressions of Emotion," *J. Personality and Social Psychology,* vol. 53, no. 4, pp. 712-717, Oct. 1987.

[29] A. Hanjalic and L.-Q. Xu, "Affective Video Content Representation and Modeling," *IEEE Trans. Multimedia,* vol. 7, no. 1, pp. 143-154, Feb. 2005.

[30] H.L. Wang and L.-F. Cheong, "Affective Understanding in Film," *IEEE Trans. Circuits and Systems for Video Technology,* vol. 16, no. 6, pp. 689-704, June 2006.

[31] M. Soleymani, J. Kierkels, G. Chanel, and T. Pun, "A Bayesian Framework for Video Affective Representation," *Proc. Int'l Conf. Affective Computing and Intelligent Interaction,* pp. 1-7, Sept. 2009.

[32] D. Liu, "Automatic Mood Detection from Acoustic Music Data," *Proc. Int'l Conf. Music Information Retrieval,* pp. 13-17, 2003.

[33] L. Lu, D. Liu, and H.-J. Zhang, "Automatic Mood Detection and Tracking of Music Audio Signals," *IEEE Trans. Audio, Speech, and Language Processing,* vol. 14, no. 1, pp. 5-18, Jan. 2006.

[34] Y.-H. Yang and H.H. Chen, "Music Emotion Ranking," *Proc. Int'l Conf. Acoustics, Speech, and Signal Processing,* pp. 1657-1660, 2009.

[35] S. Koelstra, A. Yazdani, M. Soleymani, C. Mühl, J.-S. Lee, A. Nijholt, T. Pun, T. Ebrahimi, and I. Patras, "Single Trial Classification of EEG and Peripheral Physiological Signals for Recognition of Emotions Induced by Music Videos," *Proc. Brain Informatics,* pp. 89-100, 2010.

[36] M.E. Tipping, "Sparse Bayesian Learning and the Relevance Vector Machine," *J. Machine Learning Research,* vol. 1, pp. 211-244, June 2001.

[37] J.D. Morris, "SAM: The Self-Assessment Manikin an Efficient Cross-Cultural Measurement of Emotional Response," *J. Advertising Research,* vol. 35, no. 8, pp. 63-68, 1995.

[38] M. Bradley and P. Lang, "International Affective Digitized Sounds (IADS): Stimuli, Instruction Manual and Affective Ratings," Technical Report B-2, The Center for Research in Psychophysiology, Univ. of Florida, 1999.

[39] N. Lazar, "Combining Brains: A Survey of Methods for Statistical Pooling of Information," *NeuroImage,* vol. 16, no. 2, pp. 538-550, June 2002.

[40] T.M. Loughin, "A Systematic Comparison of Methods for Combining p-Values from Independent Tests," *Computational Statistics and Data Analysis,* vol. 47, pp. 467-485, 2004.

[41] R.J. Barry, A.R. Clarke, S.J. Johnstone, C.A. Magee, and J.A. Rushby, "EEG Differences between Eyes-Closed and Eyes-Open Resting Conditions," *Clinical Neurophysiology,* vol. 118, no. 12, pp. 2765-2773, Dec. 2007.

[42] R.J. Barry, A.R. Clarke, S.J. Johnstone, and C.R. Brown, "EEG Differences in Children between Eyes-Closed and Eyes-Open Resting Conditions," *Clinical Neurophysiology,* vol. 120, no. 10, pp. 1806-1811, Oct. 2009.

[43] W. Klimesch, P. Sauseng, and S. Hanslmayr, "EEG Alpha Oscillations: The Inhibition-Timing Hypothesis," *Brain Research Rev.,* vol. 53, no. 1, pp. 63-88, Jan. 2007.

[44] H. Cole and W.J. Ray, "EEG Correlates of Emotional Tasks Related to Attentional Demands," *Int'l J. Psychophysiology,* vol. 3, no. 1, pp. 33-41, July 1985.

[45] J. Onton and S. Makeig, "High-Frequency Broadband Modulations of Electroencephalographic Spectra," *Frontiers in Human Neuroscience,* vol. 3, 2009.

[46] I. Goncharova, D.J. McFarland, J.R. Vaughan, and J.R. Wolpaw, "EMG Contamination of EEG: Spectral and Topographical Characteristics," *Clinical Neurophysiology,* vol. 114, no. 9, pp. 1580-1593, Sept. 2003.

[47] E. Harmon-Jones, "Clarifying the Emotive Functions of Asymmetrical Frontal Cortical Activity," *Psychophysiology,* vol. 40, no. 6, pp. 838-848, 2003.

[48] R.R. Cornelius, *The Science of Emotion. Research and Tradition in the Psychology of Emotion.* Prentice Hall, 1996.

[49] 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.

[50] R. McCraty, M. Atkinson, W. Tiller, G. Rein, and A. Watkins, "The Effects of Emotions on Short-Term Power Spectrum Analysis of Heart Rate Variability," *Am. J. Cardiology,* vol. 76, no. 14, pp. 1089-1093, 1995.

[51] S.D. Kreibig, "Autonomic Nervous System Activity in Emotion: A Review," *Biological Psychology,* vol. 84, no. 3, pp. 394-421, 2010.

[52] 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.

[53] 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.

[54] 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.

[55] P. Kelm, S. Schmiedeke, and T. Sikora, "Feature-Based Video Key Frame Extraction for Low Quality Video Sequences," *Proc. Int'l Workshop Image Analysis for Multimedia Interactive Services,* pp. 25-28, May 2009.

[56] Z. Rasheed, Y. Sheikh, and M. Shah, "On the Use of Computable Features for Film Classification," *IEEE Trans. Circuits and Systems for Video Technology,* vol. 15, no. 1, pp. 52-64, Jan. 2005.

[57] P. Valdez and A. Mehrabian, "Effects of Color on Emotions," *J. Experimental Psychology,* vol. 123, no. 4, pp. 394-409, Dec. 1994.

[58] R.W. Picard, *Affective Computing.* MIT Press, Sept. 1997.

[59] D. Li, I.K. Sethi, N. Dimitrova, and T. McGee, "Classification of General Audio Data for Content-Based Retrieval," *Pattern Recognition Letters,* vol. 22, no. 5, pp. 533-544, 2001.

[60] L. Lu, H. Jiang, and H. Zhang, "A Robust Audio Classification and Segmentation Method," *Proc. Ninth ACM Int'l Conf. Multimedia,* pp. 203-211, 2001.

[61] P. Boersma, "Praat, A System for Doing Phonetics by Computer," *Glot Int'l,* vol. 5, nos. 9/10, pp. 341-345, 2001.

[62] L. Chen, S. Gunduz, and M. Ozsu, "Mixed Type Audio Classification with Support Vector Machine," *Proc. Int'l Conf. Multimedia and Expo,* pp. 781 -784, July 2006.

[63] J.-S. Lee and C.H. Park, "Robust Audio-Visual Speech Recognition Based on Late Integration," *IEEE Trans. Multimedia,* vol. 10, no. 5, pp. 767-779, Aug. 2008.

Sander Koelstra received the BSc and MSc degrees in computer science from the Delft University of Technology, The Netherlands, in 2006 and 2008, respectively. Currently, he is working toward the PhD degree in the School of Electronic Engineering and Computer Science at Queen Mary University of London. His research interests include the areas of brain-computer interaction, computer vision, and pattern recognition. He is a student member of the IEEE.

Christian Mühl received the MSc degree in cognitive science from the University of Osnabrueck, Germany, in 2007. Since then, he has been working on neurophysiological mechanisms of cross-modal attention. Currently, he is working as a PhD researcher in the Human-Media Interaction group of the University of Twente, The Netherlands. His research interests include the identification of affective states by neurophysiological signals in various induction contexts. He is especially interested in the differential effects of auditory and visual affective stimulation on the activity of the brain as measured via electroencephalography.

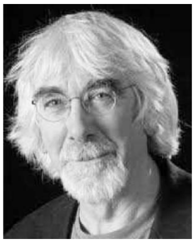
Mohammad Soleymani received the BSc and MSc degrees from the Department of Electrical and Computer Engineering, University of Tehran, in 2003 and 2006, respectively. Currently, he is working toward the doctoral degree and is a research assistant in the Computer Vision and Multimedia Laboratory (CVML), Computer Science Department, University of Geneva. His research interests include: affective computing, and multimedia information retrieval. He has been coorganizing the MediaEval multimedia benchmarking initiative since 2010. He is a student member of the IEEE.

Jong-Seok Lee received the PhD degree in electrical engineering and computer science in 2006 from KAIST, Daejeon, Korea, where he worked as a postdoctoral researcher and an adjunct professor. He worked as a research scientist in the Multimedia Signal Processing Group at the Swiss Federal Institute of Technology in Lausanne (EPFL), Lausanne, Switzerland. Currently, he is an assistant professor in the School of Integrated Technology, Yonsei University, Korea. He is an author or coauthor of more than 50 publications. His current research interests include multimedia signal processing and multimodal human-computer interaction. He is a member of the IEEE.

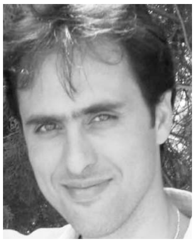
Ashkan Yazdani received the MSc degree in electrical engineering from the University of Tehran, Iran, in 2007. From 2006 to 2007, he worked as a research assistant in the Brain Signal Processing Group at the University of Tehran on the topics of EEG-based person identification, brain-computer interfacing, and fMRI signal processing. Currently, he is working toward the PhD degree in the Multimedia Signal Processing Group, Swiss Federal Institute of Technology in Lausanne (EPFL), Lausanne, Switzerland. His main research interests include EEG-based brain-computer interface systems and biomedical signal processing.

Touradj Ebrahimi received the MSc and PhD degrees in electrical engineering from the Swiss Federal Institute of Technology in Lausanne (EPFL), Lausanne, Switzerland, in 1989 and 1992, respectively. From 1989 to 1992, he was a research assistant in the Signal Processing Laboratory of EPFL. In 1990, he was a visiting researcher in the Signal and Image Processing Institute of the University of Southern California, Los Angeles, California. In 1993, he was a research engineer in the Corporate Research Laboratories of Sony Corporation in Tokyo. In 1994, he served as a research consultant at AT&T Bell Laboratories. Currently, he is working as a professor heading the Multimedia Signal Processing Group at EPFL, where he is involved with various aspects of digital video and multimedia applications. He is the coauthor of more than 100 papers and holds 10 patents. He is a member of the IEEE.

Thierry Pun received the PhD degree in image processing for the development of a visual prosthesis for the blind in 1982 from the Swiss Federal Institute of Technology, Lausanne, Switzerland. He is head of the Computer Vision and Multimedia Laboratory, Computer Science Department, University of Geneva, Switzerland. He was a visiting fellow from 1982 to 1985 at the National Institutes of Health, Bethesda, Maryland. After being a CERN fellow from 1985 to 1986 in Geneva, Switzerland, he joined the University of Geneva in 1986, where he is currently a full professor in the Computer Science Department. He has authored or coauthored more than 300 full papers as well as eight patents. His current research interests include affective computing and multimodal interaction, concern: physiological signals analysis for emotion assessment and brain-computer interaction, multimodal interfaces for blind users, data hiding, and multimedia information retrieval systems. He is a member of the IEEE.



Anton Nijholt is a full professor of human media interaction at the University of Twente (NL). He has coorganized many conferences (e.g., IVA 2009 and ACII 2009) and satellite workshops (e.g., on affective brain-computer interfacing) and has been the guest editor of many journals for special issues devoted to selections of updates of papers from these conferences and workshops. His main research interests include multimodal interaction, brain-computer interfacing, virtual humans, affective computing, and entertainment computing. He is a member of the IEEE.



Ioannis (Yiannis) Patras received the PhD degree from the Department of Electrical Engineering, Delft University of Technology, The Netherlands, in 2001. He is a senior lecturer in computer vision in the School of Electronic Engineering and Computer Science at Queen Mary University of London. He is/has been on the organizing committee of IEEE SMC 2004, Face and Gesture Recognition 2008, ICMR2011, ACM Multimedia 2013, and was the general chair of WIAMIS 2009. He is working as an associate editor on the *Image and Vision Computing Journal*. His research interests include the areas of computer vision and pattern recognition, with emphasis on human sensing and its applications in multimedia retrieval and multimodal human-computer interaction. Currently, he is interested in brain-computer interfaces and the analysis of facial and body gestures. He is a senior member of the IEEE.