Affect Recognition Based on Physiological Changes During the Watching of Music Video

ARTICLE *in* THE ACM TRANSACTIONS ON INTERACTIVE INTELLIGENT SYSTEMS · MARCH 2012

DOI: 10.1145/2133366.2133373

4 AUTHORS, INCLUDING:

Jong-Seok Lee
Yonsei University
74 PUBLICATIONS **617** CITATIONS
SEE PROFILE

Jean-Marc Vesin
École Polytechnique Fédérale de Lausanne
247 PUBLICATIONS **2,392** CITATIONS
SEE PROFILE

Touradj Ebrahimi
École Polytechnique Fédérale de Lausanne
521 PUBLICATIONS **9,223** CITATIONS
SEE PROFILE

Affect Recognition Based on Physiological Changes During the Watching of Music Videos

ASHKAN YAZDANI, Ecole Polytechnique Fédérale de Lausanne
JONG-SEOK LEE, Yonsei University
JEAN-MARC VESIN and TOURADJ EBRAHIMI, Ecole Polytechnique Fédérale de Lausanne

Assessing emotional states of users evoked during their multimedia consumption has received a great deal of attention with recent advances in multimedia content distribution technologies and increasing interest in personalized content delivery. Physiological signals such as the electroencephalogram (EEG) and peripheral physiological signals have been less considered for emotion recognition in comparison to other modalities such as facial expression and speech, although they have a potential interest as alternative or supplementary channels. This article presents our work on: (1) constructing a dataset containing EEG and peripheral physiological signals acquired during presentation of music video clips, which is made publicly available, and (2) conducting binary classification of induced positive/negative valence, high/low arousal, and like/dislike by using the aforementioned signals. The procedure for the dataset acquisition, including stimuli selection, signal acquisition, self-assessment, and signal processing is described in detail. Especially, we propose a novel asymmetry index based on relative wavelet entropy for measuring the asymmetry in the energy distribution of EEG signals, which is used for EEG feature extraction. Then, the classification systems based on EEG and peripheral physiological signals are presented. Single-trial and single-run classification results indicate that, on average, the performance of the EEG-based classification outperforms that of the peripheral physiological signals. However, the peripheral physiological signals can be considered as a good alternative to EEG signals in the case of assessing a user's preference for a given music video clip (like/dislike) since they have a comparable performance to EEG signals while being more easily measured.

Categories and Subject Descriptors: H.5.2 [**Information Interfaces and Presentation**]: User Interfaces—*Evaluation / methodology*; I.5.2 [**Pattern Recognition**]: Design Methodology—*Classifier design and evaluation*; *pattern analysis*; I.5.4 [**Pattern Recognition**]: Applications—*Signal processing*

General Terms: Algorithms, Measurement, Performance, Human Factors

Additional Key Words and Phrases: Emotion classification, EEG, physiological signals, signal processing, pattern classification, affective computing

The research leading to these results was performed within the framework of European Community's Seventh Framework Program (FP7/2007-2011) under grant agreement no. 216444 (PetaMedia). Furthermore, the authors gratefully acknowledge the support of the Swiss National Foundation for Scientific Research, and the NCCR Interactive Multimodal Information Management (IM2). This work was also supported in part by the Ministry of Knowledge Economy, Korea, under the IT Consilience Creative Program (NIPA-2010-C1515-1001-0001) and Yonsei University Research Fund of 2011.
Authors' addresses: A. Yazdani (corresponding author), Multimedia Signal Processing Group, Ecole Polytechnique Fédérale de Lausanne, EPFL/STI/IEL/GR-EB, Station 11, CH-1015, Lausanne, Switzerland; email: ashkan.yazdani@epfl.ch; J.-S. Lee, School of Integrated Technology, Yonsei University 406-840, Incheon, Korea; J.-M. Vesin, Applied Signal Processing Group, Ecole Polytechnique Fédérale de Lausanne, EPFL/STI/IEL/GR-EB, Station 11, CH-1015, Lausanne, Switzerland; T. Ebrahimi, Multimedia Signal Processing Group, Ecole Polytechnique Fédérale de Lausanne, EPFL/STI/IEL/GR-EB, Station 11, CH-1015, Lausanne, Switzerland.

DOI 10.1145/2133366.2133373 http://doi.acm.org/10.1145/2133366.2133373

ACM Reference Format:
Yazdani, A., Lee, J.-S., Vesin, J.-M., and Ebrahimi, T. 2012. Affect recognition based on physiological changes during the watching of music videos. ACM Trans. Interact. Intell. Syst. 2, 1, Article 7 (March 2012), 26 pages. DOI = 10.1145/2133366.2133373 http://doi.acm.org/10.1145/2133366.2133373

1. INTRODUCTION

For the last few years, the study of the human emotions and the recognition of various affective states have received a great deal of interest. Emotion is a psycho-physiological process or mental state triggered spontaneously rather than through conscious effort during perception of a situation or an object. It plays 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. Human emotions are very subjective and nondeterministic. The same stimulus may create different emotions in different individuals, and the same individual may express different emotions in response to the same stimulus, at different times.

Several attempts have been made to incorporate emotion in Human-Computer Interaction (HCI) systems (e.g., Fragopanagos and Taylor [2005], Jaimes and Sebe [2007], and Cowie et al. [2002]). Detection of human emotions can be accomplished by monitoring and interpreting the different cues that are given in both verbal and nonverbal communication. However, most of the current HCI systems are not able to interpret this information and suffer a lack of emotional intelligence. In other words, they are not able to identify human emotional states and to take this information into account in their decision making process for proper actions to execute. Affective computing aims at filling this gap by detecting emotional cues occurring during HCI and synthesizing emotional responses.

Furthermore, emotions are known to modulate tendencies to certain actions [Cowie et al. 2002; Sutton and Davidson 1997]. In other words, people tend to be attracted by the stimuli or situations inducing positive emotions such as happiness and joy and withdraw from the situations causing negative emotions such as fear and disgust. Therefore, emotion detection can produce an indicator for behavior prediction and it is of interest to monitor critical emotional states that could lead to potential harmful or dangerous behaviors.

More specifically, emotion assessment can be of significant interest in multimedia indexing, retrieval, and recommendation. For example, emotional tags can be obtained and assigned to the content by observing a user's emotional state during his/her multimedia consumption, which is called implicit tagging. Among various kinds of tags associated with a given content, emotional tags can be of great interest for the aim of personalized content delivery [Hanjalic and Xu 2005]. For instance, when a user feels sad, he/she may want to watch video clips containing funny stories, which will make him/her feel better. Sometimes, one may not want to watch video clips containing scenes with too much violence. In such cases, emotional tags or tags about different genres (e.g., horror, humor, etc.) can be used effectively in multimedia search and retrieval. In addition, it has been shown that the level of interest can also be assessed through analysis of facial expression [Yeasin et al. 2006], body gesture [Mota and Picard 2003], and physiological signals. Gaze detection and tracking is usually used for obtaining information about users' attention [Jaimes and Sebe 2007]. It is possible to use those cues for developing content recommendation schemes.

An appropriate modeling for emotion must be developed and used, in order to assess users' affective state. How to represent and model emotions is, however, a challenging task. Until today, numerous theorists and researchers have conducted research on this subject and consequently a large amount of literature exists today with sometimes very different solutions. Generally, there are two different families of emotion models:

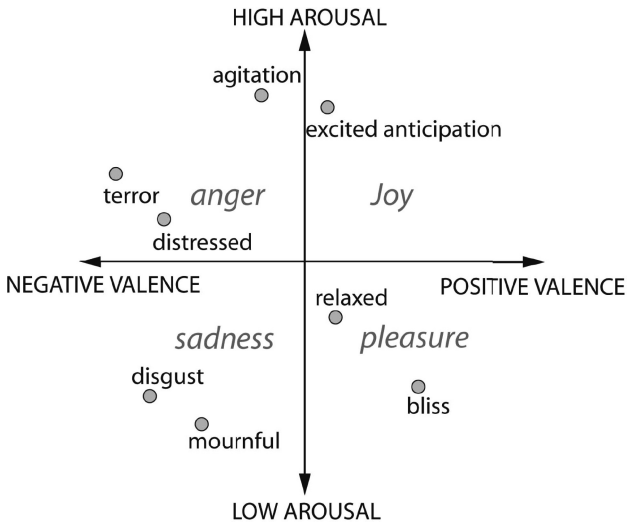

Fig. 1. Arousal-valence space and basic emotions.

the categorical models and the dimensional models. The rational for the categorical models is to have discrete basic categories of emotions from which every other emotion can be built by combining these basic emotions. The most common basic emotions are as fear, anger, sadness, joy, disgust, surprise as found by Ekman et al. [1987]. The dimensional models, on the other hand, describe the components of emotions and are often represented as a two- or three-dimensional space where the emotions are presented as points in the coordinate space of these dimensions. The goal of the dimensional model is not to find a finite set of emotions as in the categorical model but to find a finite set of underlying components of emotions [Plutchik 2001; Russell 1980]. In this work, Russell's arousal-valence dimensional model of emotion will be used to quantitatively analyze the emotions. The dimension *valence* provides information about the degree of pleasantness of the content and ranges from pleasant (positive) to unpleasant (negative). The dimension *arousal* represents the inner activation and ranges from energized to calm. In this scale, each emotional state can be placed on a two-dimensional 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 may be also included in the model [Russell 1980], which helps to distinguish between grief and rage and goes from no control to full control. Figure 1 illustrates the arousal-valence space and the distribution of basic emotions on this space [Kim and André 2008].

Until today, many research studies have investigated human facial expressions [Zeng et al. 2009], human voice and speech [Ververidis and Kotropoulos 2006], and the fusion of different modalities [Fragopanagos and Taylor 2005; Sebe et al. 2006; Cowie 2010] to extract information about subjects' affective states. Physiological signals originating from the Peripheral Nervous System (PNS) are also known to convey traces of emotion and they have been studied for the aim of emotion recognition [Lang et al. 1993; Kim and André 2008; Wang and Gong 2008; Healey 2000; Lisetti and Nasoz 2004; Chanel et al. 2009]. Kim and André [2008] investigate the potential of physiological signals as reliable channels for emotion recognition during music listening and they showed that for four emotional states of three subjects, an average recognition accuracy of 95% can be achieved. Kim et al. [2004] present a physiological signal-based emotion recognition

system induced by combination of photos and music. They show that an average correct classification ratios of 78.4% and 61.8% can be achieved for recognition of three and four categories, respectively. Lisetti and Nasoz [2004] investigate physiological changes during watching movie scenes and they show that a high recognition rate of 84% for the recognition of six emotions can be achieved.

Furthermore, a few studies have examined the feasibility of analyzing EEG signals, emanating from the Central Nervous System (CNS), for extracting information about human emotion [Kostyunina and Kulikov 1996; Krause et al. 2000; Aftanas et al. 2004; Ishino and Hagiwara 2003]. Chanel et al. [2006] study the changes in EEG signals during watching emotion evocative images and show that an accuracy of 58% for classification of three emotions can be achieved. Schaaff and Schultz [2009] use a headband EEG acquisition interface and Support Vector Machines (SVMs) to recognize three emotional categories induced by images. In Lin et al. [2010], power spectrum density of different EEG sub-bands were extracted as features during different emotions induced during listening to music and a correct classification of 82% for four emotions was achieved. Petrantonakis and Hadjileontiadis [2010] study the changes in the EEG signal of subjects when presented with images of faces expressing six basic emotions. They showed that a classification accuracy of 83% can be achieved using features based on higher- order crossings and support vector machine classification.

In Yazdani et al. [2009], we presented a P300-based Brain Computer Interface (BCI) system for emotional annotation of video contents. However, in that work, the emotions of the user are not captured by means of analyzing his/her EEG signals but instead, he/she selects a basic emotion after each video using the developed BCI system. Furthermore, in our previous works [Yazdani et al. 2010lb, 2010a], we explored the possibility of detecting curiosity (scientific interest) of different subjects by analyzing their brain electrical activities. More precisely, a BCI system was developed to retrieve scientifically interesting images from an image database only by analyzing users' EEG signals while they were watching very fast sequences of images of the database.

The results and achievements of the mentioned prior works have motivated the creation of novel databases containing emotional expressions in various modalities. Current publicly available databases are often comprised of video (facial expressions, body gestures), audio (speech and voices in different languages), and audiovisual data. One of the first databases of physiological signals for analysis of emotion was presented in Healey [2000]. In that work, a collection of multi-parameter recordings from 24 healthy volunteers, while they were driving on a prescribed route including city streets and highways in and around Boston, Massachusetts, was performed. The objective of this study was to investigate the feasibility of automated recognition of stress on the basis of the recorded signals, which include electrocardiogram (ECG), electromyogram (EMG), skin resistance measured on the hand and foot, and respiration[1]. Moreover, a publicly available multimodal emotional database which includes EEG, peripheral physiological responses, and facial expressions is the eNTERFACE 2005 emotional database presented in Savran et al. [2006]. This data was recorded during emotion elicitation using emotional images from the international affective picture system [Lang et al. 2008]. This database is composed of two sets. In the first set, EEG, peripheral physiological signals, videos of subjects' facial expression and functional near-infra-red spectroscopy (fNIRS) data were recorded from five male subjects. The second set contains fNIRS and videos of subjects' facial expressions taken from 16 male and female subjects.

Most of the mentioned prior works investigated emotions induced by single modalities (music, movie clips, photos). This article presents our work on classification of

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

positive/negative valence, high/low arousal, and like/dislike, induced in users during viewing of music video clips, based on EEG and peripheral physiological signals. The major contributions of this article are twofold. First, the database acquired and studied in our work is introduced and is made publicly available.[2] Second, we present a novel asymmetry index based on Relative Wavelet Entropy (RWE) to acknowledge the asymmetric distribution of EEG energies over the right and left hemispheres of the brain. Then, the results of single-trial and single-run emotion classification using EEG and peripheral physiological signals are presented, compared, and discussed. Finally, the ability to generalize the proposed methodology for recognizing emotions of new subjects will be explored.

In the current work, music video clips are used as audiovisual stimuli in order to elicit different emotions. To this end, a relatively large set of music video clips (70 clips) was gathered. A subjective test was then performed to select the most appropriate test material. For each video, a two-minutes highlight was extracted for the experiment. Six participants were asked to participate in the experiment and their physiological signals (EEG and peripheral physiological signals) were recorded while they were watching the 20 selected music video clips. Participants were also asked to rate each video in terms of arousal, valence, and like/dislike.

Our previous work [Koelstra et al. 2010] presents the preliminary results of analyzing these signals. The present work provides a more detailed description of the database acquisition procedure and presents more efficient signal processing and classification methodology for improvement of the preliminary results.

The acquired database includes the following items:

(1) the recordings of EEG and peripheral physiological signals;
(2) the participants' ratings of arousal, valence, and like/dislike;
(3) the list of video clips used[3]; and
(4) the subjective ratings from the initial online assessment of 70 candidate video clips (refer to Section 2.1).

The layout of the article is as follows. In Section 2, the experimental protocol including the selection of stimuli and the experiment setup is described in detail. Section 3 provides the signal processing and data analysis methods used in this study. The results of single-trial and single-run classification are given in Section 4. The conclusion of this work follows in Section 5.

2. EXPERIMENTAL PROTOCOL

In this research, we employ the widely used valence-arousal scale proposed by Russell [1980] in order to quantitatively measure emotions. In this scale, each emotional state is placed on a two-dimensional plane with arousal and valence as the horizontal and vertical axes. Arousal is the psychological and physiological state of being awake or reactive to stimuli and ranges from inactive (e.g., uninterested, bored) to active (e.g., alert, excited). Valence represents the intrinsic attractiveness or aversiveness of a stimuli and ranges from unpleasant (e.g., sad, stressed) to pleasant (e.g., happy, elated).

In the following sections, the procedures for test material selection and physiological data acquisition are explained.

[2]http://mmspg.epfl.ch/emotion_dataset/.
[3]Due to licensing issues, we are not able to include the actual videos, but YouTube links are given where possible.

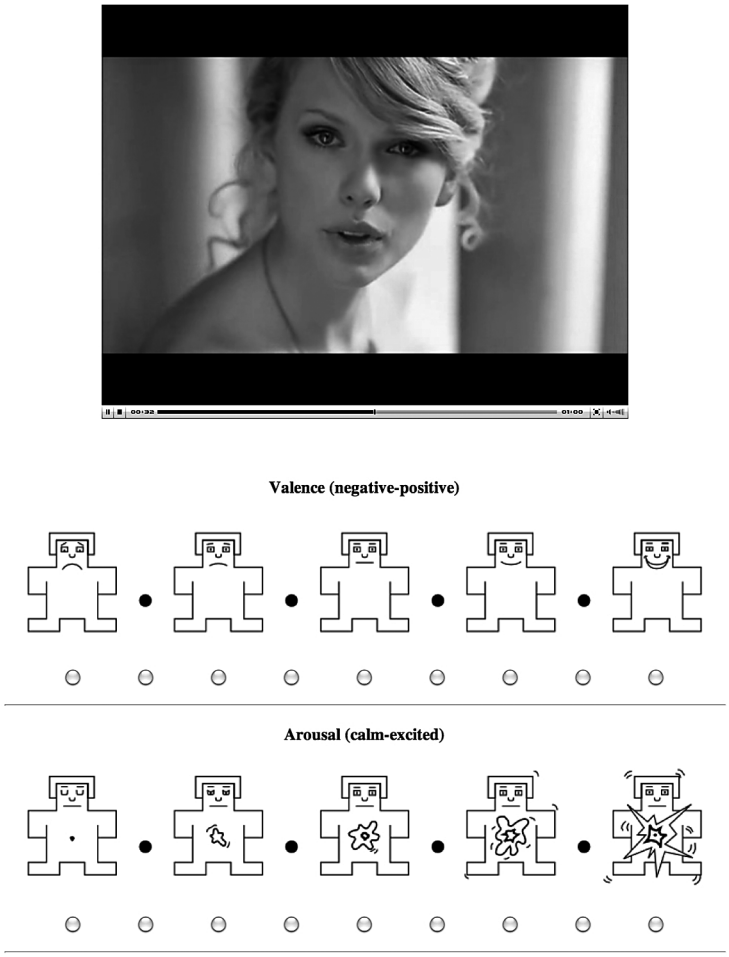

Fig. 2. Screenshot of the Web interface for subjective emotion assessment.

2.1. Stimuli Selection

In this study, music video clips are selected as audiovisual stimuli for the goal of emotion elicitation. Therefore, the first stage of our work was to carefully create an appropriate test set of music video clips. The objective of this selection procedure was to ensure that clips inducing various levels of valence and arousal are included in the final dataset. To this end, 70 candidate music video clips spanning diverse genres, ages, and styles were collected manually. From this collection, the final set of 20 clips were chosen through a Web-based subjective emotion assessment test. During this online subjective test, each subject watched the video clips one by one and were asked to rate the levels of induced emotion in terms of valence and arousal on discrete 9-point scales, as shown in Figure 2. On average, each video clip was rated by 11 subjects.

Figure 3 shows the ratings averaged over the subjects on the valence-arousal plane. Five emotional categories in the plane were considered, from each of which we chose four representative video clips for inclusion in the final test set: positive valence and high arousal (V_+A_+), positive valence and low arousal (V_+A_-), negative valence and high arousal (V_-A_+), negative valence and low arousal (V_-A_-), and neutral categories,

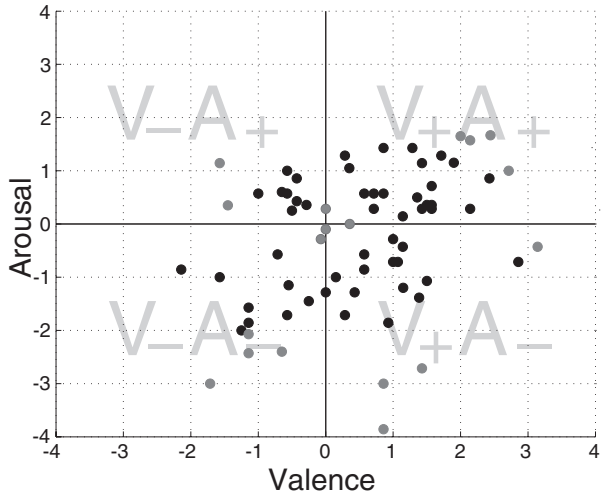

Fig. 3. Subjective test results for selection of video stimuli. The chosen clips are marked with the red color.

which roughly correspond to emotional states of joy, pleasure, anger, sadness, and neutralness, respectively [Kim and André 2008]. For the neutral category, the four video clips whose distances to the origin of the plane are the smallest were chosen. For each of the other four categories, we chose four video clips showing large discriminability (i.e., distance) from the other four categories in the valence-arousal plane. In the V_-A_+ category, however, the clips were located relatively close to the origin of the plane (i.e., the neutral category) and only two video clips could be considered as representative in this category. Thus, each of them was split into two parts and separately used in the experiments. Finally, the first two-minute segments of the selected 20 video clips were used for physiological signal acquisition.

Cohen's kappa was used to examine the inter-rater agreement in the subjective rating results. The binary-quantized ratings (i.e., positive and negative valence or arousal) were used to measure Cohen's kappa for each pair of subjects. For the 70 candidate video clips, the mean and standard deviation values of Cohen's kappa were 0.17 ± 0.19 and 0.13 ± 0.17 for valence and arousal, respectively, whereas they were 0.26 ± 0.34 and 0.26 ± 0.34 for the selected 20 clips. Thus, it is observed that the stimuli selection was performed in a way that the selected clips are the ones showing higher agreement between subjects than others in an overall sense. However, the mean kappa values are still quite small, which demonstrates the inherent difficulty of the affect recognition problem that is dealt with in this article, that is, the induced affect for each person is subjective and varies over subject significantly. In order to properly recognize such subjective emotional responsiveness, the ground truths of the affect recognition experiments presented shortly are the self-assessment ratings that the participants provided after signal acquisition.

2.2. Experimental Setup

The experiments were performed in a laboratory environment with controlled temperature and illumination. EEG and peripheral physiological signals were recorded using a Biosemi ActiveTwo system[4] on a dedicated recording laptop (Pentium M, 1.8 GHz). Stimuli were presented on a dedicated stimulus laptop (P4, 3.2GHz) that sent

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

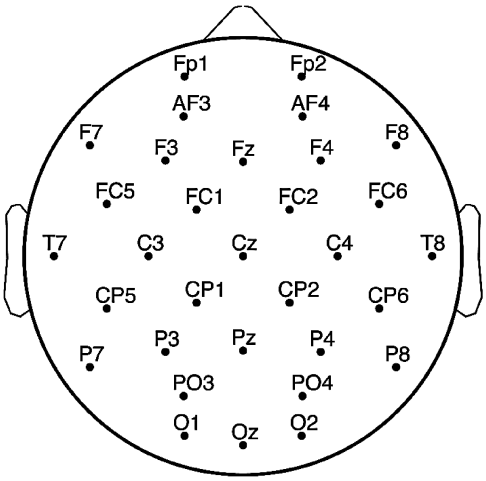

Fig. 4. Electrode configuration (32 channels) used in the experiments.

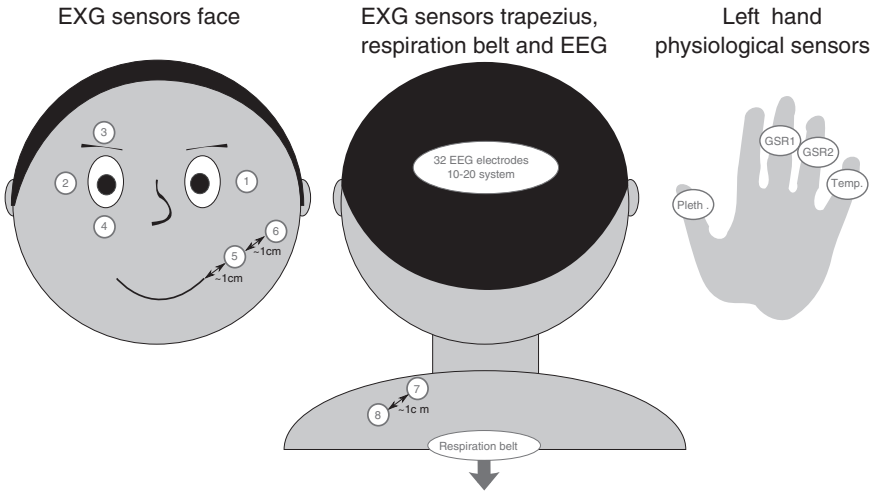

Fig. 5. Placement of peripheral physiological sensors. Four electrodes were used to record EOG (①-④) and 4 for EMG (⑤-⑧). In addition, GSR, blood volume pressure, temperature, and respiration were measured.

synchronization markers directly to the recording PC. For stimulus display and recording the user's ratings, the "Presentation" software by Neurobehavioral systems[5] was used. In order to minimize eye movements, all video stimuli were shown with a width of 640 pixels, filling approximately a quarter of the screen. Thirty-two active AgCl electrodes were placed according to the international 10-20 system (Figure 4), from which the EEG data were recorded at a rate of 512 Hz. At the same time, 13 peripheral physiological signals were also recorded (see Section 3.2).

During the data acquisition experiments, several physiological signals were recorded along with the EEG signals, namely, Galvanic Skin Response (GSR), respiration, skin temperature, blood volume pulse by plethysmograph, EMGs of zygomaticus major and trapezius muscles (2 channels each), and 4-channel electrooculogram (EOG). Figure 5

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

illustrates the electrode placement for acquisition of peripheral physiological signals. All the signals were recorded at a sampling rate of 512 Hz.

GSR, also known as skin conductance, measures the electrical conductance of the skin. It varies with the moisture level controlled by the sympathetic nervous system and, thus, is used to capture the affective state, especially the arousal. It has been shown that a change in the magnitude of GSR and the intensity of the emotional experience are well associated in terms of arousal [Lang 1995]. In our experiments, GSR was measured by placing two electrodes on the distal phalanges of the middle and index fingers.

The breathing activity of the subjects was recorded by using a stretch sensor around their abdomen. The amount of stretch in the rubber band of the sensor is measured as a voltage change. In general, a decreased respiration rate is related to relaxation of a subject, whereas negative emotions can cause irregular respiration patterns and momentary cessation of respiration may be due to surprising events and tense situations.

The skin temperature is known to be related to the emotional state, that is, arousing, negative emotions cause a decrease in temperature, whereas calm, positive emotions tend to increase the temperature [McFarland 1985]. A sensor was attached to the subjects' little fingers to record the skin temperature.

A plethysmograph sensor was positioned on the thumb of a subject in order to measure the blood volume pulse. It has been reported that the heart rate and its variability are subject to change according to the affective state of a subject, for example, anger, fear, and sadness cause increased heart rates [Ekman et al. 1983] and pleasantness increases peak heart rate response [Lang et al. 1993].

Four sensors were used to record the EMG signals. Two of them were placed on the trapezius muscle of the neck to monitor head movements, and the other two were on the zygomaticus major muscle to detect the subject's laughing or smiling.

Finally, the EOG signals were recorded by using four sensors around the eyes of each subject, from which activities related to eye blinking were captured. It is known that the eye blinking rate is related to anxiety.

Six participants (4 males and 2 females), aged between 27 and 35, participated in the experiment. Prior to the experiment, they were given a set of instructions informing them of the experimental protocol and the meaning of the different scales used for self-assessment. more specifically, the instructions included the detailed description of the test, familiarization with different signal artifacts created during head and body movements, and how to assess the induced emotions. An experimenter was also present to answer any question. 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 participant performed a practice run to familiarize himself/herself with the system. In this unrecorded run, a short music video clip was shown, followed by self-assessment by the participant. Figure 6 shows a participant shortly before the start of the experiment.

The experiment started with a two-minute baseline recording, during which a fixation cross was displayed on the monitor and the participant was asked to relax. Then, the 20 selected video clips were presented randomly in 20 separate runs, each run consisting of the following steps:

(1) five-second baseline recording (display of a fixation cross);
(2) two-minute display of the music video; and
(3) self-assessment for arousal, valence, and liking.

At the last step of each run, participants performed a self-assessment of their levels of arousal, valence, and liking. Self-Assessment Manikins (SAM) [Morris 1995] were used to visualize the scales (see Figure 7). For the liking scale, thumbs down/thumbs

Fig. 6. A participant shortly before the experiment.

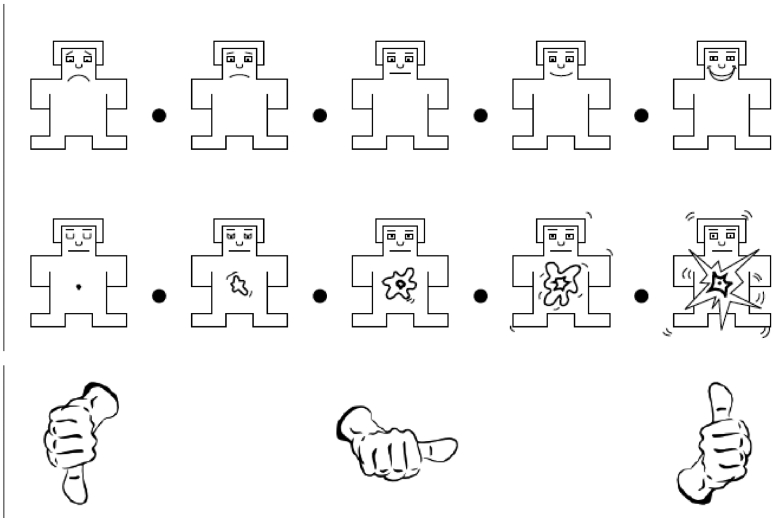

Fig. 7. Images used for self-assessment of arousal, valence, and liking.

up symbols were used. The valence scale ranges from unhappy or sad to happy or joyful. The arousal scale ranges from calm or bored to stimulated or excited. A third scale asks for participants' personal appreciation of the video (that is, how much they liked it). 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.

The manikins were displayed in the center of the screen along with the numbers between 1 and 9 shown below. Participants moved the mouse horizontally below the numbers and clicked to indicate their self-assessment level. Participants were informed that they could click not only exactly on a number but also in between two numbers, making the self-assessment a continuous scale.

Table I. Correlations between the Rating Scales (valence, arousal, and like/dislike) and the Order of the Presentation of Stimuli [Koelstra et al. 2010]

	Valence	Arousal	Like/Dislike	Order
Valence	1	0.46	0.66	−0.24
Arousal	–	1	0.56	−0.17
Like/Dislike	–	–	1	−0.18
Order	–	–	–	1

3. DATA ANALYSIS

In this section, the signal processing methods for feature extraction of EEG and peripheral physiological signals are described.

Preliminary statistical analysis of the acquired signals revealed a relatively strong correlation between EEG sub-bands in some channels and subjective values of valence, arousal, and like/dislike [Koelstra et al. 2010]. Moreover, analysis of subjective ratings was performed to validate the affect induction approach and rule out possible threats to reliability (e.g., due to extreme habituation or fatigue). To this end, the Spearman correlation between the ratings and order of stimuli was computed. Table I presents the results of this analysis.

As can be seen from Table I, a moderate correlation was found between the ratings on the valence, arousal, and like/dislike scales. This can be explained by the fact that most people usually like music video clips which arouse and/or induce positive emotions. However, despite the correlations between the scales, the results suggest that the participants could differentiate between the concepts of valence, arousal, and like/dislike. Furthermore, no significant correlation between the stimulus order and the ratings was observed. This indicates that effects of habituation and fatigue were kept to an acceptable minimum.

3.1. EEG Signal Processing

3.1.1. Preprocessing. Before extracting features from EEG signals and learning a classification function, several preprocessing operations were applied to the recorded data in the order stated next.

(1) *Filtering*. In order to remove the slow drifts and high-frequency noises from the acquired data, a sixth-order Butterworth bandpass filter with the cut-off frequencies of 0.6 Hz and 100 Hz was used. Filtering of the input sequence was performed in both forward and reverse time directions to remove all phase distortion, effectively doubling the filter order.

(2) *Downsampling*. The EEG data was then downsampled from 512 Hz to 256 Hz. The downsampling process filters the input data with a lowpass filter and then resamples the resulting smoothed signal at a lower rate.

(3) *Artifact removal*. The recorded EEG signals are often contaminated with other noncerebral artifactual signals such as eye blinking, eye movements, and muscle movements. These artifacts can potentially cause large amplitude outliers in the EEG and subsequently result in a notable deterioration in classification accuracy. In this article, an orthogonal projection approach was used to remove any potential electrooculogram (EOG) and EMG artifacts from the recorded signal. Let Y denote the recorded EEG signal from a given electrode. Also, let X denote the artifact subspace, a matrix whose columns represent vertical EOG, horizontal EOG, and EMG from the right trapezius, which correspond to eye movement, blinking, and head movement, respectively. These columns are linearly independent vectors in $\mathbb{R}^n$ with n the number of samples. The resultant artifactual component N that modulates Y could be obtained by orthogonal

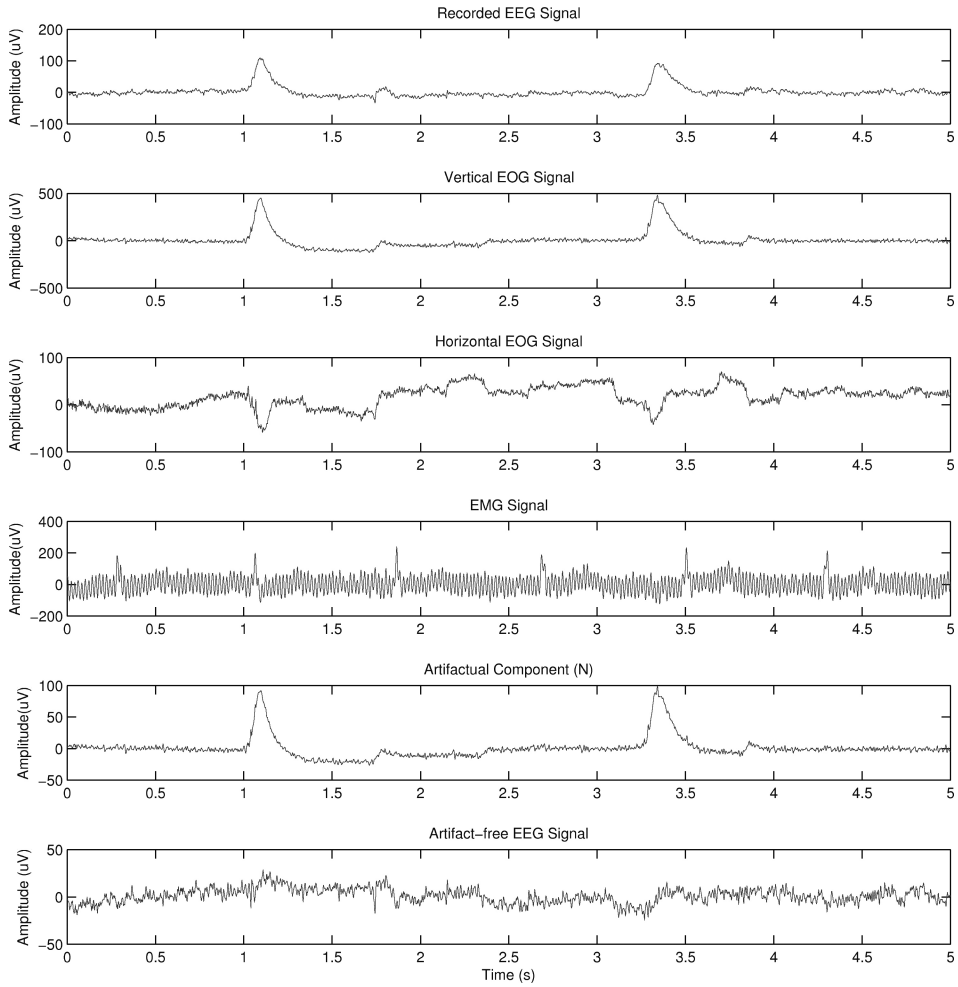

Fig. 8. Recorded EEG signal from FP1 channel containing EOG and EMG artifacts and the resultant artifact-free signal after artifact removal.

projection of Y on the X subspace.

$$N = X(X^T X)^{-1} X^T Y \tag{1}$$

The matrix $A = X(X^T X)^{-1} X^T$ is called the projection matrix for the artifact subspace. Thus, the artifact-free EEG ($\tilde{Y}$) signal can be obtained by subtracting N from Y. We have

$$\tilde{Y} = Y - X(X^T X)^{-1} X^T Y \tag{2}$$

or

$$\tilde{Y} = WY, \qquad W = (I_n - X(X^T X)^{-1} X^T) \tag{3}$$

where I_n is the identity matrix of size n and W is called the artifact removal matrix. Figure 8 illustrates the results of this artifact removal for an example window of a contaminated data. As can be seen in this figure, the artifactual component N was computed from contaminated EEG signal and artifact subspace X through Eq. (1).

This approach was repeated for all electrodes and the resultant artifact-free signal was used for further signal analysis.

3.1.2. Feature Extraction. After preprocessing, the EEG signal was broken down into distinct segments such that each segment represented the EEG signal acquired during each run (refer to Section 2.2 for the definition of run). More precisely, each segment began with the presentation onset of its corresponding video clip and lasted for the 120 seconds of video presentation. In order to extract descriptive and discriminating features from the EEG signal, the wavelet transform was used. This transform represents the signal in both time and frequency and provides precise information about transient events occurring in the signal. It is known effective for representing various aspects of signals such as trends, discontinuities, and repeated patterns when other signal processing approaches fail or are not as effective [Kolev et al. 1997]. Furthermore, the wavelet time-frequency representation does not make any assumptions about signal stationarity and is capable of detecting dynamic changes due to its localization properties. Moreover, the computational time is relatively short since a fast wavelet transform in a multi-resolution framework can be used. A wavelet family $\Psi_{a,b}$ is a set of elementary functions generated by dilation and translation of a mother wavelet $\Psi(t)$:

$$\Psi_{a,b}(t) = |a|^{\frac{-1}{2}} \Psi\left(\frac{t-b}{a}\right) \quad \text{such that} \quad a, b \in \mathbb{R}, a \neq 0. \tag{4}$$

Here, a is the scaling parameter that stretches or shrinks the mother wavelet and b is the translation parameter that moves the mother wavelet to different time positions at any scale without changing its shape. The Continuous Wavelet Transform (CWT) of a signal is defined as the correlation between the signal and the wavelet family $\Psi_{a,b}$. In a special case where a and b take only discrete values $a_j = 2^{-j}$ and $b_{j,k} = 2^{-j}k$ with $j, k \in \mathbb{Z}$, an orthonormal Hilbert space $L^2(\mathbb{R})$ consisting of finite energy signals can be constituted by the following family.

$$\Psi_{j,k}(t) = 2^{j/2}\Psi(2^j t - k) \tag{5}$$

Given a discrete time stochastic process S with length M, the associated Discrete Wavelet Transform (DWT) coefficients can be obtained through

$$C_j(k) = \langle S, 2^{j/2}\psi(2^j t - k)\rangle, \tag{6}$$

where $j = -m, \ldots, -1$ and $m = \log_2 M$. This correlation gives information on the signal at time $2^{-j}k$ and scale 2^{-j}. The number of wavelet coefficients obtained at each decomposition level equals $2^j M$.

The energy at each decomposition level can be computed using the coefficients of the detail signal

$$E_j = \sum_k |C_j(k))|^2 \tag{7}$$

and consequently the total energy of S can be obtained by

$$E_{total} = \|S\|^2 = \sum_{j<0}\sum_k |C_j(k))|^2 = \sum_{j<0} E_j. \tag{8}$$

Thus, the relative wavelet energy for each decomposition level can be defined as

$$p_j = \frac{E_j}{E_{total}}. \tag{9}$$

Clearly, the sum of relative wavelet energies of all decomposition levels equals to one ($\sum_j p_j = 1$) and their distribution $\{p_j\}$ can be considered as a time-scale density.

The concept of Shannon's entropy [Shannon 2001], which plays a central role in information theory, is sometimes used as measure of uncertainty. The entropy of a random variable is defined in terms of its probability distribution and can be shown to be a good measure of randomness or uncertainty. The total Wavelet Entropy (WE) [Blanco et al. 1998] of S can be defined as

$$H_{WT}(p) = -\sum_{j<0} p_j \ln p_j. \tag{10}$$

This metric provides information about the degree of order/disorder of the random process S. One can easily observe that in the case of a totally random process, which represents a very disordered signal, the wavelet coefficients for all decomposition levels will have significant values and will be of the same order. In this case, the relative wavelet energies of all decomposition levels are almost equal and the total WE will take its maximum value. On the contrary, if S is a very ordered process (e.g., a signal with a very narrow band spectrum), it will have its major contribution in only one decomposition level and produces insignificant wavelet coefficients for other decomposition levels. Consequently, the relative wavelet energy values of all decomposition levels will almost be equal to zero, except for the level which contains all the energy of the signal. The value of the total WE in this case will be close to zero. Therefore, for a given signal S, the value of WE quantifies the degree of order of the signal and the lower this value, the greater the level of information in the signal.

Additionally, the Kullback-Leibler divergence, also known as Kullback-Leibler relative entropy, can be used to define the Relative Wavelet Entropy (RWE). This relative entropy is a measure of dissimilarity between two probability distributions $\{p_j\}$ and $\{q_j\}$ and is defined as

$$H_{WT}(p \parallel q) = \sum_{j<0} p_j \ln \frac{p_j}{q_j}\ . \tag{11}$$

RWE is defined for two probability distributions (more precisely, RWE of distribution $\{p_j\}$ with respect to distribution $\{q_j\}$) and the more different the two distributions are, the higher is the value of RWE . In the extreme case of $p_j \equiv q_j$ (perfectly equal distributions), the value of RWE is zero.

Relative wavelet energy has been used as a feature extraction technique in some EEG processing applications such as studies on epilepsy [Guo et al. 2009] and WE has been studied before in some EEG processing applications such as the analysis of grand averaged Event-Related Potentials (ERPs) and EEG changes during closed eye and open eye states [Yordanova et al. 2002; Rosso et al. 2001]. In this study, the relative wavelet energies of each electrode together with RWE of symmetrical electrode pairs are extracted as features. The latter feature can be considered as a novel asymmetry index for detecting any asymmetry in the distribution of energy over the right and left brain hemispheres due to a change of emotional state.

3.2. Peripheral Physiological Signal Processing

The physiological signals were recorded at a sampling rate of 512 Hz. Typical one-minute examples of the acquired raw signals are shown in Figure 9. From the signals, features for recognition were extracted as follows.

First, the signals were partitioned into multiple temporal segments. Physiological changes caused by affective states are usually observed on a longer term in comparison to the changes in EEG signals [Picard et al. 2001]. In order to capture such changes appropriately, relatively long moving windows were used for segmentation. We tested the recognition performance for various combinations of window parameters (i.e., length

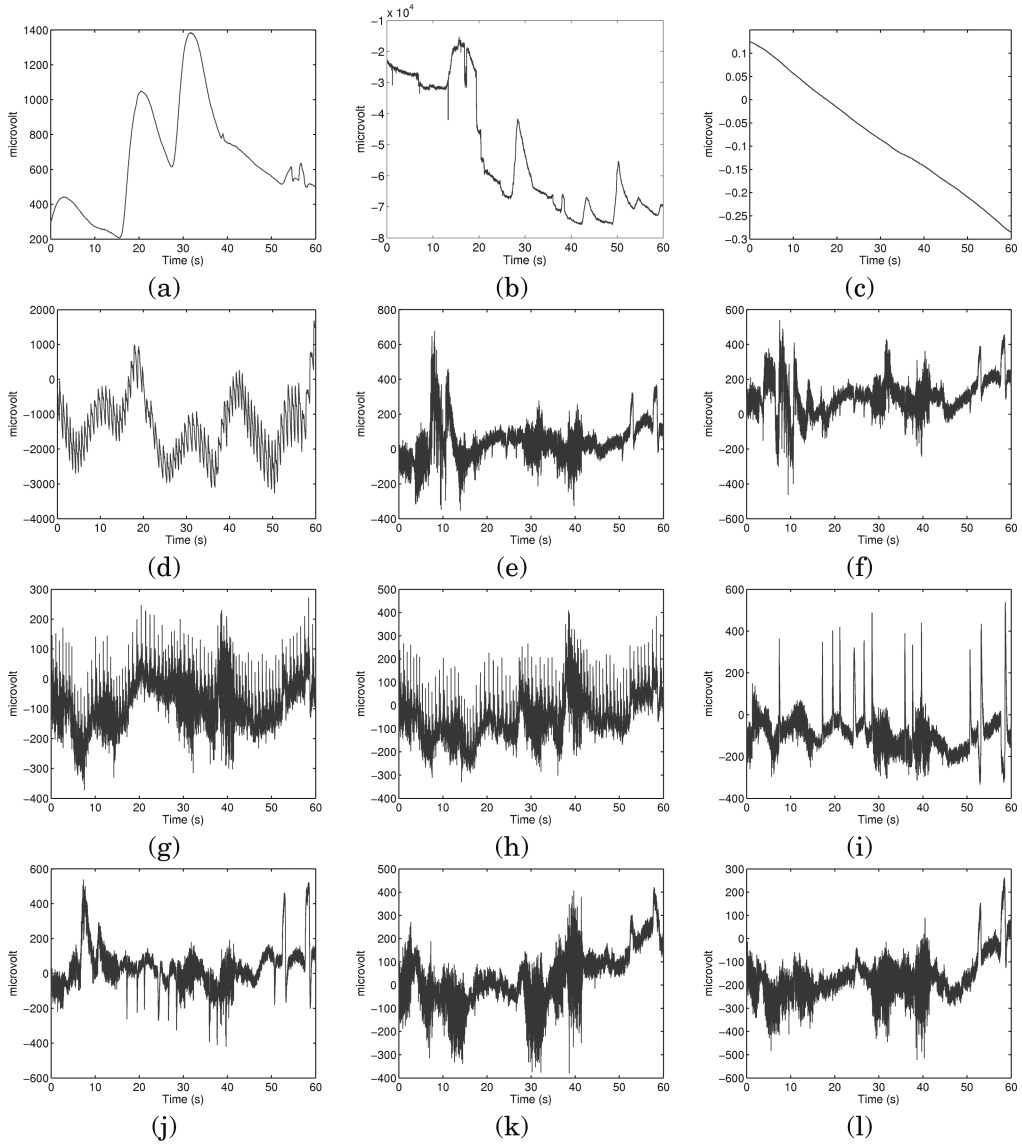

Fig. 9. Typical examples of the recorded physiological signals: (a) GSR; (b) respiration; (c) skin temperature; (d) blood volume pressure; (e)-(f) EMG on the zygomaticus major muscle; (g)-(h) EMG on the trapezius muscle; (i)-(l) EOG.

and moving rate): the length varied from 10 to 120 seconds and the moving rate ranged between 5 and 30 seconds. The best performance was observed by using a window having a length of 60 seconds and moving at every 15 seconds (so that neighboring segments have a 45-seconds overlap), which produced five segments for each music video clip.

Then, for each segment, 30-dimensional features were extracted. We extracted the mean and standard deviation of each signal. In addition, the respiration rate, heart beat rate, and eye blinking rate were estimated from the breathing activity, blood

Table II. Different Decomposition Levels and Their Corresponding Frequency Ranges

Decomposition Level	Frequency Range (Hz)	Corresponding EEG Subband
D1	64–100	–
D2	32–64	Gamma
D3	16–32	Beta
D4	8–16	Alpha
D5	4–8	Theta
A5	0.6–4	Delta

volume pulse, and vertical EOG signals, respectively, and their mean and standard deviation values were used as features.

4. RESULTS AND DISCUSSION

Three different binary classification problems were posed: positive/negative valence, low/high arousal, and low/high liking. To this end, the participants' ratings by self-assessment during the experiment were used as the ground truth. The ratings for each of these scales were thresholded into two classes (low/high or negative/positive). On the 9-point rating scales, the threshold was simply placed in the middle. It should be noted that participant 1 had to to be excluded for the arousal and liking targets, as this participant assigned high arousal rating to 17 out of 20 and high liking ratings to 19 out of 20 video clips. As a result, we did not have enough samples to train the classifier for low arousal and low liking ratings for this participant. All other participants rated the videos in a more balanced manner.

In this section, the classification results of EEG and peripheral physiological signals using the methodology detailed in the previous section are presented, compared, and discussed. To this end, the classification results are reported for both single-trial classification and single-run classification schemes. The former scheme refers to classification of short segments of EEG and peripheral physiological signals, whereas the latter scheme investigates the number of music video clips for which the measures of arousal, valence, and like/dislike would be correctly classified based on the processing of EEG and peripheral physiological signals. Furthermore, the study results of the generalization possibility of the proposed approach to other subjects will be presented.

In order to evaluate the classification performance, a leave-one-trial-out cross-validation approach is utilized. More precisely, as mentioned in Section 2.2, for each participant, 20 runs (trials) were acquired and each run was broken down into 24 single trials. In order to validate the classification performance, one run was left out as testing set and a classifier was trained only based on single trials of the remaining (19) runs. This was repeated until each run is taken out as a test run once. The rationale for this way of cross-validation is that the single trials of one given run are not used both as training and test samples in the cross-validation.

4.1. Single-Trial Classification

Prior to feature extraction and classification, the EEG signal for each run was broken down into nonoverlapping windows of 5-seconds long. This window length was selected after a search for different integer window lengths between four and ten. These windows are hereinafter referred as EEG single trials. Therefore, for each run, a total number of 24 single trials were created. DWT using the orthogonal cubic spline mother wavelet was then applied to the data of each single trial to decompose the EEG signal into different sub-bands. Among several alternatives, cubic spline functions were used as they are symmetric and orthogonal, and combine smoothness with numerical advantages [Unser 2002]. With five levels of decomposition ($j = -5, \ldots, -1$), the coefficients for the theta, alpha, beta, and gamma sub-bands (refer to Table II) was achieved. Figure 10 illustrates a typical EEG single trial and its corresponding sub-bands.

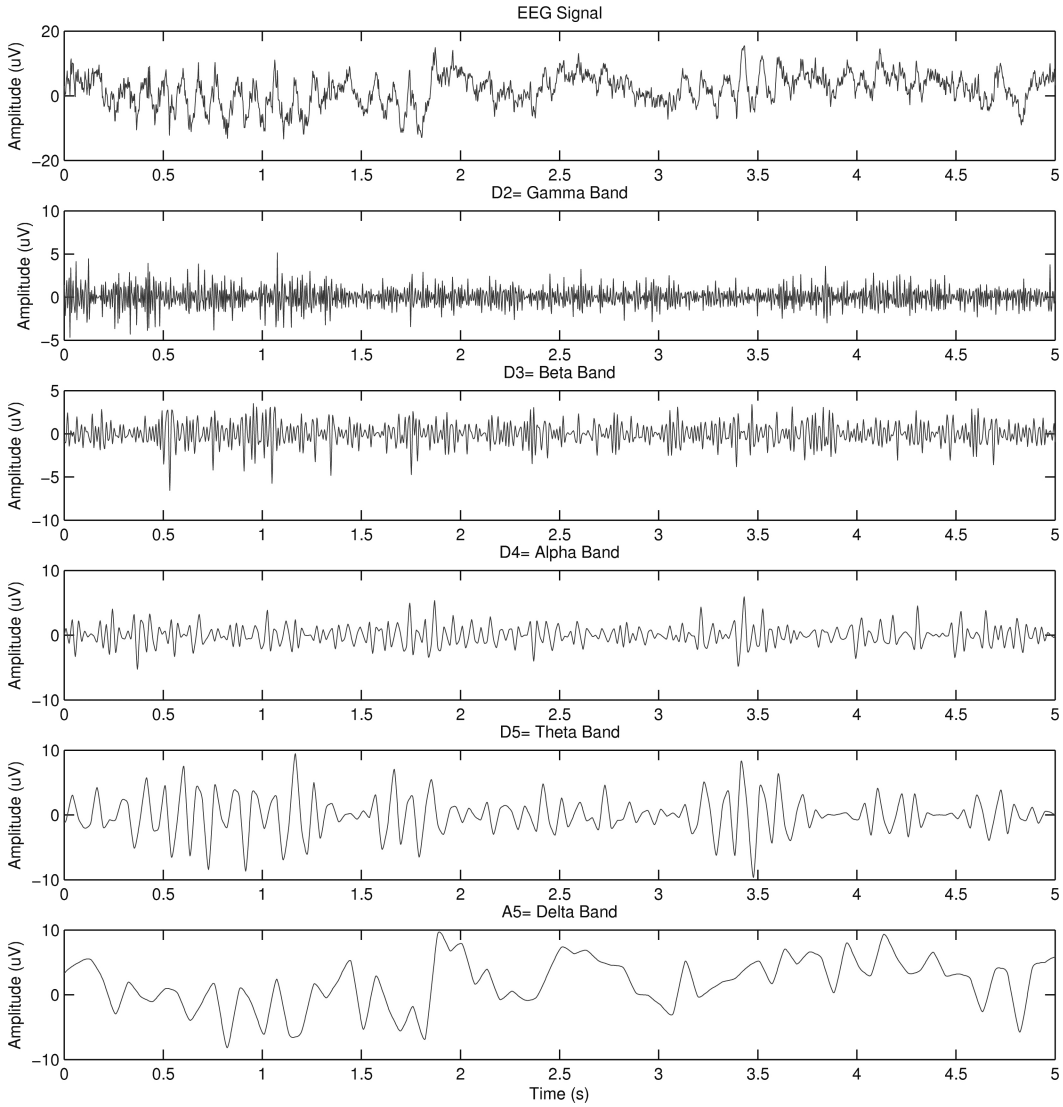

Fig. 10. A typical EEG single trial from the Pz channel and its sub-bands.

To construct a feature vector from each single trial, the values of relative energies for each electrode were computed separately and concatenated together. Furthermore, a review of over 70 published studies on the relationship between emotion or emotion-related constructs and asymmetries in EEG activity over the frontal cortex suggests asymmetries in frontal EEG activity in response to emotional stimuli and changes in emotional state [Coan and Allen 2004]. In order to take into account the possible asymmetry in the brain activities due to emotional stimuli, a novel asymmetry index based on relative wavelet entropy of symmetrical pairs of electrodes on the right and left brain hemispheres was computed using Eq. (11). If the relative wavelet energy (probability) distributions of a given symmetrical pair of electrodes are exactly equal ($p_i = q_i,\ i = \{gamma, beta, alpha, theta, delta\}$), then the value of RWE of this pair is

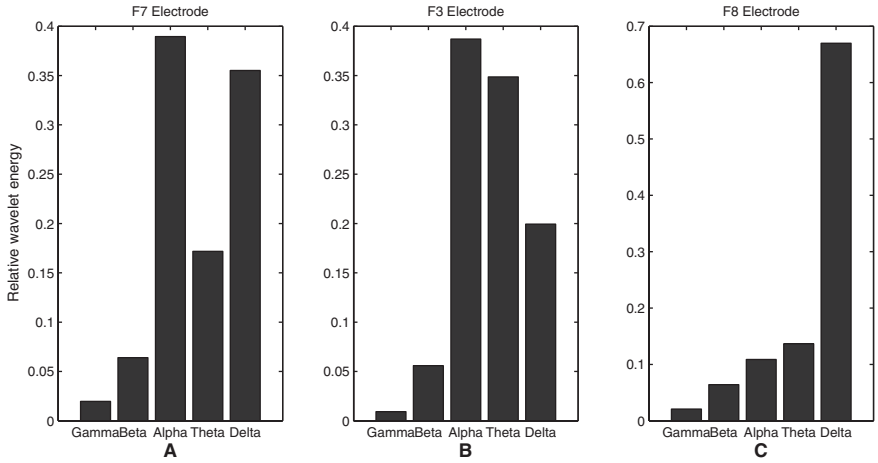

Fig. 11. Relative wavelet energy distributions corresponding to different EEG sub-bands. Distribution A, $\{p_i\} = \{0.02, 0.06, 0.39, 0.17, 0.36\}$; B, $\{p_i\} = \{0.01, 0.06, 0.39, 0.35, 0.20\}$; C, $\{p_i\} = \{0.02, 0.06, 0.11, 0.14, 0.67\}$. The wavelet entropy values for these distributions are $H_{WT}(A) = 1.86$, $H_{WT}(B) = 1.82$, and $H_{WT}(C) = 1.50$ and the RWE values are $H_{WT}(A \parallel B) = 0.15$ and $H_{WT}(A \parallel C) = 0.44$.

equal to zero. The larger the value of RWE, the more asymmetric the EEG activity for a given electrode pair.

Figure 11 presents the relative wavelet energy distributions corresponding to five wavelet resolution levels for three different electrodes (F7, F3, and F8) for the case of a music video clip which induces positive valence. As can be seen, the relative wavelet energy distributions of channels F7 and F8 are very different, and hence the corresponding value of asymmetry index is relatively high. In the case of electrode pair F7 and F3 both located on the left hemisphere of the scalp, the relative wavelet energy distributions are less different and consequently the value of the asymmetry index is lower.

The RWE values of all symmetrical pairs of electrodes on the right and left hemisphere were extracted and integrated with the feature vector created using values of relative energies. Thus, the dimensionality of the feature vectors was $(N_e \times N_s) + N_a$, where $N_e = 32$ denotes the number of electrodes, $N_s = 5$ denotes the number of sub-bands of each EEG signal, and $N_a = 14$ denotes the number of symmetrical electrode pairs. Consequently, the total dimension of a feature vector extracted from a given EEG single trial was $(32 \times 5) + 14 = 174$.

A Support Vector Machine (SVM) classifier with radial basis function kernels was used for the classification of single trials. In order to perform the leave-one-trial-out cross-validation, all single trials of a given run were left out and the SVM was trained using the remaining single trials from different runs, which was repeated for each run. Therefore, for testing each run, the training was performed using 456 trials (24 single trial $\times$ 19 runs). Table III presents the results of this single-trial classification. It can be seen that average accuracies of 69.58%, 73.66%, and 70.25% were achieved for classification of valence, arousal, and like/dislike, respectively.

In the case of the physiological signals, the signals were segmented by using a moving window of 60 seconds with 15-second shifts, as explained in Section 3.2. These segments are referred as peripheral physiological single trials. From each segment, a 30-dimensional feature vector was extracted.

The relative importance of each feature among the 30 extracted was examined as follows. The discriminating power of each feature for each target type (valence, arousal,

Table III. Two Class Single-Trial EEG Classification Accuracy Values for the Valence, Arousal, and Like/Dislike Targets

	P1	P2	P3	P4	P5	P6	Avg.
Valence	63.3	67.1	70.0	58.7	77.5	80.8	69.6
Arousal	–	63.7	75.8	78.7	77.9	72.1	73.7
Like/Dislike	–	70.8	73.3	75.8	69.2	62.1	70.2

Table IV. Two Class Single-Trial Peripheral Physiological Signals Classification Accuracy Values for the Valence, Arousal, and Like/Dislike Targets

	P1	P2	P3	P4	P5	P6	Avg.
Valence	45	53	53	54	62	67	55.7
Arousal	–	45	64	60	40	61	54
Like/Dislike	–	71	67	74	57	64	66.6

The accuracy values are from 5×20 single trials for all 20 leave-one-trial-out sessions for each participant, thus the precision of the values is different from that shown in Table III.

or like/dislike) was defined as

$$d = \frac{(m_+ - m_-)^2}{s_+^2 + s_-^2}, \tag{12}$$

where m_+ and m_- denote the mean feature values for the positive and negative (or high and low) classes, respectively, and s_+^2 and s_-^2 the variances of the feature values for the two classes, respectively. A large value of d indicates that the corresponding feature induces a large distance between the centers of the samples of the two classes and small deviations in the distributions of the two sample sets, and hence leads to better separability. Figure 12 compares the average discriminating powers of the extracted features over all leave-one-trial-out sessions. It is observed that the feature components showing high discriminating powers vary according to the target type. For valence, the 11th and 13th components (i.e., the mean heart beat interval and the mean value of a zygomaticus major EMG signal) have the highest discriminating powers; the 8th, 15th, and 29th components (i.e., the standard deviation of the skin temperature, the mean value of a zygomaticus major EMG signal, and the mean eye blinking rate) are effective for arousal; for like/dislike, the 3rd and 22nd components (i.e., the mean value of the respiration signal and the standard deviation of an EOG signal) show relatively high discriminating powers.

As in the case of EEG, an SVM classifier with radial basis function kernels was used for single-trial classification of the physiological signals. The protocol of the leave-one-trial-out cross-validation remained the same except that, due to the different number of segments for peripheral physiological signals (5 segments for each run), 95 samples from 19 trials were used for training for each test run.

Table IV shows the results of the physiological signal-based single-trial classification. On average, the accuracies obtained for valence, arousal, and like/dislike were 55.7%, 54%, and 66.6%, respectively. Unlike the case of EEG, the performance was the best for like/dislike classification, which will be discussed further in Section 4.3. In addition, it should be noted that the performance of the EEG-based and physiological-signal-based systems in Tables III and IV are not directly comparable because the signal segmentation (single-trial extraction) was performed using different window lengths and thus the performances were measured using different numbers of samples.

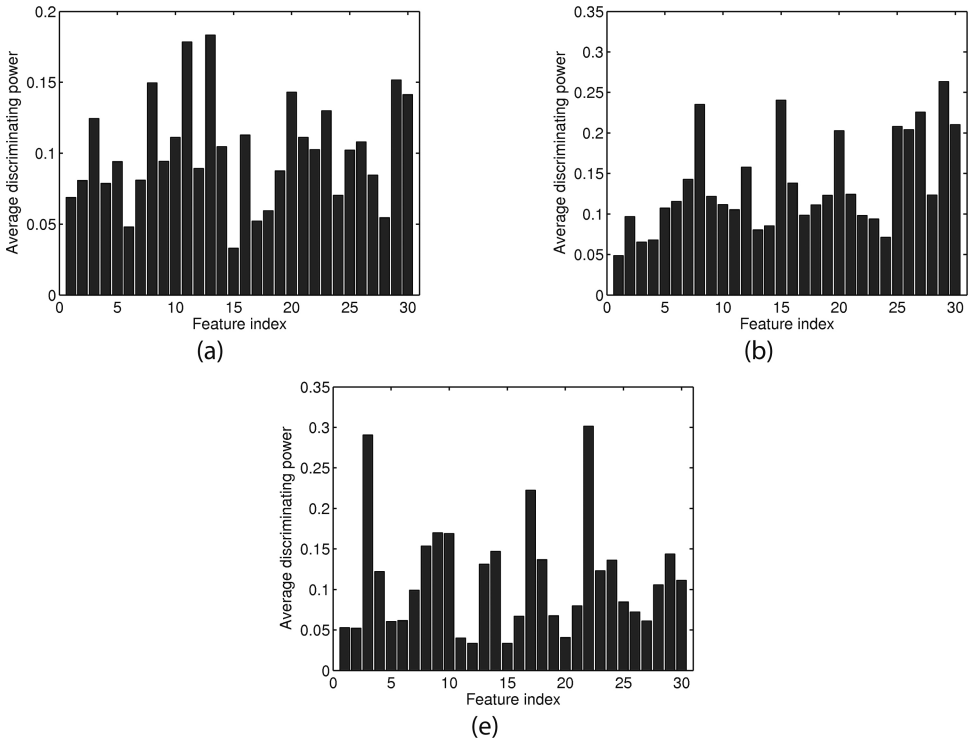

Fig. 12. Discriminating power of each feature extracted from the physiological signals for (a) valence, (b) arousal, and (c) like/dislike.

4.2. Single-Run Classification

The rationale for this classification scheme can be explained as follows. In fact, the EEG and peripheral physiological signal single trials are the segments of the data acquired for a given run. These single trials were formed in order that more training samples are available for learning a classification function. Emotional states, however, do not necessarily change on the time scale of single trials. In other words, at the end of each run the participants report their self-assessments and all the single trials of the run are labeled with these self-assessments accordingly. Therefore, there may exist several single trials during which the emotional states did not change much when compared with that of the baseline, while they are labeled with participants' self-assessments of the whole run. These imperfectly labeled single trials cause deterioration in classification accuracy.

In order to perform the single-run classification of EEG and peripheral physiological signals, the results of single-trial classification of these signals were aggregated as follows. For a given video, the rate of correctly classified single trials ($r = \frac{n_c}{n_t}$) was computed, where n_c denotes the number of correctly classified single trials and n_t denotes the total number of the single trials of a run. We assume the signals acquired during presentation of a run are correctly classified if $r > 0.5$, which shows that more than half of the single trials of the run were correctly classified. Table V presents the results of this classification scheme for both EEG and peripheral signals. For instance, the valence classification results for participant 5 indicate that for 80% of the video clips (16 out of 20 video clips) the EEG-based classified valence values match the

Table V. Results of Two Class Single-Run Classification of EEG and Peripheral Physiological Signals for the Valence, Arousal, and Like/Dislike Targets

	P1		P2		P3		P4		P5		P6		Avg.	
	EEG	Phy.	EEG	Phy.	EEG	Phy.	EEG	Phy.	EEG	Phy.	EEG	Phy.	EEG	Phy.
Valence	65	50	60	55	70	50	60	50	80	70	80	70	69.2	57.5
Arousal	–	–	60	45	80	65	85	55	95	35	90	65	82	53
Like/Dislike	–	–	70	75	80	65	75	75	75	55	70	70	74	68

Table VI. Results of Subject-Independent Two Class Single-Trial Classification of EEG and Peripheral Physiological Signals for the Valence, Arousal, and Like/Dislike Targets

	P1		P2		P3		P4		P5		P6		Avg.	
	EEG	Phy.	EEG	Phy.	EEG	Phy.	EEG	Phy.	EEG	Phy.	EEG	Phy.	EEG	Phy.
Valence	68.3	39	62.9	61	60.8	60	56.2	56	49.2	41	74.6	32	62	48.2
Arousal	58.3	52	57.5	51	63.7	52	43.7	32	49.6	52	58.3	53	55.2	48.7
Like/Dislike	63.7	57	61.7	45	68.3	56	65.4	69	55.8	60	57.5	55	62.1	57

Table VII. Results of Subject-Independent Two Class Single-Run Classification of EEG and Peripheral Physiological Signals for the Valence, Arousal, and Like/Dislike Targets

	P1		P2		P3		P4		P5		P6		Avg.	
	EEG	Phy.	EEG	Phy.	EEG	Phy.	EEG	Phy.	EEG	Phy.	EEG	Phy.	EEG	Phy.
Valence	75	40	60	60	60	60	55	60	45	35	75	30	61.7	48.2
Arousal	55	55	55	50	70	55	35	35	50	55	55	55	53.3	50.8
Like/Dislike	70	65	60	40	65	65	65	65	50	60	55	55	60.7	58.3

self-assessments and for 70% of the video clips (14 out of 20 video clips) the peripheral-physiological-signals-based classified valence match the self-assessment.

4.3. Generalization

In order to explore the feasibility of developing a general-purpose affect recognition system, which can be trained using data acquired from limited number of participants and can be applied for a naive subject who never used the system before, a leave-one-participant-out cross-validation scheme was used. In other words, a classifier was trained based on physiological signals acquired from five participants and was tested on the data acquired from the remaining one subject. This was repeated until each participant was a test participant once. The results of single-trial and single-run classification using this approach is presented in Table VI and TableVII, respectively.

As can be seen in these tables, for each subject relatively lower correct classification rates will be achieved when his/her physiological signals are not used in the training phase. However, the obtained results for most of the participants suggest that EEG signals can potentially be used to develop a subject-independent affect recognition and peripheral physiological signals can be used for determining the like/dislike measure.

4.4. Discussion

The results of single-trial EEG classification show that emotional valence, arousal, and like/dislike can all be classified with acceptable accuracies. It can be also inferred that on average, arousal can be estimated with higher accuracy in comparison with valence and like/dislike. However, this is not valid for all participants. For instance, for participant 6 the accuracy for valence is much higher than that for arousal and like/dislike. Furthermore, for all valence, arousal, and like/dislike measures, the classification accuracy varies among participants, which indicates that the classification performance is participant-dependent and it will be a challenging problem to build a general classifier for all participants.

The single-trial classification results using the peripheral physiological signals show that like/dislike classification can be done with the highest accuracy among the three target types, which is quite close to the accuracy produced using the EEG signal. As in the case of EEG, the results are highly participant-dependent; depending on the participant, the accuracies for the three targets vary by roughly $\pm 10\%$ around the mean values. The highest accuracies were obtained with participant 6 for valence, participant 3 for arousal, and participant 4 for like/dislike.

When the EEG-based and peripheral physiological signal-based single-trial classification results are compared, it may be observed that the EEG-based classification generally performs better than the physiological-signal-based one. We conducted the Wilcoxon rank sum test [Wilcoxon 1945] in order to compare their subject-wise performance statistically, under the null hypothesis that the classification accuracy values for the two modalities for each participant are independent samples from identical distributions with equal medians, against the alternative that they do not have equal medians. The test results showed that the performance difference for classification of valence and arousal is statistically significant, while no significant difference was found for liking/disliking, both at a significance level of 5%. Thus, it can be concluded that, from the viewpoint of the individual participant performance, the EEG-based classification outperforms the physiological signal-based one only for valence and arousal, although the performance on average for like/dislike is better in the former than the latter.

The results of single-run classification of EEG signals show that on average, for almost 70% of the music video clips, the valence and like/dislike measures can be correctly extracted. An accuracy of 80% was obtained for the case of arousal classification. Similar to the EEG single-trial classification, it can be seen that on average, arousal can be classified with higher accuracy when compared with valence and like/dislike. This is valid for all participants, except for participant 2. Furthermore, for participants 5 and 6 , significant classification accuracies of 95% and 90% were achieved.

The best performance for the peripheral physiological signal-based single-run classification was obtained for like/dislike (68% in Table V). The highest classification accuracies were obtained with participants 5 and 6 for valence, participants 3 and 6 for arousal, and participants 2 and 4 for like/dislike, which shows highly participant-dependent performance, as in the case of single-trial classification.

Overall, it is observed that the EEG-signal-based single-run classification outperforms the peripheral physiological signal based one for all three targets. However, the Wilcoxon rank sum tests revealed that the difference of their participant-wise performance was statistically significant only for valence at a significance level of 5%. Again, this implies that, although the EEG-based classification produces higher accuracies than the peripheral physiological signal-based one in an overall sense, the former outperforms the latter only for classification of valence in terms of participant-wise performance.

In order to further compare the single-run classification performance of the EEG and peripheral physiological signals, their performance for all leave-one-trial-out experiments is examined in terms of the contingency matrix for each target, as shown in Figure 13. The four numbers in each matrix stand for the numbers of video clips that the two signal modalities (EEG and peripheral physiological signals) classify correctly, one signal modality classifies correctly but the other one incorrectly, and both classify incorrectly [Gillick and Cox 1989]. We applied McNemar's test [McNemar 1947] to statistically compare the two modalities with the null hypothesis that, given that only one of the two modalities makes an error, it is equally likely to be either one, which means that one is not better than the other one. The test results in Figure 13 show that the difference in performances of the two classifiers is significant only for arousal

		Physiological	
		Incorrect	*Correct*
EEG	*Incorrect*	20	17
	Correct	31	52

(a)

		Physiological	
		Incorrect	*Correct*
EEG	*Incorrect*	15	3
	Correct	32	50

(b)

		Physiological	
		Incorrect	*Correct*
EEG	*Incorrect*	11	15
	Correct	21	53

(c)

Fig. 13. Contingency matrices comparing the EEG and physiological signal-based classification systems for (a) valence, (b) arousal, and (c) like/dislike.

at a significance level of 5%, which is due to the small number of samples that were incorrectly classified by using the EEG signals but correctly classified by using the physiological signals (Figure 13(b)). This implies that the EEG-signal-based classification clearly outperforms the physiological counterpart for arousal, which can be also seen in Table V (82% versus 53%). On the other hand, the two modalities have complementary characteristics for valence and like/dislike, which is observed from relatively large values in the off-diagonal entries in Figure 13(a) and (c). Such characteristics can be exploited for constructing an integrated classification system using the two modalities simultaneously in order to obtain improved performance.

From the single-trial and single-run classification results, participant-wise consistency in the two classification systems is also observed. In other words, a participant showing high accuracies for EEG-signal-based classification tends to show high accuracies for physiological-signal-based classification. This may come from various participant-dependent effects, such as differences in the participants' concentration or emotional sensitivity.

Finally, it is worth mentioning that there exists a trade-off between measurement complexity and accuracy when the two signal modalities are compared. While the EEG signal shows better performance than the physiological signals, the latter have an advantage in terms of sensor placement, set-up time, and comfort. Therefore, considering that the accuracies of the two modalities are comparable for like/dislike in our experiments, using peripheral physiological signals can be a good alternative to estimate a user's emotional preference for a given music video clip at reduced complexity and discomfort.

5. CONCLUSION

In this work, we have studied the potential of EEG and peripheral physiological signals for the aim of analysis of music video clip-based induced emotions. We have presented

a database for the analysis of spontaneous emotions containing EEG and peripheral physiological signals recorded from six participants, in which each participant watched and rated their emotional response to 20 music videos along the scales of arousal and valence, as well as their liking of the videos. This database is made publicly available and it is our hope that other researchers will try their methods and algorithms on this highly challenging database.

A novel EEG feature extraction technique based on RWE was introduced to capture the asymmetric energy distribution of EEG signals over the right and left hemispheres of the brain. Single-trial and single-run classification schemes were performed for the scales of arousal, valence, and liking by using features extracted from the EEG and peripheral physiological signals. Furthermore, the possibility of generalizing the proposed methodology in this article, to new naive subjects was assessed. The classification results confirmed that there is a relatively abundant amount of information in both EEG and peripheral physiological signals regarding users' emotional states. These results are very promising, however the performance of the proposed methodology should be tested with more participants. Thus, we hold the work described in this article as only the beginning of a large project. We intend to develop a more extensive video clip database so that they can elicit stronger and more diverse emotions in participants and thus increase the accuracy of the emotion recognition techniques. A larger number of participants can assess the generality of the results obtained in this work and the feasibility of developing a general classifier, which can work for any naive participant without any training sessions. Other modalities such as multimedia content features would also be used to extract the emotional information of the test stimuli. Last but not least, we plan to fuse the peripheral physiological and EEG modalities in order to better exploit the relative strengths of each modality.

ACKNOWLEDGMENTS

The authors would like to thank Christian Muehl, Mohammad Soleymani, and Sander Koelstra for their help and participation in data acquisition.

REFERENCES

AFTANAS, L., REVA, N., VARLAMOV, A., PAVLOV, S., AND MAKHNEV, V. 2004. Analysis of evoked EEG synchronization and desynchronization in conditions of emotional activation in humans: Temporal and topographic characteristics. *Neurosci. Behav. Physiol. 34,* 8, 859–867.

BLANCO, S., FIGLIOLA, A., QUIROGA, R., ROSSO, O., AND SERRANO, E. 1998. Time-Frequency analysis of electroencephalogram series. III. Wavelet packets and information cost function. *Phys. Rev. E 57,* 1, 932–940.

CHANEL, G., KIERKELS, J., SOLEYMANI, M., AND PUN, T. 2009. Short-term emotion assessment in a recall paradigm. *Int. J. Hum.-Comput. Stud. 67,* 8, 607–627.

CHANEL, G., KRONEGG, J., GRANDJEAN, D., AND PUN, T. 2006. Emotion assessment: Arousal evaluation using EEG's and peripheral physiological signals. In *Proceedings of the Conference on Multimedia Content Representation, Classification and Security*. 530–537.

COAN, J. AND ALLEN, J. 2004. Frontal EEG asymmetry as a moderator and mediator of emotion. *Biol. Psychol. 67,* 1-2, 7–50.

COWIE, R. 2010. *Emotion-Oriented Systems: The Humaine Handbook*. Springer.

COWIE, R., DOUGLAS-COWIE, E., TSAPATSOULIS, N., VOTSIS, G., KOLLIAS, S., FELLENZ, W., AND TAYLOR, J. 2002. Emotion recognition in human-computer interaction. *IEEE Signal Process. Mag. 18,* 1, 32–80.

EKMAN, P., FRIESEN, W. V., O'SULLIVAN, M., CHAN, A., DIACOYANNI-TARLATZIS, I., HEIDER, K., KRAUSE, R., LECOMPTE, W. A., PITCAIRN, T., AND RICCI-BITTI, P. E. 1987. Universals and cultural differences in the judgments of facial expressions of emotion. *J. Person. Social Psychol. 53,* 4, 712–717.

EKMAN, P., LEVENSON, R. W., AND FRIESEN, W. V. 1983. Autonomic nervous system activity distinguishes among emotions. *Sci. 221*, 1208–1210.

FRAGOPANAGOS, N. AND TAYLOR, J. 2005. Emotion recognition in human-computer interaction. *Neural Netw. 18,* 4, 389–405.

GILLICK, L. AND COX, S. J. 1989. Some statistical issues in the comparison of speech recognition algorithms. In *Proceedings of the International Conference Acoustics, Speech and Signal Processing*. 532–535.

GUO, L., RIVERO, D., SEOANE, J., AND PAZOS, A. 2009. Classification of EEG signals using relative wavelet energy and artificial neural networks. In *Proceedings of the ACM/SIGEVO Summit on Genetic and Evolutionary Computation*. 177–184.

HANJALIC, A. AND XU, L. 2005. Affective video content representation and modeling. *IEEE Trans. Multimedia 7,* 1, 143–154.

HEALEY, J. A. 2000. Wearable and automotive systems for affect recognition from physiology. Ph.D. thesis, MIT.

ISHINO, K. AND HAGIWARA, M. 2003. A feeling estimation system using a simple electroencephalograph. In *Proceedings of the IEEE International Conference Systems, Man and Cybernetics*. Vol. 5. 4204–4209.

JAIMES, A. AND SEBE, N. 2007. Multimodal human-computer interaction: A survey. *Comput. Vis. Image Understand. 108,* 1-2, 116–134.

KIM, J. AND ANDRÉ, E. 2008. Emotion recognition based on physiological changes in music listening. *IEEE Trans. Pattern Anal. Mach. Intell. 30,* 12, 2067–2083.

KIM, K., BANG, S., AND KIM, S. 2004. Emotion recognition system using short-term monitoring of physiological signals. *Med. Biol. Engin. Comput. 42,* 3, 419–427.

KOELSTRA, S., YAZDANI, A., SOLEYMANI, M., MÜHL, C., LEE, J.-S., NIJHOLT, A., PUN, T., EBRAHIMI, T., AND PATRAS, I. 2010. Single trial classification of EEG and peripheral physiological signals for recognition of emotions induced by music videos. In *Proceedings of the International Conference Brain Informatics*. Springer, 89–100.

KOLEV, V., DEMIRALP, T., YORDANOVA, J., ADEMOGLU, A., AND ISOGLU-ALKAÇ, Ü. 1997. Time-frequency analysis reveals multiple functional components during oddball P300. *NeuroRep. 8,* 8, 2061–2065.

KOSTYUNINA, M. AND KULIKOV, M. 1996. Frequency characteristics of EEG spectra in the emotions. *Neurosci. Behav. Physiol. 26,* 4, 340–343.

KRAUSE, C., VIEMERÖ, V., ROSENQVIST, A., SILLANMÄKI, L., AND ÅSTRÖM, T. 2000. Relative electroencephalographic desynchronization and synchronization in humans to emotional film content: An analysis of the 4-6, 6-8, 8-10 and 10-12 Hz frequency bands. *Neurosci. Lett. 286,* 1, 9–12.

LANG, P. 1995. The emotion prob: Studies of motivation and attention. *Amer. Psychol. 50,* 5, 372–385.

LANG, P., BRADLEY, M., AND CUTHBERT, B. 2008. International affective picture system (IAPS): Affective ratings of pictures and instruction manual. Tech. rep. A-8, University of Florida, Gainesville, FL.

LANG, P., GREENWALD, M., BRADELEY, M., AND HAMM, A. 1993. Looking at pictures- affective, facial, visceral, and behavioral reactions. *Psychophysiol. 30,* 3, 261–273.

LIN, Y., WANG, C., JUNG, T., WU, T., JENG, S., DUANN, J., AND CHEN, J. 2010. Eeg-Based emotion recognition in music listening. *IEEE Trans. Biomed. Engin. 57,* 7, 1798–1806.

LISETTI, C. L. AND NASOZ, F. 2004. Using noninvasive wearable computers to recognize human emotions from physiological signals. *EURASIP J. Appl. Signal Process.*, 1, 1672–1687.

MCFARLAND, R. A. 1985. Relationship of skin temperature changes to the emotions accompanying music. *Biofeedback Self-Regul. 10,* 3, 255–267.

MCNEMAR, E. L. 1947. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika 12*, 153–157.

MORRIS, J. D. 1995. SAM: the self-assessment manikin. An efficient cross-cultural measurement of emotional response. *J. Advertis. Res. 35,* 8, 63–68.

MOTA, S. AND PICARD, R. 2003. Automated posture analysis for detecting learner's interest level. In *Proceedings of the Computer Vision and Pattern Recognition Workshop*. 49–49.

PETRANTONAKIS, P. AND HADJILEONTIADIS, L. 2010. Emotion recognition from eeg using higher order crossings. *IEEE Trans. Inf. Technol. Biomed. 14,* 2, 186–197.

PICARD, R. W., VYZAS, E., AND HEALEY, J. 2001. Toward machine emotional intelligence: Analysis of affective physiological state. *IEEE Trans. Pattern Anal. Mach. Intell. 23,* 10, 1175–1191.

PLUTCHIK, R. 2001. The nature of emotions. *Amer. Sci. 89*, 344.

ROSSO, O., BLANCO, S., YORDANOVA, J., KOLEV, V., FIGLIOLA, A., SCHÜRMANN, M., AND BAAR, E. 2001. Wavelet entropy: a new tool for analysis of short duration brain electrical signals. *J. Neurosci. Methods 105,* 1, 65–75.

RUSSELL, J. A. 1980. A circumplex model of affect. *J. Personal. Social Psychol. 39,* 6, 1161–1178.

SAVRAN, A., CIFTCI, K., CHANEL, G., MOTA, J. C., VIET, L. H., SANKUR, B., AKARUN, L., CAPLIER, A., AND ROMBAUT, M. 2006. Emotion detection in the loop from brain signals and facial images. In *Proceedings of the eNTERFACE 2006 Workshop*.

SCHAAFF, K. AND SCHULTZ, T. 2009. Towards emotion recognition from electroencephalographic signals. In *Proceedings of the International Conference Affective Computing and Intelligent Interaction and Workshops*. 1–6.

SEBE, N., COHEN, I., GEVERS, T., AND HUANG, T. 2006. Emotion recognition based on joint visual and audio cues. In *Proceedings of the 18th International Conference on Pattern Recognition*. Vol. 1. IEEE, 1136–1139.

SHANNON, C. 2001. A mathematical theory of communication. *ACM SIGMOBILE Mob. Comput. Comm. Rev. 5,* 1, 3–55.

SUTTON, S. AND DAVIDSON, R. 1997. Prefrontal brain asymmetry: A biological substrate of the behavioral approach and inhibition systems. *Psychol. Sci. 8,* 3, 204–210.

UNSER, M. 2002. Splines: A perfect fit for signal and image processing. *IEEE Signal Process. Mag. 16,* 6, 22–38.

VERVERIDIS, D. AND KOTROPOULOS, C. 2006. Emotional speech recognition: Resources, features, and methods. *Speech Comm. 48,* 9, 1162–1181.

WANG, J. AND GONG, Y. 2008. Recognition of multiple drivers' emotional state. In *Proceedings of the International Conference Pattern Recognition*. 1–4.

WILCOXON, F. 1945. Individual comparisons by ranking methods. *Biomet. Bull. 1,* 6, 80–83.

YAZDANI, A., LEE, J., AND EBRAHIMI, T. 2009. Implicit emotional tagging of multimedia using eeg signals and brain computer interface. In *Proceedings of the 1st SIGMM Workshop on Social Media*. ACM, 81–88.

YAZDANI, A., VESIN, J., IZZO, D., AMPATZIS, C., AND EBRAHIMI, T. 2010a. The impact of expertise on brain computer interface based salient image retrieval. In *Proceedings of the IEEE Annual International Conference on Engineering in Medicine and Biology Society (EMBC)*, IEEE, 1646–1649.

YAZDANI, A., VESIN, J., IZZO, D., AMPATZIS, C., AND EBRAHIMI, T. 2010b. Implicit retrieval of salient images using brain computer interface. In *Proceedings of the 17th IEEE International Conference on Image Processing (ICIP)*. IEEE, 3169–3172.

YEASIN, M., BULLOT, B., AND SHARMA, R. 2006. Recognition of facial expressions and measurement of levels of interest from video. *IEEE Trans. Multimedia 8,* 3, 500–508.

YORDANOVA, J., KOLEV, V., ROSSO, O., SCHÜRMANN, M., SAKOWITZ, O., ÖZGÖREN, M., AND BASAR, E. 2002. Wavelet entropy analysis of event-related potentials indicates modality-independent theta dominance. *J. Neurosci. Methods 117,* 1, 99–109.

ZENG, Z., PANTIC, M., ROISMAN, G. I., AND HUANG, T. S. 2009. A survey of affect recognition methods: Audio, visual, and spontaneous expressions. *IEEE Trans. Pattern Anal. Mach. Intell. 31,* 1, 39–58.

Received December 2010; revised July 2011; accepted October 2011