Multi-Modal Emotion Recognition in Response to Videos

Mohammad Soleymani, *Student Member, IEEE,* Maja Pantic, *Senior Member, IEEE,*
Thierry Pun, *Member, IEEE,*

Abstract—This paper presents a user-independent emotion recognition method with the goal of recovering affective tags for videos using electroencephalogram (EEG), pupillary response and gaze distance. We first selected 20 video clips with extrinsic emotional content from movies and online resources. Then EEG responses and eye gaze data were recorded from 24 participants while watching emotional video clips. Ground truth was defined based on the median arousal and valence scores given to clips in a preliminary study using an online questionnaire. Based on the participants' responses, three classes for each dimension were defined. The arousal classes were calm, medium aroused and activated and the valence classes were unpleasant, neutral and pleasant. One of the three affective labels of either valence or arousal was determined by classification of bodily responses. A one-participant-out cross validation was employed to investigate the classification performance in a user-independent approach. The best classification accuracy of 68.5% for three labels of valence and 76.4% for three labels of arousal were obtained using a modality fusion strategy and a support vector machine. The results over a population of 24 participants demonstrate that user-independent emotion recognition can outperform individual self-reports for arousal assessments and do not underperform for valence assessments.

Index Terms—Emotion recognition, EEG, Pupillary reflex, Pattern classification, Affective computing.

1 Introduction

Emotions play an important role in viewers' content selection and consumption. With the rapid expansion of digital multimedia content, alternative methods to the existing explicit tagging are needed to enrich the pool of tagged content. When a user watches video clips or listens to music, he/she may experience certain feelings and emotions [1], [2], [3] which manifest through bodily and physiological cues, e.g., pupil dilation and contraction, facial expressions, e.g., frowning, and changes in vocal features, e.g., laughter.

In order to translate user's bodily and behavioral reactions to emotions, reliable emotion assessment techniques are required. Emotion assessment is a challenging task; even users are not always able to express their emotion with words and the emotion self-reporting error is not negligible. This makes it difficult to define a ground truth. Affective self-reports might be held in doubt because users cannot always remember all the different emotions they had felt during watching a video, and/or might misrepresent their feelings due to self presentation, e.g., a user wants to show he is courageous whereas in reality he was scared [4]. The emotion recognition system provides us with an alternative that reduces the effort of deciding on the right label and on defining the right questions or methods to assess emotions explicitly.

One of the most accepted and well-known theories which explains the process of emotional experience is appraisal theory. According to this theory, cognitive judgment or appraisal of situation is a key factor in the emergence of emotions [5], [6], [7]. According to Orthoney, Clore and Collins (OCC) [6] emotions are experienced with the following scenario. First, there is a perception of an event, object or an action. Then, there will be an evaluation of events, objects or actions according to personal wishes or norms. Finally, the perception and evaluation result in a specific emotion. Considering this scenario for an emotional experience in response to multimedia content, emotions arise first through sympathy with the presented emotions in the content. During appraisal process for an emotional experience in response to multimedia content, a viewer examines events, situations and objects with respect to their novelty, pleasantness, goal, attainability, copability, and compatibility with his/her norms. Then, viewer's perception induces specific emotions which changes the viewer's physiological responses, motor actions, and feelings [8].

Scherer [9] categorized emotions into utilitarian and aesthetic emotions. Emotional responses to videos are a mixture of both utilitarian and aesthetic emotions with an emphasize on the later one. Existence of aesthetic emotional responses discourages simply using six well-known basic emotions in the context of emotion understanding of videos.

- *Mohammad Soleymani and Thierry Pun are with the Computer Vision and Multimedia Laboratory, University of Geneva, Battelle Campus, Building A, Rte. de Drize 7, Carouge(GE) CH - 1227, Switzerland. Email: mohammad.soleymani@unige.ch*
- *Maja Pantic is with the Department of computing, Imperial College London, SW7 2AZ, UK e-mail: (see http://www.doc.ic.ac.uk/~maja)*
- *Maja Pantic is also with the Faculty of Electrical Engineering, Mathematics and computer science, University of Twente, the Netherlands.*

Digital Object Indentifier 10.1109/TAC.2011.37 1949-3045/11/$26.00 © 2011 IEEE

1.1 Background

Emotional responses to multimedia content have been studied from three perspectives. There has been a research trend towards estimating emotions from multimedia content [10], [11], [12]. On the other hand, recognizing emotions induced by videos has been studied in the affective computing community [13], [14], [15], [3], [16]. The emotion recognition has been also used in applications such as detecting topical relevance, or summarizing videos [17], [13], [18]. This paper presents an emotion recognition method using EEG signals and eye gaze data in response to videos using users' bodily responses.

Hanjalic and Xu [10] 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 based features based on their relation to the valence-arousal space that was defined as an affect model. Irie et al. [19] proposed a latent topic model by defining affective audio-visual words in the content of movies to detect emotions in movie scenes. They extracted emotion-category-specific audio-visual features named affective audio-visual words. These higher level features were used to classify movie scenes using a latent topic driving model. This model takes into account temporal information which is the effect of the emotion from precedent scene to improve affect classification. The probability of emotional changes between consecutive scenes was also used in [20] to improve emotional classification of movie scenes using content features.

Emotional characteristics of videos have also improved music and image recommendation. Shan et al. [12] used affective characterization using content analysis to improve film music recommendation. Tkalčič et al. showed how affective information can improve image recommendation [11]. In their image recommendation scenario, affective scores of images from the international affective picture system (IAPS) [21] were used as features for an image recommender. They conducted an experiment with 52 participants to study the effect of using affective scores. The image recommender using affective scores showed a significant improvement in the performance of their image recommendation system.

Joho et al. [17], [13] developed a video summarization tool using facial expressions. A probabilistic emotion recognition based on facial expressions was employed to detect emotions of 10 participants watching eight video clips. The participants were asked to mark the highlights of the video with an annotation tool after the experiments. The expression change rate between different emotional expressions and the "pronounce level" or amount of expression were used as features to detect personal highlights in the videos. The pronounce levels they used was ranging from highly expressive emotions, surprise and happiness, to no expression or neutral. They have also extracted two affective content-based features which were audio energy and visual change rate from videos to create an affective curve in the same way as the affective highlighting method proposed by Hanjalic [10]

There has been long standing research on emotion assessment from physiological signals [1], [22], [23], [24], [14], [25]. Amongst these studies, few of them achieved notable results using video stimuli. Lisetti and Nasoz used physiological response to recognize emotion in response to movie scenes [14]. The movie scenes elicited six emotions, namely sadness, amusement, fear, anger, frustration and surprise. They achieved a high recognition rate of 84% for the recognition of these six emotions. However the classification was based on the analysis of the signals in response to pre-selected segments in the shown video known to be related to highly emotional events.

Takahashi [15] recorded EEG and peripheral physiological signals from 12 participants. He then classified the responses to emotional videos into five classes namely, joy, sadness, disgust, fear, and relax. He achieved the accuracy of 41.7% using EEG signals. However the feature level fusion of EEG signals and peripheral physiological signals failed to improve the classification accuracy.

An affective characterization for movie scenes using peripheral physiological signals as well as multimedia content features was proposed by Soleymani et al. [3]. Eight participants watched 64 movie scenes and self-reported their emotions. Affective correlates between different physiological and content features were studied. A linear regression trained by relevance vector machines (RVM) was utilized to estimate each clip's affect from physiological and content features.

Koelstra et al. [16] recorded EEG and peripheral physiological signals of six participants in response to music videos. Participants rated their felt emotions by means of arousal, valence and like/dislike rating rating. The emotional responses of each participant was classified into two classes of low/high arousal, low/high like/dislike, and low/high valence. The average classification rates varied between 55% and 58% which is slightly above random level.

In a more recent study, Kolodyazhniy et al. [26] used peripheral physiological signals to recognize neutral, fear and sadness responses to movie excerpts. During the presentation of videos to the participants, they introduced startle stimuli using randomly generated white noise sounds to boost physiological responses. Their system was able to recognize sadness, fear and neutral emotional states with the recognition rate of 77.5% in a participant-independent approach.

Eye gaze and pupillary responses has been used extensively to measure attention. However, we are not aware of research on how emotions affect eye gaze while watching videos; therefore the eye gaze itself has not been used for emotion recognition. The pupillary response is the measurement of pupil diameter over

time. Pupil can dilate or constrict in response to light, cognitive, attentional and emotional stimuli [27], [28]. Gao et al. [29] showed the significance of using pupillary reflex for emotion assessment after reducing the light effect using a real-time feedback.

1.2 Potential application

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 [12], [11], [30].

Currently, social media websites encourage users to tag their content. However, the users' intent when tagging multimedia content does not always match the information retrieval goals. A large portion of user defined tags are either motivated by the goal of increasing the popularity and reputation of a user in an online community or based on individual and egoistic judgments [31]. Implicit tagging does not interrupt users while listening or watching a video. Moreover, in presence of a reliable implicit tagging measurement method, determined tags carry less irrelevant and inaccurate information.

Users do not evaluate media content on the same criteria. Some might tag multimedia content with words to express their emotion while others might use tags to describe the content. For example, a picture receive different tags based on the objects in the image, the camera by which the picture was taken or the emotion a user felt looking at the picture. Scherer defines this by intrinsic and extrinsic appraisal [9]. Intrinsic appraisal is independent from the current goals and values of the viewer while extrinsic or transactional appraisal leads to feeling emotions in response to the stimuli. For example, the content's intrinsic emotion of a picture with a smiling face is happiness whereas this person might be a hatred figure to the viewer and the extrinsic appraisal leads to unpleasant emotions. What we want to detect is the later one that is the emotion felt by the viewer.

The goal of implicit affective tagging is thus to monitor the reactions of a person in response to a particular multimedia content and automatically recognize the corresponding tag. These responses can be used to reliably generate affective tags. A scheme of implicit tagging scenario versus explicit tagging is shown in Fig. 1.

In the proposed implicit tagging scenario, multimedia content will be tagged based on the bodily reactions of users recorded by a physiological acquisition device and an eye gaze tracker. The reactions can be used both to find tags common to a population and to develop a personal profile possibly in combination with user preferences and browsing history. With the recently

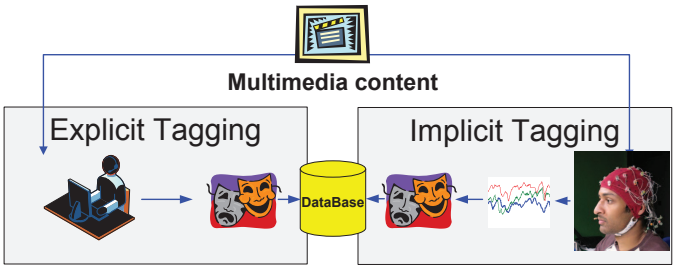


Fig. 1. Implicit affective tagging vs. explicit tagging scenarios. The analysis of the bodily responses replace the direct interaction between user and the computer. Therefore, user do not have to be distracted for tagging the content.

marketed physiological devices such as Neurosky[1], and Emotiv helmet[2], physiological interfaces are likely going to be the emerging human computer interfaces of the future.

1.3 Research questions

Although individual differences are always present in emotional reactions, there usually also exists one more common affective response in a population to a multimedia content. For example the scenes from a drama movie induce sadness in most of the people. For a dramatic scene, sadness can be considered the popular or frequently felt emotion. In this paper, we investigate and show the feasibility and comparable performance of a user-independent emotion recognition to detect the dominant or commonly selected affective tags. The implicit tagging application limited our choices for modalities to cues which are measurable while participants are sitting and are mostly in passive mode. The research questions that this paper investigates are:

1) Is it possible to design an accurate and user-independent classification protocol to recognize emotions from pupillary reflex, EEG signals and other bodily responses in response to video content?
2) Can non-verbal affective cues replace affective self-report with comparable emotion recognition performance and no requisite of direct user inputs?

The rest of the paper is organized as follows. The emotion model, video dataset, experimental protocol and physiological dataset classification are discussed in Section 2. Section 3 presents and discusses the experimental results and compares their performance to self-reports. The current methods' open issues and future work are discussed in Section 4. The paper is concluded in Section 5.

1. http://www.neurosky.com/
2. http://www.emotiv.com/

2 MATERIAL AND METHODS

2.1 Emotional model

Although the most straightforward way to represent an emotion is to use discrete labels such as fear, and joy, label-based representations have some disadvantages. The main one being that labels are not cross-lingual: emotions do not have exact translations in different languages, e.g., "disgust" does not have an exact translation in Polish [32]. Psychologists therefore often represent emotions or feelings in an n-dimensional space (generally 2- or 3-dimensional). The most famous such space, which is used in the present study and originates from cognitive theory, is the 2D valence-arousal or pleasure-arousal space [33]. The valence scale ranges from unpleasant to pleasant. The arousal scale ranges from passive to active or excited.

2.2 Preliminary study

In the preliminary study 21 commercially produced movies were first segmented into their scenes. Scenes longer than two minutes were divided into shorter two minutes long excerpts. From these excerpts, 155 emotional video clips containing full or part of movie scenes were manually selected. The 155 selected videos were shown to more than 50 participants; each video clip received 10 annotations in average [34]. The preliminary study was conducted utilizing an online affective annotation system in which the participants were able to use a web interface to report their emotions in response to the videos played by a web-based video player (see Fig. 2). In case of using videos from online repositories, the full length videos were used in the dataset.

In the preliminary study the participants were thus asked to self-assess their emotion by reporting the felt arousal (ranging from calm to excited/activated) and valence (ranging from unpleasant to pleasant) on nine points scale as well as emotional keywords. 14 video clips were chosen based on the preliminary study from the clips which received the highest number of emotional keyword tags in different emotion categories which are listed in the Table 1. Videos were selected to cover different emotional responses (see Fig. 3). Three other popular video clips from online resources were added to this set (two for joy/happiness and one for disgust). Three past weather forecast reports (retrieved from youtube.com) were also used as neutral emotion clips. The videos from online resources were added to the dataset to enable us to distribute some of the emotional video samples with the recorded multi-modal dataset described below. Table 1 gives the emotional labels, titles, and sources of the emotional video clips.

The median arousal and valence was used to determine ground truth labels with the following procedure. First, the values assessed by the online questionnaire were centered and then three equal length intervals were defined on the assessment range ($arousal, valence \in [1, 9]$). The labels assigned to all videos are given in Table 1. The distribution of online self emotions for the selected videos is shown in Fig. 3.

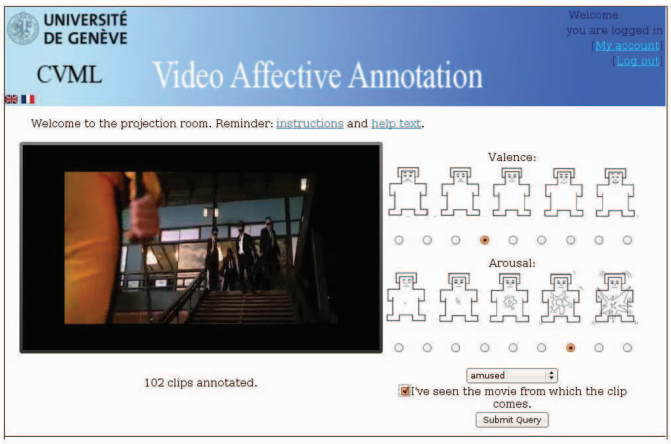


Fig. 2. A snapshot of the online affective annotation system. Arousal and valence were assessed using SAM manikins. Participants were able to give their emotional tag using a drop down menu.

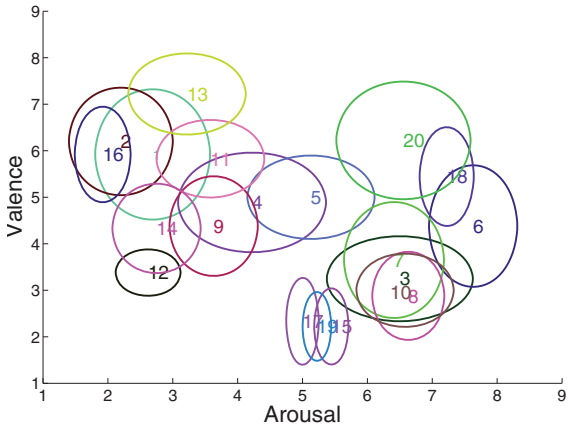


Fig. 3. Stimulus videos are shown in the valence-arousal plane. The center of the ellipses represents the mean arousal and valence and the horizontal and vertical radius represents the standard deviation of the online assessments. The clip codes are printed at the center of each ellipse.

Ultimately, 20 videos were selected to be shown which were between 34.9s to 117s long ($M = 81.4s$, $SD = 22.5s$). Psychologists recommended videos from one to ten minutes long for elicitation of a single emotion [35], [2]. Here, the video clips were kept as short as possible to avoid multiple emotions or habituation to the stimuli while keeping them long enough to observe the effect.

2.3 Experiment Protocol and Setup

A multi-modal recording setup was arranged to record facial videos, audio and vocal expressions, eye gaze,

TABLE 1
The video clips are listed the with their sources. The emotion labels are: calm (Cal.), medium aroused (Med.), activated (Act.), unpleasant (Unp.), neutral valence (Neu.), Pleasant (Pls.).

Code	Emotion Labels	Video clips sources
1	Act., Unp.	Hannibal
2	Act., Unp.	The Pianist
3	Med., Pls.	Mr. Bean's holiday
4	Act., Neu.	Ear worm (blip.tv)
5	Med., Neu.	Kill Bill VOL I
6	Med., Pls.	Love actually
7	Med., Pls.	Mr. Bean's holiday
8	Cal., Pls.	The thin red line
9	Med., Neu.	The shining
10	Med., Pls.	Love actually
11	Act., Unp.	The shining
12	Med., Unp.	Gangs of New York
13	Act., Unp.	Silent hill
14	Med., Unp.	The thin red line
15	Cal., Neu.	AccuWeather New York weather report (youtube.com)
16	Act., Unp.	American history X
17	Cal., Neu.	AccuWeather Detroit weather report (youtube.com)
18	Act., Pls.	Funny cats (youtube.com)
19	Cal., Neu.	AccuWeather Dallas weather report (youtube.com)
20	Act., Pls.	Funny (blip.tv)

and physiological signals simultaneously (see Fig. 4). The experiment was controlled by the Tobii studio software (http://www.tobii.com). In order to synchronize different modalities, device generated time stamps were recorded along with audio and physiological signals. These time stamps consisted of time series with square shaped periodic signal (60Hz) representing the moments when the cameras' shutters were open to capture each frame. The synchronization method and hardware setup details are given in Lichtenauer et al. [36].

The Biosemi active II system[3] with active electrodes was used for physiological signals acquisition. Physiological signals including ECG, EEG (32 channels), galvanic skin response (GSR), respiration amplitude, and skin temperature were recorded while the videos were shown to the participants. Peripheral physiological signals, facial videos and vocal expressions modalities were not employed in this paper; therefore all the results of this paper are only based on EEG signals, pupillary response and gaze distance recorded by eye gaze tracker. However, the classification protocol presented in this paper can be applied to a wide variety of modalities and is not limited to the utilized modalities.

Participants were asked to report their felt emotions by indicating their felt arousal and valence on a nine points scale. To simplify the interface a keyboard was provided with only nine numerical keys and the participant could answer each question by pressing one of the nine.

30 participants with different cultural and education backgrounds volunteered to participate in response to a campus wide call for volunteers at Imperial College, London. Out of the 30 young healthy adult participants, 17 were female and 13 were male; ages varied between 19 to 40 years old ($M = 26.06,\ SD = 4.39$). Participants had different educational background from undergraduate students to post-docs with different English proficiency from intermediate to native speakers. The data recorded from six participants were not analyzed due to technical problems, poor signal quality and unfinished data collection. Hence, the analysis results of this paper are only based on the responses recorded from 24 participants. The results in the Section 3 of this paper are only based on Eye gaze data and EEG signals. This database is freely available to the academic community, and is easily accessible through a web-interface[4].

Fig. 4. The experimental setup. 6 video cameras were recording facial expressions during the experiment. The modified keyboard is visible in front of the participant.

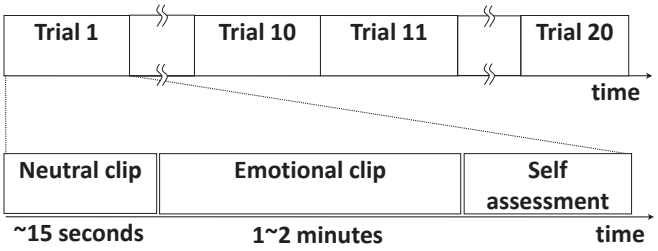


Fig. 5. Each trial started by a 15s neutral clip and continued by playing one emotional clip. The self-assessment was done at the end of each trial. There were 20 trials in each session of experiment.

3. http://www.biosemi.com

4. http://mahnob-db.eu

The participants were informed about the experiment and their rights with a verbal introduction, by email and through a consent form. Participants were trained about the interface before the experiment and during the setup time. The participants were also introduced to the meaning of arousal, valence in the self-assessment procedure, and to the nature of the video content.

In emotional-affective experiments the bias from the emotional state needs to be removed. For this purpose before each emotional video a short neutral clip randomly selected from the clips provided by the Stanford psychophysiology laboratory [2] was shown to the participants.

Each trial started with a short neutral clip. After watching the short neutral clip, one of the 20 video clips was played. Video clips were played from the dataset in random order. After watching the video clip, the participant filled in the self-assessment form which appeared automatically. In total, the time interval between the start of a trial and the end of the self-reporting phase was approximately two and half minutes. This interval included playing the neutral clip, playing the emotional clip, performing the self-assessment. Running of the whole protocol took in average 50 minutes in addition to 30 minutes setup time (see Fig. 5).

2.4 Preprocessing and Feature Extraction

2.4.1 EEG signals

Psychological studies regarding the relations between emotions and the brain are uncovering the strong implication of cognitive processes in emotions [37], [38]. As a result, the EEG signals carry valuable information about the participants' felt emotions.

Electroencephalogram signals were recorded with a 1024Hz sampling rate and later downsampled to 256Hz to reduce the memory and processing costs. EEG signals were recorded using active AgCl electrodes placed according to the international 10-20 system. The layout of EEG electrodes on the cap are shown in Fig. 6. The unwanted artifacts, trend and noise were reduced prior to extracting the features from EEG data by pre-processing the signals. Drift and noise reduction were done by applying a 4-45Hz band-pass filter. Other artifacts such as muscular activity was kept at minimum level by instructing the participants to minimize their movements while videos were playing. Biosemi active electrodes record EEG signals referenced to common mode sense electrode (CMS) as a part of its feedback loop. In order to gain the full common-mode rejection ratio (CMRR) at 50Hz, EEG signals should be re-referenced to another reference. EEG signals were thus re-referenced to the average reference to maximize signal to noise ratio.

The spectral power of EEG signals in different bands was found to be correlated with emotions [39], [40], [25]. Power spectral density (PSD) from different bands were computed using fast Fourier transform (FFT) and Welch algorithm [41]. In this method, the signal is split into overlapping segments and the PSD is estimated by averaging the periodograms. The averaging of periodograms results in smoother power spectrum. The PSD of each electrode's EEG signals was estimated using 15s long windows with 50% overlapping.

TABLE 2
This table list all the features extracted from eye gaze data and EEG signals.

Eye gaze data	Extracted features
Pupil diameter	standard deviation, spectral power in the following bands:]0, 0.2]Hz,]0.2, 0.4]Hz,]0.4, 0,6]Hz and]0.6, 1]Hz
Gaze distance	approach time ratio, avoidance time ratio, approach rate
Eye blinking	blink depth, blinking rate, length of the longest blink, time spent with eyes closed
EEG	theta, slow alpha, alpha, beta, and gamma PSD for each electrode. The spectral power asymmetry between 14 pairs of electrodes in the four bands of alpha, beta, theta and gamma.

The logarithms of the PSD from theta ($4Hz < f < 8Hz$), slow alpha ($8Hz < f < 10Hz$), alpha ($8Hz < f < 12Hz$), beta ($12Hz < f < 30Hz$) and gamma ($30Hz < f$) bands were extracted from all 32 electrodes as features. In addition to power spectral features, the difference between the spectral power of all the 14 symmetrical pairs on the right and left hemisphere was extracted to measure the possible asymmetry in the brain activities due to the valence of an emotional stimuli [42], [39]. The asymmetry features were extracted from all mentioned bands except slow alpha. The total number of EEG features of a trial for 32 electrodes is $14 \times 4 + 32 \times 5 = 216$ features. A list of extracted EEG features is given in Table 2.

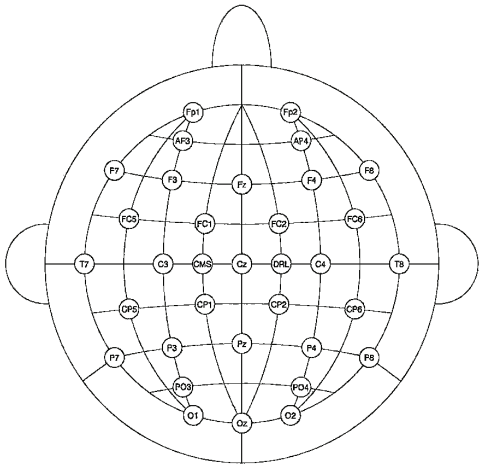


Fig. 6. The EEG cap layout for 32 EEG in addition to two reference electrodes. Retrieved from Biosemi website (http://www.biosemi.com).

2.4.2 Eye gaze data

The X120 Tobii[5] eye gaze tracker provides the position of the projected eye gaze on the screen, the pupil diameter, the moments when the eyes were closed and the instantaneous distance of the participant's eyes to the gaze tracker device positioned below the screen. The eye gaze data was sampled at 60Hz. The blinking moments are also extractable from eye gaze data. The eye gaze itself is highly dependent on the content and therefore it was not used directly for emotion recognition. However, pupil diameter has been shown to change in different emotional states [27], [28].

A linear interpolation was used to replace the missing pupil diameter samples due to eye blinking. Then the average diameter of right and left eye pupil was used as the pupil diameter time series. The major cause of pupil diameter variation comes from lighting; therefore the participants' responses to the same video (stimuli) in the controlled lighting environment follow similar patterns. There are different parametric models for pupillary light reflex [43], [44]. However, these parametric models are not error free and calculating their numerous parameters is rather difficult without specific light reflex experiment. It has been shown that the pupillary light reflex magnitude changes with age and between different people [43]. Most of the participants in our experiment were young, in their twenties; therefore the aging effect assumed to be negligible. The difference between the magnitudes can be reduced by normalizing the pupil diameter time series. Consequently we extracted the light reflex using a non-parametric estimation from the data. This common lighting reflex pattern was estimated for each video using principal component analysis (PCA).

If Y is the $M \times N_p$ matrix containing the centered and normalized pupillary responses to the same video from N_p participants and M samples, then Y consists of three components:

$$Y = X + Z + E \tag{1}$$

X is the lighting response which is the strongest effect in the signal. Z is the parasympathetic emotional and attentional response and E is the noise originated from measurement. These three components are originated from independent sources and the decorrelating characteristic of principal component analysis (PCA) is able to separate these three. First, Y was decomposed using principal component analysis (PCA) into N_p components. The first principal component is assumed to be a close estimation of the lighting reflex. The normalized principal component was then removed from normalized time series. Then the remaining residual part includes $Z + E$.

$$Y = UDV^T \tag{2}$$

$$A_p = UD \tag{3}$$

$$S_p = V^T \tag{4}$$

$$Y_1 = A_{p1} S_{p1} \tag{5}$$

$$Y_R = Y - Y_1 \tag{6}$$

If we decompose Y using singular value decomposition (SVD) U is a matrix with eigen vectors of YY^T as its column. D is a diagonal matrix whose diagonal values are the eigen values of YY^T. Finally the columns of V are the eigen vectors of Y^TY (see Equation 2). From the principal components of Y, A_p we can reconstruct the first principal component or the light reflex pattern Y_1 (see Equation 3). To remove the light reflex component, Y_1, from all the time series, it is enough to subtract it from the original data (see Equation 5 and 6). Y_R is the residual part which contains the emotional and attentional pattern in addition to the noise.

After removing the linear trend, the power spectrum of the pupil diameter variation was computed. Standard deviation and spectral features were extracted from the pupil diameter. The Hippus effect is the small oscillations of eye pupil diameter between 0.05 to 0.3Hz and with the amplitude of 1 mm [43], [45]. Hippus effect has been shown to be present when one is relaxed or passive. In the presence of mental activity the effect will disappear. The Hippus effect is extracted by the first two power spectral features which are covering up to 0.4 Hz. The rate of eye blinking is shown to be correlated with anxiety [46]. From the eye blinks, the eye blinking rate, the average and maximum blink duration were extracted as features. In addition to the eye blinking features the amount of time the participants spent with his/her eyes closed was also used as a feature to detect possible eye closing due to unpleasant emotions.

Although the participants were asked not to move during the experiment, there were small head movements which manifested itself in the distance between participants' eyes and the eye gaze tracker. The distance of the participant to the screen and its changes provide valuable information about the participants' posture. The total change in the distance of the user to the gaze tracker, gaze distance, was calculated to measure the possible approach and avoid phenomena. The amount of time the participant spent per trial getting close or far from the screen was computed as well. These features were named approach and avoidance ratio to represent the amount of time participant spent getting close or going far from the screen. The frequency of the participants' movement towards the screen during each trial, approach rate, was also extracted. Ultimately 12 features were extracted from the eye gaze data. A summary of all extracted features is given in Table 2.

5. http://www.tobii.com

2.4.3 Feature normalization

All the extracted features were numerical. To reduce the between participant differences, it is necessary to normalize the features. Maximum-Minimum normalization was applied on each feature of the features set separately on each participant's signals. In this normalization the minimum value for any given feature is subtracted from the same feature of a participant and the results were divided by the difference between the maximum and minimum values.

2.5 Emotion Classification

With the proposed inter-participant emotion recognition, the goal is to find the emotional class with the highest agreement within a population. The most popular emotional class or tag can satisfy a larger population of viewers in a video retrieval scenario. For each video from the dataset, the ground truth was thus defined by computing the median of arousal and valence scores given on a nine point scale. The median values were then categorized into three classes with equal intervals. According to this definition, we can name these classes calm, medium aroused, and activated for arousal and unpleasant, neutral, and pleasant for valence.

A SVM classifier with RBF kernel was employed to classify the samples using features from each of the two modalities. Prior to classification, a feature selection was used to select discriminative features as follows. First, a one way ANOVA test was done on only the training set for each feature with the class as the independent variable. Then any feature for which the ANOVA test was not significant ($p > 0.05$) was rejected. This feature selection criterion was hence re-calculated for each cross validation's iteration. A leave-one-participant-out cross validation technique was used to validate the user-independent classification performance. At each step of cross validation, the samples of one participant were taken out as test set and the classifier was trained on the samples from the rest of the participants. This cross validation was employed to imitate the effect of introducing a new user to our emotion recognition system. This process was repeated for all participants' data.

2.5.1 Modality fusion strategy

Classification in different modalities can be fused at both feature level and decision level. We applied these two fusion strategies and reported their results. With the feature level fusion, the feature vectors from different modalities were concatenated to form a larger feature vector. The feature selection and classification methods were then applied to the new feature set. However with the decision level fusion, classification was performed on each modality separately and the classification outcomes were fused to generate the fusion results. In [15] feature level fusion of EEG and peripheral physiological signals did not improve the single modality results. On the other hand, Chanel et al. [25] showed how a fusion strategy improved the emotion recognition accuracy by fusing the results from EEG and peripheral features at decision level. Our results (see Section 3) shows how in contrary to feature level fusion (FLF), decision level fusion (DLF) significantly outperforms the best single modality for arousal classification and do not underperform for valence classification.

In addition to the superior classification performance obtained by multi-modal strategy, in the absence of one of the modalities due to temporary problems or artifacts, the system can still continue working as single modality emotion detection. The adaptability of the system to remove and add new modalities can be achieved without re-training the classifiers using the DLF. The adaptability and scalability of the DLF strategy gives it another advantage over FLF.

Here we used two modalities which are EEG and eye gaze data. The results of the classification over two modalities were fused to obtain the multi-modal fusion results. If the classifiers provide confidence measures on their decisions, combining decisions of classifiers can be done using a summation rule. The confidence measure summation fusion was used due to its simplicity and its proved performance for emotion recognition according to [25]. Other decision combination methods including product of confidence measures, decision template fusion, Dempster-Shafer, Bayesian belief integration [47], weighted sum and weighted product [48] did not give superior results.

In this paper, the probabilistic outputs of classifiers are used as a measure of confidence. The sum rule is thus defined as follows for a given trial:

$$g_a = \frac{\sum\limits_{q \in Q} P_q(\omega_a|x_i)}{\sum\limits_{a=1}^{K} \sum\limits_{q \in Q} P_q(\omega_a|x_i)} = \sum_{q \in Q} \frac{1}{|Q|} P_q(\omega_a|x_i) \qquad (7)$$

In Equation 7, g_a is the summed confidence interval for affect class ω_a . Q is the ensemble of the classifiers chosen for fusion, $|Q|$ the number of such classifiers and $P_q(\omega_a|x_i)$ is the posterior probability of having class ω_a the sample is x_i according to classifier q. The final choice is done by selecting the class ω_a with the highest g_a. It can be observed that g_a can also be viewed as a confidence measure on the class, ω_a, given by the fusion of classifiers.

There are two problems employing SVM classifiers in this fusion scheme. First, they are intrinsically only two-class classifiers and secondly, their output is uncalibrated so that it is not directly usable as a confidence value in the case one wants to combine outputs of different classifiers or modalities. To tackle the first problem, the one versus all approach is used where one classifier is trained for each class (N classifier to train) and the final choice is done by majority voting. For the second problem, Platt [49] proposes to model the probability of being in one of the two classes knowing the output value of the SVM by using a sigmoid fit, while Wu et al.

[50] proposes a solution to extend this idea to multiple classes. In this study we used the MATLAB libSVM implementation [51] of the Platt and Wu algorithms to obtain the posterior probabilities, $P_q(\omega_a|x_i)$.

3 EXPERIMENTAL RESULTS

The experiments were performed in a laboratory environment with controlled temperature and illumination; 24 participants viewed 20 video clips each. 467 samples were gathered over a potential dataset of $24 \times 20 = 480$ samples; the 13 missing ones were unavailable due to technical difficulties.

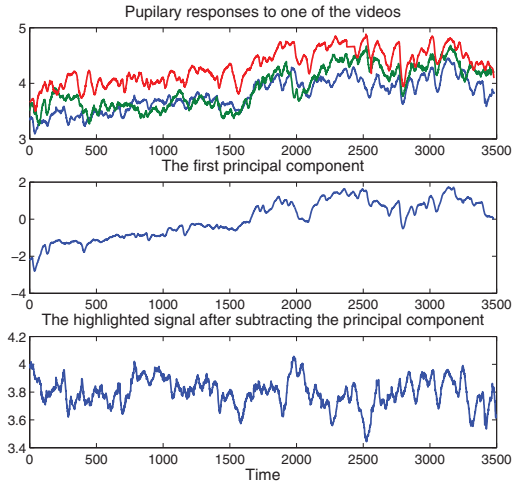
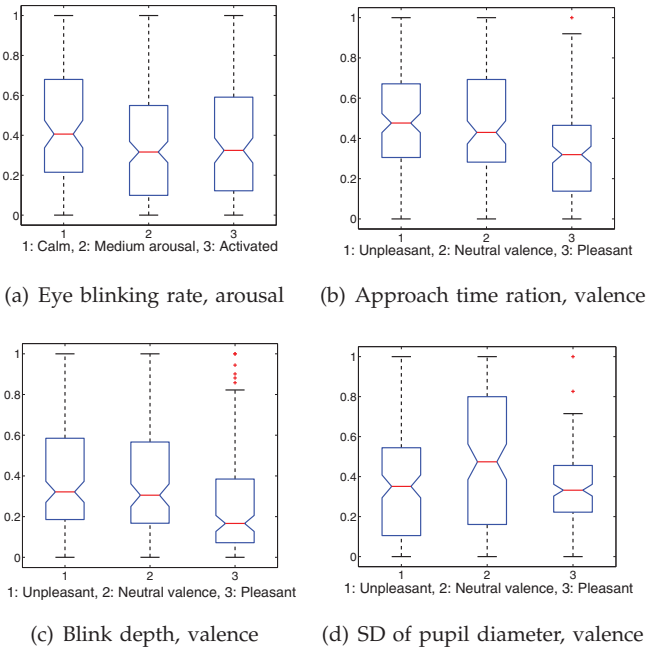


Fig. 7. From top to bottom: on the first plot there is an example of pupil diameter measures from three different participants in response to one video. The second plot shows the first principal component extracted by PCA from the time series shown in the first plot (the lighting effect). The bottom plot shows the pupil diameter of the blue signal in the first plot after reducing the lighting effect.

For the pupillary responses to each video the most significant component was extracted using PCA and then removed from the pupillary time series. In the example given in Fig. 7, examples of the pupillary responses, extracted pupillary lighting reflex and the residual component after removing the light reflex are given. The normalized variance or eigen-values of the first component were found to be significantly larger than the rest of the components. The first principal component carried in average more than 50% of the variance in the data.

In order to study the discrimination abilities of the eye gaze data features, a one way analysis of variance test was performed on the features. The difference between the mean of features in different arousal or valence categories was found significant ($p < 0.05$). The significance of one way ANOVA shows that there is at least a significant difference between the means of the samples from two classes out of three. The box plots of four features



(a) Eye blinking rate, arousal (b) Approach time ration, valence

(c) Blink depth, valence (d) SD of pupil diameter, valence

Fig. 8. Box plots of four different gaze data features in three emotional conditions. (a) Eye blinking rate for arousal classification (b) Approach time ratio for valence classification (c) Blink depth, average blink time, for valence classification (d) STD of pupil diameter for valence classification. One way ANOVA results showed a significant difference beween features mean of different classes ($p < 0.05$)

namely, eye blinking rate, approach rate, maximum blink length, and standard deviation of pupillary responses are shown in Fig. 8. In average eye blinking rate was higher in calmer videos (see Fig. 8(a)). The amount of time participants spent getting closer to the screen is lower for the pleasant category. This shows that they had a tendency to seat more upright while watching more pleasant videos (see Fig. 8(b)). On the other hand, the maximum blink length or depth is higher for unpleasant videos. This is due to the fact that participants kept their eyes closed for some moments while watching unpleasant videos (see Fig. 8(c)). Pupillary response's standard deviation is also shown to be higher during neutral scenes (see Fig. 8(d)).

To find the best discriminative EEG features, the linear discrimination criterion was calculated. This parameter is the the between class variance divided by within class variance for any given feature (see Table 3). For arousal classification, PSD in alpha bands of occipital electrodes was found to be the most discriminant features. In contrary for valence beta and gamma bands of temporal electrodes are more informative. The between class to within class variance ratios are higher for the best arousal EEG features. The higher linear discrimination criterion for best arousal features explains the superior classification rate for arousal dimension (see Table 4).

Regarding the self-reports, we computed the aver-

TABLE 3
10 best EEG features for arousal and valence classification based on linear discrimination criterion. The between class variance to within class variance ratios, $\sigma_{bw}^2/\sigma_{wn}^2$ are also given.

Arousal classification			Valence classification		
Band	Electrode/s	$\sigma_{bw}^2/\sigma_{wn}^2$	Band	Electrode/s	$\sigma_{bw}^2/\sigma_{wn}^2$
Slow α	PO4	0.18	β	T8	0.08
α	PO4	0.17	γ	T8	0.08
θ	PO4	0.16	β	T7	0.07
Slow α	PO3	0.15	γ	T7	0.06
θ	Oz	0.14	γ	P8	0.05
Slow α	O2	0.14	γ	P7	0.05
Slow α	Oz	0.14	θ	Fp1	0.04
θ	O2	0.13	β	CP6	0.04
θ	FC6	0.13	β	P8	0.04
α	PO3	0.13	β	P7	0.04

TABLE 4
The classification rate and F1 scores of emotion recognition for different modalities.

Modality	Classification rate		Average F1	
dimension	arousal	valence	arousal	valence
EEG	62.1%	50.5%	0.60	0.50
Eye gaze	71.1%	66.6%	0.71	0.66
Feature level fusion (FLF)	66.4%	58.4%	0.65	0.55
Decision level fusion (DLF)	76.4%	68.5%	0.76	0.68
Self-reports with SAM manikins	55.7%	69.4%	0.57	0.70
Random level	33.3%	33.3%	0.36	0.40

age pair-wise Cohen's kappa for keyword based annotations. A fair multi-rater agreement was found on emotional keywords (9 keywords) with $\kappa = 0.32$. The correlation between arousal and valence ratings between participants was also computed. The correlation between arousal and valence ratings given by different participants on nine points scales were $mean(\rho) = 0.45,\ SD(\rho) = 0.25$ and $mean(\rho) = 0.73,\ SD(\rho) = 0.12$ respectively. Therefore, there was a higher inter-rater agreement on valence comparing to arousal.

3.1 Emotion Recognition Feasibility

Referring to the first research question, the results have shown that it is possible to accurately recognize emotions with a user-independent approach. The classification accuracy measures are summarized in Table 4. The traditional F-score which combines precision and recall by their harmonic mean was also computed for each emotion category to give an overall evaluation of classification performance (Equation 8). The F1 score varies between zero and one. The random level is 0.5 for binary classification and balanced classes; values closest to 1 indicate a better performance.

$$F1 = \frac{2 \times precision \times recall}{precision + recall} \tag{8}$$

Precision and recall can be only defined for one class; hence, the F1 scores were calculated from the results of one versus all classification schemes for each class separately. As a result, the expectation of F1 scores of a uniform random classifier are calculated and given in Table 4. The classification rate of both three class classifications are defined as the percentage of correctly classified samples.

For the SVM classifier, the size of the kernel, γ, was selected between $[0.01, 10]$, based on the average F1 score using a 20-fold cross validation on the training set. The C parameter that regulates the tradeoff between error minimization and margin maximization is empirically set to 1. Classifications were first performed with the goal of recovering the three classes with a leave-one-participant-out cross validation scheme. Regarding the single modality classification of arousal and valence in three classes, we obtained 62.1% and 50.5% accuracy from EEG signals and 71.1% and 66.6% accuracy from eye gaze data (see Table 4). Although EEG classification results are inferior to the eye gaze data, they are comparable to the state of the art classification rates considering the inter-participant classification scheme [52], [25].

The FLF did not improve the best single modality, gaze data, results. However, the modality fusion strategy using the DLF improved the best SVM classification rate for arousal up to 76.4%. The DLF did not underperform for valence classification. To test the statistical significance of the classification performance, a paired t-test was used to compare F1 scores of the DLF on one side and the self reports and the best single modality, eye gaze data, on the other side. The F1 scores from each participant's samples were compared and the improvement over arousal classification comparing to eye gaze data and self reports were found significant ($p < 0.01$). However, the difference between the eye gaze, DLF, and self reports F1 scores on valence classification was not found statistically significant. The confidence levels of the classification results from the two modalities were added to find the class with the highest summed confidence.

The confusion matrices for each modality show how they performed on each emotion category (Tables 5.a - 5.j). In these confusion matrices the row represents the classified label and each column represents the ground truth. Only for activated category EEG classification performed as well as gaze data modality. However, the fusion of both with the exceptions of neutral valence class outperformed gaze data results (see tables 5.a - 5.d and 5.f - 5.i). The DLF outperformed the feature level fusion for all categories except unpleasant (see tables 5.c, 5.d, 5.h, and 5.i).

The use of different stimuli and emotion classes make it difficult to directly compare the results to similar work. Here, we compare the obtained accuracy with the most

TABLE 5
Confusion matrices of different classification schemes (row: classified label; column: ground truth). The numbers on the first row and the first column of tables a, b, c, d and e represents: 1. calm, 2. medium aroused, 3. activated and for tables f, g, h, i, and j represents: 1. unpleasant 2. neutral valence 3. pleasant. The confusion matrices relate to classification using (a, f) EEG signals (b, g) Eye gaze data (c, h) Feature level fusion (FLF) (d, i) Decision level fusion (DLF) (e, j) Self reports

Arousal

	1	2	3
1	44	15	11
2	22	111	39
3	30	60	135

(a) EEG

	1	2	3
1	60	7	9
2	10	136	40
3	26	43	136

(b) Eye gaze data

	1	2	3
1	49	14	10
2	15	117	31
3	32	55	144

(c) FLF

	1	2	3
1	62	6	8
2	8	146	28
3	26	34	149

(d) DLF

	1	2	3
1	63	35	11
2	24	88	65
3	9	63	109

(e) Self reports

Valence

	1	2	3
1	87	56	52
2	23	52	15
3	54	31	97

(f) EEG

	1	2	3
1	108	46	26
2	36	77	12
3	20	16	126

(g) Eye gaze data

	1	2	3
1	139	83	56
2	7	36	10
3	18	20	98

(h) FLF

	1	2	3
1	115	48	22
2	27	75	12
3	22	16	130

(i) DLF

	1	2	3
1	126	40	4
2	38	91	53
3	0	8	107

(j) Self reports

similar existing studies. Kolodyazhniy et al. [26] used videos extracted from movies in a user independent strategy to classify three classes; namely, neutral, sadness and fear. They obtained 77.5% recognition rate from peripheral physiological responses while introducing random startles. Their results on three classes is almost at the same level of our arousal classification performance. In a gaming protocol Chanel et al. [52] achieved the accuracy of 63% in a user-independent approach on the recognition of three classes; namely, boredom, engagement, anxiety. These three classes can be translated to our three arousal levels. Our results are inferior to the ones by Lisetti and Nasoz [14] on six classes. However, we used different modalities and videos and therefore these results are not directly comparable.

3.2 Emotion recognition and self-reports

Referring to the second research question, the agreement between participants' self-reports and ground truth is shown in the confusion matrix given in Table 5.e and Table 5.j. The columns of this table represent how the videos of each class defined by ground truth were individually self-reported. For example, the first column of this table represent how many of the samples which were actually in class one were classified into different classes.

In order to measure the agreement between individual self-reports and the ground truth, the self-reported arousal and valence scores on nine point scale were translated into three levels. These levels were then treated like classified labels and the classification rate was computed. This was done by considering that the goal of each participant is to label a video clip by the correct label, the most common tag. The classification rate for individually self-reported labels was 55.7% for arousal which is inferior to the worst classifier's result. Although, looking at the inter-annotation agreement, participants found that it easier to self-report pleasantness, the classification rate for valence is not significantly lower than the self-report rate. Therefore, the accuracy of obtained tags via classification is comparable to the individually reported labels.

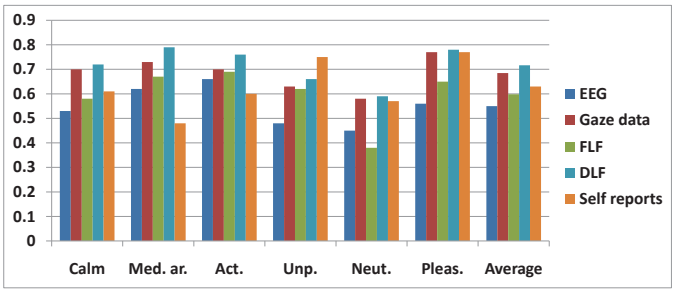


Fig. 9. This bar chart shows the F1 score for classification results of each class from different modalities.

Fig. 9 summarizes the comparison of different classification strategies showing the F1 scores for each category and on average. Looking at the bars on the most right side of the chart, only EEG results are inferior to the explicit self-report agreements using self assessment manikins.

4 Discussion

In this section, we discuss limitations of the current study and present the open issues. Physiological responses can vary due to non-emotional changes, such as circadian rhythms, ambient temperature, body posture and other psychophysiological factors such as attention , anticipation and mental effort [53]. Emotion recognition from bodily responses is therefore confounded by contextual factors. Moreover, like other similar studies [26], the generalization of the results are limited by the videos shown to the participants.

The inter-annotation agreement for arousal self-reports is lower comparing to keyword based self-assessments. In a real-case scenario for an explicit tagging system, using words will be easier for an ordinary user and leads to higher between participant agreement in comparison to arousal and valence reported by self assessment manikins (SAM). However, emotional keywords

are difficult to translate and might not exist with the exact same meaning in different languages [32]. Emotion detection can overcome those difficulties with keeping the accuracy at the same level. In this paper, participants were asked to explicitly choose an emotional indicator to form the ground truth. In a real application, with the existence of a reliable user-independent emotion recognition method, the self-reporting phase can be eliminated.

Arousal classes were in average detected with higher accuracy using EEG signals comparing to valence labels. This might be due to higher visual and auditory variance of the arousal variant videos comparing to valence variant ones. Exciting scenes usually contain fast movements and loud noises which manifest themselves both in EEG signals and pupillary responses, whereas the difference between pleasant and unpleasant responses can be hidden in the semantics. The direct bodily responses to different stimuli can increase the variance in responses and improve the emotion recognition results. For example, faster changing video induces a different response in occipital cortex activities comparing to a more static video.

The DLF superior classification rate for arousal and its similar performance for valence classification shows that the proposed emotion classification can replace the self-reporting of single participants for detecting popular emotional tags for this dataset. These popular emotional tags are defined by the emotions felt by the majority of users watching the same video. After detecting emotional classes, they can be stored with other metadata attached to each video. Emotional labels can be converted to scores for arousal and valence for each video. The emotional scores can be then used, as in the image recommendation applications [11], to improve a video recommender's performance. In future, the recognized emotional labels should be added as features to a video recommendation system to study the effect of introducing emotional labels on those systems. This effect can be determined by assessing users' satisfaction from a recommendation or retrieval system with and without emotional information. The emotion detection can be also used indirectly as a tool to detect topical relevance in information retrieval systems [18].

The determination of ground truth is based on the participants' feedback in the online preliminary experiment. In order to measure the agreement between the popular responses of the two separate populations of the preliminary assessments and the experiments, we computed the median valence and arousal reported during the experiment and compared the labels based on the recorded participants' popular response. Only three arousal labels out of 40 labels were changed from activated to medium arousal or vice-versa. No valence label has changed when comparing two populations. This is due to valence dimension's higher inter-annotator agreement.

This study still has open issues that need to be considered in the future. In a real case scenario, any new user will need to have a few minutes of signals recorded to provide reliable values for feature normalization. The estimation of pupillary response to lighting is an open issue which needs more investigation. Although we assumed that the lighting pupillary responses are similar between participants, there will be a large difference in case of introducing users from a different age group. Therefore, the parameters of a reliable model for pupillary reflex to lighting, such as [43], should be determined before introducing a new user to the system. Alternatively, the real time lighting effect similar to [29] can be employed to remove the lighting effect. A larger video set and larger number of participant can be considered to increase the generalization of the developed emotion classifier in response to videos.

In this paper, the length of the experimental session limited the number of videos we could show to each participant. The number of participants in this paper is large and diverse enough comparing to similar studies [1], [25]. However this population only consisted of young students which limits the trained algorithm to this particular group. These limitations might worsen the results in case of introducing a new genre of video which was not present in the current video set. To generalize and train such a system, the recruited participants should be as close as possible to the target audience of the video retrieval or recommendation systems.

Emotions can co-occur and/or last for very short moments. This puts using a single emotional label for a video clip under question. In order to address this issue, self-reporting should include the possibility to indicate different emotions and their degree of strength. This is possible by using questionnaires such as positive and negative affect schedule (PANAS) [54] or Geneva emotion wheel [9]. However, these emotional self-reporting methods are more complex and make the experiment longer. In future, multiple or co-occurring emotions should be assessed using a more sophisticated self reporting tool.

5 Conclusions

This paper showed the performance of an inter-participant emotion recognition tagging approach using participants' EEG signals, gaze distance and pupillary response as affective feedbacks. The feasibility of an approach to recognize emotion in response to videos is shown. Although the results were based on a fairly small video dataset due to experimental limitations the promising accuracy can be scalable to more samples from a larger population. The improved performance using multi-modal fusion techniques leads to the conclusion that by adding other modalities, such as facial expressions, accuracy as well as robustness should further improve. Results from our previous studies [3] showed that there is a significant difference between peoples' emotional self assessments in response to videos. However, there usually exists one most popular emotional

tag for which there is significant agreement in a population. This "most popular emotion" has been shown to be detectable with monitoring users' bodily responses. Moreover, the population tags give the retrieval system higher chance of success in a given population. We have shown that it is possible to design an accurate and user-independent classification protocol to recognize emotions from pupillary reflex, EEG signals in response to video content. Moreover, we have shown that for the utilized video dataset, the non-verbal affective cues can replace affective self-report with comparable emotion recognition performance and no requisite of direct user inputs. We can thus answer positively to our two research questions.

ACKNOWLEDGMENTS

The work of Soleymani and Pun was supported in part by the Swiss National Science Foundation and in part by the European Community's Seventh Framework Programme (FP7/2007-2011) under grant agreement Petamedia no. 216444. The data acquisition part of this work was supported by the European Research Council under the ERC Starting Grant agreement no. ERC-2007-StG-203143 (MAHNOB). The work of Pantic is supported in part by the European Community's 7th Framework Programme (FP7/2007-2013) under the grant agreement no 231287 (SSPNet). The authors would like to thank Dr. Jeroen Lichtenauer and Jozef Doboš (Imperial College, London), Prof. Didier Grandjean and Dr. Guillaume Chanel (University of Geneva) for their valuable scientific comments and technical support during the experiments.

REFERENCES

[1] J. Kim and E. André, "Emotion Recognition Based on Physiological Changes in Music Listening," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 30, no. 12, pp. 2067–2083, 2008.

[2] J. Rottenberg, R. D. Ray, and J. J. Gross, *Emotion elicitation using films*, ser. Series in affective science. Oxford University Press, 2007, pp. 9–28.

[3] M. Soleymani, G. Chanel, J. J. M. Kierkels, and T. Pun, "Affective Characterization of Movie Scenes Based on Content Analysis and Physiological Changes," *International Journal of Semantic Computing*, vol. 3, no. 2, pp. 235–254, June 2009.

[4] R. W. Picard and S. B. Daily, "Evaluating Affective Interactions: Alternatives to Asking What Users Feel," in *CHI Workshop on Evaluating Affective Interfaces: Innovative Approaches*, 2005.

[5] K. R. Scherer, "Studying the emotion-antecedent appraisal process: An expert system approach," *Cognition & Emotion*, vol. 7, no. 3, pp. 325–355, 1993.

[6] A. Ortony, G. L. Clore, and A. Collins, *The Cognitive Structure of Emotions*. Cambridge University Press, July 1988.

[7] D. Sander, D. Grandjean, and K. R. Scherer, "A systems approach to appraisal mechanisms in emotion," *Neural Netw.*, vol. 18, no. 4, pp. 317–352, May 2005.

[8] W. Wirth and H. Schramm, "Media and Emotions," *Communication research trends*, vol. 24, no. 3, pp. 3–39, 2005.

[9] K. R. Scherer, "What are emotions? And how can they be measured?" *Social Science Information*, vol. 44, no. 4, pp. 695–729, December 2005.

[10] A. Hanjalic and L.-Q. Xu, "Affective video content representation and modeling," *Multimedia, IEEE Transactions on*, vol. 7, no. 1, pp. 143–154, 2005. [Online]. Available: http://dx.doi.org/10.1109/TMM.2004.840618

[11] M. Tkalčič, U. Burnik, and A. Košir, "Using affective parameters in a content-based recommender system for images," *User Modeling and User-Adapted Interaction*, pp. 1–33–33, September 2010.

[12] M. K. Shan, F. F. Kuo, M. F. Chiang, and S. Y. Lee, "Emotion-based music recommendation by affinity discovery from film music," *Expert Syst. Appl.*, vol. 36, no. 4, pp. 7666–7674, September 2009.

[13] H. Joho, J. M. Jose, R. Valenti, and N. Sebe, "Exploiting facial expressions for affective video summarisation," in *Proceeding of the ACM International Conference on Image and Video Retrieval*, ser. CIVR '09. New York, NY, USA: ACM, 2009.

[14] C. L. Lisetti and F. Nasoz, "Using noninvasive wearable computers to recognize human emotions from physiological signals," *EURASIP J. Appl. Signal Process.*, vol. 2004, no. 1, pp. 1672–1687, January 2004.

[15] K. Takahashi, "Remarks on Emotion Recognition from BioPotential Signals," in *in 2nd Int. Conf. on Autonomous Robots and Agents, 2004*, 2005.

[16] 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," in *Brain Informatics*, ser. Lecture Notes in Computer Science, Y. Yao, R. Sun, T. Poggio, J. Liu, N. Zhong, and J. Huang, Eds. Berlin, Heidelberg: Springer, 2010, vol. 6334, ch. 9, pp. 89–100.

[17] H. Joho, J. Staiano, N. Sebe, and J. Jose, "Looking at the viewer: analysing facial activity to detect personal highlights of multimedia contents," *Multimedia Tools and Applications*, pp. 1–19, October 2010.

[18] I. Arapakis, I. Konstas, and J. M. Jose, "Using facial expressions and peripheral physiological signals as implicit indicators of topical relevance," in *Proceedings of the seventeen ACM international conference on Multimedia*, ser. MM '09. New York, NY, USA: ACM, 2009, pp. 461–470.

[19] G. Irie, T. Satou, A. Kojima, T. Yamasaki, and K. Aizawa, "Affective Audio-Visual Words and Latent Topic Driving Model for Realizing Movie Affective Scene Classification," *Multimedia, IEEE Transactions on*, vol. 12, no. 6, pp. 523 –535, October 2010.

[20] M. Soleymani, J. J. M. Kierkels, G. Chanel, and T. Pun, "A Bayesian framework for video affective representation," in *Proceedings of the International Conference on Affective Computing and Intelligent interaction (ACII 2009)*, September 2009, pp. 1–7.

[21] P. Lang, M. Bradley, and B. Cuthbert, "International affective picture system (IAPS): Affective ratings of pictures and instruction manual," University of Florida, Gainesville, Florida, US, Tech. Rep. A-8, 2008.

[22] P. J. Lang, M. K. Greenwald, M. M. Bradley, and A. O. Hamm, "Looking at pictures: Affective, facial, visceral, and behavioral reactions," *Psychophysiology*, vol. 30, no. 3, pp. 261–273, 1993.

[23] J. Wang and Y. Gong, "Recognition of multiple drivers' emotional state," in *In ICPR'08: The 19th International Conference on Pattern Recognition*, 2008.

[24] J. A. Healey, "Wearable and automotive systems for affect recognition from physiology," Ph.D. dissertation, MIT, 2000.

[25] G. Chanel, J. J. M. Kierkels, M. Soleymani, and T. Pun, "Short-term emotion assessment in a recall paradigm," *International Journal of Human-Computer Studies*, vol. 67, no. 8, pp. 607–627, August 2009.

[26] V. Kolodyazhniy, S. D. Kreibig, J. J. Gross, W. T. Roth, and F. H. Wilhelm, "An affective computing approach to physiological emotion specificity: Toward subject-independent and stimulus-independent classification of film-induced emotions," *Psychophysiology*, vol. 7, no. 48, pp. 908–922, 2011.

[27] Bradley, M. Margaret, Miccoli, Laura, Escrig, A. Miguel, Lang, and J. Peter, "The pupil as a measure of emotional arousal and autonomic activation," *Psychophysiology*, vol. 45, no. 4, pp. 602–607, July 2008.

[28] T. Partala and V. Surakka, "Pupil size variation as an indication of affective processing," *International Journal of Human-Computer Studies*, vol. 59, no. 1-2, 2003.

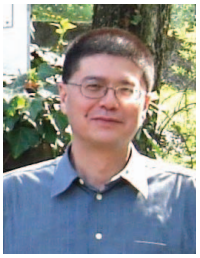
[29] Y. Gao, A. Barreto, and M. Adjouadi, "Monitoring and processing of the pupil diameter signal for affective assessment of a computer user," in *Proceedings of the 13th International Conference on Human-Computer Interaction. Part I: New Trends*. Berlin, Heidelberg: Springer-Verlag, 2009, pp. 49–58.

[30] J. J. M. Kierkels, M. Soleymani, and T. Pun, "Queries and tags in affect-based multimedia retrieval," in *ICME'09: Proceedings of the 2009 IEEE international conference on Multimedia and Expo*. Piscataway, NJ, USA: IEEE Press, 2009, pp. 1436–1439.

[31] M. Pantic and A. Vinciarelli, "Implicit Human-Centered Tagging," *IEEE Signal Processing Magazine*, vol. 26, no. 6, pp. 173–180, November 2009.
[32] J. A. Russell, "Culture and the Categorization of Emotions," *Psychological Bulletin*, vol. 110, no. 3, pp. 426–450, 1991.
[33] J. A. Russell and A. Mehrabian, "Evidence for a three-factor theory of emotions," *Journal of Research in Personality*, vol. 11, no. 3, pp. 273–294, September 1977.
[34] M. Soleymani, J. Davis, and T. Pun, "A collaborative personalized affective video retrieval system," in *Affective Computing and Intelligent Interaction and Workshops, 2009. ACII 2009. 3rd International Conference on*, sep 2009.
[35] A. Schaefer, F. Nils, X. Sanchez, and P. Philippot, "Assessing the effectiveness of a large database of emotion-eliciting films: A new tool for emotion researchers," *Cognition & Emotion*, vol. 24, no. 7, pp. 1153–1172, 2010.
[36] J. Lichtenauer, M. Valstar, J. Shen, and M. Pantic, "Cost-Effective Solution to Synchronized Audio-Visual Capture Using Multiple Sensors," in *AVSS '09: Proceedings of the 2009 Sixth IEEE International Conference on Advanced Video and Signal Based Surveillance*. Washington, DC, USA: IEEE Computer Society, 2009, pp. 324–329.
[37] R. Adolphs, D. Tranel, and A. R. Damasio, "Dissociable neural systems for recognizing emotions," *Brain and Cognition*, vol. 52, no. 1, pp. 61–69, June 2003.
[38] A. R. Damasio, T. J. Grabowski, A. Bechara, H. Damasio, L. L. B. Ponto, J. Parvizi, and R. D. Hichwa, "Subcortical and cortical brain activity during the feeling of self-generated emotions," *Nature Neuroscience*, vol. 3, no. 10, pp. 1049–1056, October 2000.
[39] R. J. Davidson, "Affective neuroscience and psychophysiology: toward a synthesis." *Psychophysiology*, vol. 40, no. 5, pp. 655–665, September 2003.
[40] L. I. Aftanas, N. V. Reva, A. A. Varlamov, S. V. Pavlov, and V. P. Makhnev, "Analysis of evoked EEG synchronization and desynchronization in conditions of emotional activation in humans: temporal and topographic characteristics." *Neuroscience and behavioral physiology*, vol. 34, no. 8, pp. 859–867, October 2004.
[41] P. D. Welch, "The Use of Fast Fourier Transform for the Estimation of Power Spectra: A Method Based on Time Averaging Over Short, Modified Periodograms," *IEEE Transactions on Audio and Electroacoustics*, vol. 15, pp. 70–73, 1967.
[42] S. K. Sutton and R. J. Davidson, "Prefrontal Brain Asymmetry: A Biological Substrate of the Behavioral Approach and Inhibition Systems," *Psychological Science*, vol. 8, no. 3, pp. 204–210, 1997.
[43] V. F. Pamplona, M. M. Oliveira, and G. V. G. Baranoski, "Photorealistic models for pupil light reflex and iridal pattern deformation," *ACM Trans. Graph.*, vol. 28, no. 4, pp. 1–12, 2009.
[44] A. Longtin and J. Milton, "Modelling autonomous oscillations in the human pupil light reflex using non-linear delay-differential equations," *Bulletin of Mathematical Biology*, vol. 51, no. 5, pp. 605–624, September 1989.
[45] H. Bouma and L. C. J. Baghuis, "Hippus of the pupil: Periods of slow oscillations of unknown origin," *Vision Research*, vol. 11, no. 11, 1971.
[46] F. H. Kanfer, "Verbal rate, eyeblink, and content in structured psychiatric interviews," *Journal of Abnormal and Social Psychology*, vol. 61, no. 3, pp. 341–347, 1960.
[47] D. Ruta and B. Gabrys, "An Overview of Classifier Fusion Methods," *Computing and Information Systems*, vol. 7, no. 1, pp. 1–10, 2000.
[48] L. I. Kuncheva, *Combining Pattern Classifiers: Methods and Algorithms*. Wiley-Interscience, july 2004.
[49] J. C. Platt, *Probabilities for SV Machines*. MIT Press, 2000, pp. 61–74.
[50] T. F. Wu, C. J. Lin, and R. C. Weng, "Probability Estimates for Multi-class Classification by Pairwise Coupling," *J. Mach. Learn. Res.*, vol. 5, pp. 975–1005, 2004.
[51] C. Chang and C. Lin, "LIBSVM: a Library for Support Vector Machines," 2001.
[52] G. Chanel, C. Rebetez, M. Bétrancourt, and T. Pun, "Emotion assessment from physiological signals for adaptation of game difficulty," *Systems, Man and Cybernetics, Part A: Systems and Humans, IEEE Transactions on*, vol. PP, no. PP, p. In Press, 2011.
[53] S. D. Kreibig, G. Schaefer, and T. Brosch, *Psychophysiological response patterning in emotion: Implications for affective computing*. Oxford, UK: Oxford University Press, 2010, ch. 2.4, pp. 105–130.
[54] D. Watson, L. A. Clark, and A. Tellegen, "Development and validation of brief measures of positive and negative affect: the PANAS scales." *Journal of personality and social psychology*, vol. 54, no. 6, pp. 1063–1070, June 1988.

Mohammad Soleymani received both his B.Sc. and M.Sc. degrees from the department of Electrical and Computer Engineering, University of Tehran, Iran, in 2003 and 2006 respectively. He is now a doctoral student and research assistant at the Computer Vision and Multimedia Laboratory, Computer Science Department, University of Geneva, Switzerland. His research interests include affective computing and multimedia information retrieval. He has been co-organizing the MediaEval multimedia benchmarking initiative since 2010. He is a member of IEEE and the HUMAINE association.

Maja Pantic is Professor in Affective and Behavioural Computing at Imperial College London, Department of Computing, UK, and at the University of Twente, Department of Computer Science, the Netherlands. She is one of the world's leading experts in the research on machine understanding of human behavior including vision-based detection, tracking, and analysis of human behavioral cues like facial expressions and body gestures, and multi-modal human affect/mental state understanding. She has published more than 100 technical papers in these areas of research. In 2008, for her research on Machine Analysis of Human Naturalistic Behavior (MAHNOB), she received European Research Council Starting Grant as one of 2% best young scientists in any research field in Europe. In 2011, Prof. Pantic was awarded Roger Needham Award made annually for a distinguished research contribution in computer science by a UK based researcher within ten years of their PhD. She currently serves as the Editor in Chief of Image and Vision Computing Journal and as an Associate Editor for the IEEE Transactions on Systems, Man, and Cybernetics Part B (TSMC-B). She has also served as the General Chair for several conferences and symposia including the IEEE FG 2008 and the IEEE ACII 2009. She is a Senior Member of the IEEE.



Thierry Pun (IEEE Member, EE Eng. 1979, PhD 1982) is head of the Computer Vision and Multimedia Laboratory, Computer Science Department, University of Geneva, Switzerland. He received his Ph.D. in image processing for the development of a visual prosthesis for the blind in 1982, at the Swiss Federal Institute of Technology, Lausanne, Switzerland. He was visiting fellow from 1982 to 1985 at the National Institutes of Health, Bethesda, USA. After being CERN Fellow from 1985 to 1986 in Geneva, Switzerland, he joined the University of Geneva, in 1986, where he currently is full professor at the Computer Science Department. He has authored or co-authored about 300 full papers as well as eight patents. His current research interests, related to affective computing and multimodal interaction, concern: physiological signals analysis for emotion assessment and brain-computer interaction, multi-modal interfaces for blind users, data hiding, multimedia information retrieval systems.