Affective Characterization of Movie Scenes Based on Multimedia Content Analysis and User's Physiological Emotional Responses

Soleymani, Mohammad; Chanel, Guillaume; Kierkels, Joep Johannes Maria; Pun, Thierry

Affective Characterization of Movie Scenes Based on Multimedia Content Analysis and User's Physiological Emotional Responses

Mohammad Soleymani Guillaume Chanel Joep J.M. Kierkels Thierry Pun
Computer Vision and Multimedia Laboratory, Computer Science Department
University of Geneva
Battelle Building A, Rte. De Drize 7,
CH - 1227 Carouge, Geneva, Switzerland
{mohammad.soleymani, guillaume.chanel, joep.kierkels, thierry.pun}@ unige.ch

Abstract

In this paper, we propose an approach for affective representation of movie scenes based on the emotions that are actually felt by spectators. Such a representation can be used for characterizing the emotional content of video clips for e.g. affective video indexing and retrieval, neuromarketing studies, etc. A dataset of 64 different scenes from eight movies was shown to eight participants. While watching these clips, their physiological responses were recorded. The participants were also asked to self-assess their felt emotional arousal and valence for each scene. In addition, content-based audio- and video-based features were extracted from the movie scenes in order to characterize each one. Degrees of arousal and valence were estimated by a linear combination of features from physiological signals, as well as by a linear combination of content-based features. We showed that a significant correlation exists between arousal/valence provided by the spectator's self-assessments, and affective grades obtained automatically from either physiological responses or from audio-video features. This demonstrates the ability of using multimedia features and physiological responses to predict the expected affect of the user in response to the emotional video content.

Keywords: Multimedia indexing and retrieval, affective personalization and characterization, emotion recognition and assessment, affective computing, physiological signals analysis.

1. Introduction

The amount of available digital multimedia content has greatly increased during the last decade. Powerful and novel multimedia indexing and retrieval methods have thus become essential to sift through such abundance. In this paper we propose to use the emotion that is actually felt by a given spectator as an indexing feature, in addition to more classical features like those based on video analysis of the media content. In order to demonstrate that for movie scenes affect can be represented by grades we compared self-assessment of the emotional content of scenes, with affective grades automatically estimated from physiological responses and multimedia content analysis.

The affective and emotional preferences of a user play an important role in multimedia content selection. Imagine you feel bored and you are looking for an entertaining movie. How can a system understand your affective preferences? What are your real affective preferences? These questions are hard to answer, because user emotional preferences depend on many aspects such as context, culture, sex, age, etc. A "personal content delivery" [1] system which considers one's emotional preferences should answer these needs. This paper introduces an affective representation method that can operate at the core of such a system.

To estimate affect, physiological responses are valued for not interrupting users for self reporting phases. In addition, affective self-reports might be held in doubt because the participant cannot remember all the different emotions he/she had during the experiment, and/or might misrepresent his/her feelings due to self presentation (*i.e.* the participant wants to show he/she is courageous whereas in reality he/she was scared) or for pleasing the experimenter [2]. Self-assessment is however necessary as ground truth, to show that the physiological measurements are valid and also to train the affect representation system. Finally, while self reports are unable to represent dynamic changes, physiological measurements give the ability of measuring the user responses dynamically [3].

Affect based video content characterization requires the understanding of the intensity and type of affect which is expected to be evoked in the user (audience) while watching a movie/video. There are only a limited number of studies on content-based affective representation/understanding of movies, and these

mostly rely on self-assessments or population averages to obtain the emotional content of a movie [1;4].

Wang and Cheong [4] used content-based audio and video features to classify basic emotions elicited by movie scenes. They classified audio, into music, speech and environment signals and processed them separately to shape an audio affective vector. They combined this vector with video-based features such as key lighting, and visual excitement to generate a scene affective vector, which was classified and labeled with emotions. Hanjalic et al. [1] introduced "personalized content delivery" as a valuable tool in affective indexing and retrieval systems. They first selected video- and audio- content based features based on their relation to the arousal and valence space that was defined as an affect model for affect ([5]; see also Section 2 of this paper). Combining these features, they then estimated arising emotions in this space. While the arousal and valence grades could be used separately for indexing, they combined those grades by following their temporal pattern in this arousal/valence space. This allowed determining an affect curve shown to be useful for extracting video highlights in a movie or sport video.

Affective systems require methods for automatically assessing user's emotional state. Computerized emotion assessment gained interest over the last years. Most of current methods focus on facial expressions and speech analysis. However, these methods cannot always be relied upon since users are not always speaking or turning their head towards the camera lens. With the advancement of wearable systems for recording peripheral physiological signals, it is becoming more practically feasible to employ these signals in an easy-to-use human computer interface [6;7]. We therefore concentrated on the use of peripheral physiological signals for assessing emotion, namely: galvanic skin resistance (GSR), blood pressure which provided heart rate, respiration pattern, and skin temperature. In order to record facial muscles activity we also used electromyograms (EMG) from the Zygomaticus major and Frontalis muscles. At this stage of the study, we opted for not using electroencephalograms (EEG) due to the cumbersomeness of the apparatus and acquisition protocols, although EEG's have been shown to be very useful for assessing emotions [6;8-11].

This paper demonstrates a first step towards benefiting from actual physiological responses for creating affect-based tools. Personalized emotional profiles can be determined and subsequently used for affect based video indexing. Peripheral physiological signals were first recorded for monitoring the arousal/valence grades of participants' emotion while they were watching a movie scene. In order to understand the user's emotional behavior, sets of features extracted from the physiological signals were linearly combined to obtain an estimate for the arousal and valence grades. These grades, assessed while watching movie scenes, can be used as a new dimension of information in a user's personal affective profile. Multimedia content-based features were also extracted from the scenes by audio and video processing. The correlation between the self-assessed arousal/valence values and those computed from physiological features was determined, as well as the correlation between these self-assessed arousal/valence values and those obtained from multimedia features. The correlation between the physiological signals and the multimedia features was also investigated to determine which multimedia features give rise to which type of emotion. All correlations are shown to be significant: physiological responses of participants can characterize video scenes, and audio-visual features can fairly reliably be used to predict the spectator's felt emotion. The variation between participants of those content-based features that were the most correlated with self-assessment demonstrates the need for considering personal preferences in affective indexing of multimedia contents. Finally it can be noted that we did not focus on temporal changes in arousal and valence space, rather we investigated the average affect related to each movie segments of interest (scenes).

The remainder of this paper is organized as follows. Section 2 presents some background on representation of affect and on the arousal/valence model to represent emotions. Section 3 elaborates on data acquisition, feature extraction and selection, and how features are combined for representation. The experimental results are given in Section 4 and finally conclusions are presented in Section 5.

2. Affective representation

Emotions are not discrete phenomena but rather continuous ones. Psychologists therefore 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 space. Valence represents the way one judges a situation, from unpleasant to pleasant; arousal expresses the degree of felt excitement, from calm to exciting. Cowie used the valence/activation space (similar to the valence/arousal space) to model and assess emotions from speech [7;12]. Although such spaces do not provide any verbal description, a point in such space can be mapped to a categorical feeling label.

In order to record their felt emotions, participants were asked to grade each movie scene by arousal and valence grades using self-assessment Manikins (SAM) [13]. The arousal grade represented the level of arousal or excitement felt when watching the scene while the valence grade represents the felt pleasantness.

3. Material and methods

3.1. Overview

A video dataset of 64 movie scenes was created (see Section 3.3) from which content-based low-level features were extracted. Experiments were conducted during which physiological signals were recorded from spectators. After each scene, the spectator self-assessed his/her arousal and valence levels. To reduce the mental load of the participants, the protocol divided the show into 2 sessions of 32 movie scenes each. Each of these sessions lasted approximately two hours, including setup. Eight healthy participants (three female and five male, from 22 to 40 years old) participated in the experiment. Thus, after finishing the experiment three types of affective information about each movie clip were available:

- multimedia content-based information extracted from audio and video signals;
- physiological responses from spectators' bodily reactions (due to the autonomous nervous system) and from facial expressions;
- self-assessed arousal and valence, used as 'ground truth' for the true feelings of the spectator.

Since video scenes were showed in random order, the occurrence of high and low arousal and valence values in the self-assessed vectors (64 elements each) does not depend on the order in which scenes were presented.

Next, we aim at demonstrating how those true feelings about the movie scenes can be obtained by using the information that is either extracted from audio and video signals or contained within the recorded physiological signals. To this end, features that are likely to be influenced by affect have been extracted from the audio and video content as well as from the physiological signals. Thus a (single) feature vector composed of 64 elements highlights a single characteristic (for instance, average sound energy) of the 64 movie scenes. In a similar way feature vectors were extracted from the physiological signals. As one may expect, a single feature, e.g. average sound energy, may not be equally relevant to the affective feelings of different participants. In order to personalize the set of all extracted features, an additional operation called relevant-feature selection has been implemented. During the relevant-feature selection for arousal, the correlation between the single-feature vectors and the self-assessed arousal vector is determined. Only the features with high absolute correlation coefficient ($|\rho|$ above 0.25 and p-value below 0.05) were subsequently used for estimating arousal. A similar procedure was performed for valence. It will be shown that accurate estimates of the self-assessed arousal and valence can be obtained based on the relevant feature vectors for physiological signals as well as from the relevant feature vectors for audio and video information.

3.2. Experiments

The participants were first informed about the experiment, the meaning of arousal and valance, the self-assessment procedure, and the video content. In emotional-affective experiments the bias of the emotional state (participants' mood) needs to be removed. To allow leveling of feature values over time a baseline is recorded at each trial start by showing one short 30s. neutral clip randomly selected from clips provided by the Stanford psychophysiology laboratory [14].

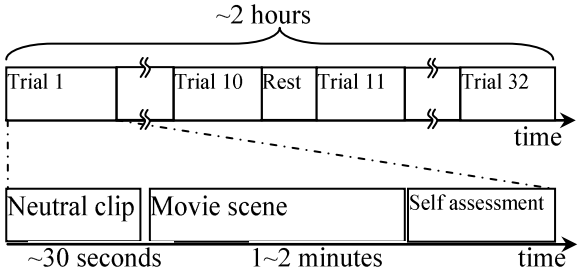

Figure 1. Experimental protocol.

Figure 1 presents the experimental protocol and its timing. Each trial started with the user pressing the "I am ready" key which started the neutral clip playing. After watching the neutral clip, one of the movie scenes was played. Movie scenes were selected from the dataset in random order. After watching the movie scene, the participant filled in the self-assessment form which popped up automatically. In total, the time interval between the starts of consecutive trials was approximately three to four minutes. This interval included playing the neutral clip, playing the selected scene, performing the self-assessment, and the participant-controlled rest time.

In the self-assessment step for evaluating arousal and valence, the SAM Manikin pictures with a slider to facilitate self-assessment of arousal and valence were used (see Figure 2). The sliders correspond to a numerical range of [0, 1] while the numerical scale was not shown to the participants.

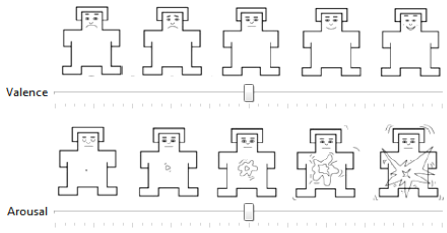

Figure 2. Arousal and valence self-assessment: SAM manikins and sliders.

3.3. Data

3.3.1. Movie scenes dataset

To create the video dataset, we extracted video scenes from eight movies selected either according to similar studies (e.g. [1;4;6;14]), or from recent famous movies. The movies included four major genres: drama, horror, action, and comedy. Video clips used for this study are from the following: Saving Private Ryan (action), Kill Bill, Vol. 1 (action), Hotel Rwanda (drama), The Pianist (drama), Mr. Bean's Holiday (comedy), Love Actually (comedy), The Ring, Japanese version (horror) and 28 Days Later (horror). The extracted scenes, eight for each movie, had durations of approximately one to two minutes each and contained an emotional event (judged by the authors).

3.3.2. Physiological signals

Peripheral signals and facial expression EMG signals were recorded for emotion assessment. EMG signals from the right Zygomaticus major muscle (smile, laughter) and right Frontalis muscle (attention, surprise) were used as indicators of facial expressions. Galvanic skin resistance (GSR), skin temperature, breathing pattern (using a respiration belt) and blood pressure (using a plethysmograph) were also recorded. All physiological data was acquired via a Biosemi Active-two system with active electrodes, from Biosemi Systems (http://www.biosemi.com). The data were recorded with a sampling frequency of 1024 Hz in a sound-isolated Faraday cage. Examples of recorded physiological signals in a surprising scene are given in Figure 3. The GSR and respiration signals were respectively smoothed by a 512 and a 256 points averaging filters to reduce the high frequency noise. EMG signals were filtered by a Butterworth band pass filter with a lower cutoff frequency of 4 Hz and a higher cutoff frequency of 40 Hz.

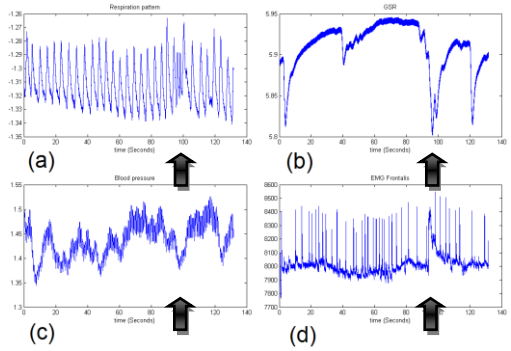

Figure 3. Physiological response (participant 2) to a surprising action scene. The following raw physiological signals are shown: respiration pattern (a), GSR (b), blood pressure (c), and Frontalis EMG (d). The surprise moment is indicated by an arrow.

3.4. Feature extraction

3.4.1. Audio and video content-based features

Sound has an important impact on user's affect. For example according to the findings of Picard [15], loudness of speech (energy) is related to evoked arousal, while rhythm and average pitch are related to valence. The audio channels of the movie scenes were extracted and encoded into monophonic information (MPEG layer 3 format) at a sampling rate of 48 kHz, and their amplitude range was normalized in [-1, 1]. All of the resulting audio signals were normalized to the same amplitude range before further processing. A total of 79 low-level audio features were determined for each of the audio signals. These features, listed in Table 1, are commonly used in audio and speech processing and audio classification [16;17].

Wang et al [4] demonstrated the relationship between audio type's proportions and affect, where these proportions refer to the respective duration of music, speech, environment, and silence in the audio signal of a video clip. To determine the three important audio types (music, speech, and environment), silence was first identified by comparing the audio signal energy of each sound sample with a pre-defined threshold empirically set at 5×10^{-7}. After removing silence, the remaining audio signals were classified by the three classes support vector machine (SVM). We implemented a three class audio type classifier using support vector machines (SVM with polynomial kernel) operating on audio low-level features in a time window of one second. Despite some classes overlapping (e.g. presence of a musical background during a dialogue), the classifier was usually able to recognize the dominant audio type. The SVM was trained utilizing more than 3 hours of audio, extracted from movies and labeled manually. The classification

results were used to form 4 bins (3 audio types and silence) normalized histogram; these histogram values were used as affective features for the affective representation. MFCC (Mel frequency cepstral coefficients), LPCC (Linear prediction cepstral coefficients) and the pitch of audio signals were extracted using the PRAAT software package [18].

Movie scenes have been segmented at the shot level using the OMT shot segmentation software [19]. Video clips were encoded into MPEG-1 format to extract motion vectors and I frames for further feature extraction. We used the OVAL library (Object-based Video Access Library) [20] to capture video frames and extract motion vectors.

From a movie director's point of view, lighting key [4;22] and color variance [22] are important parameters to evoke emotions. We therefore extracted lighting key from frames in the HSV space by multiplying the average value (V in HSV) by the standard deviation of the values (V in HSV). Color variance was obtained in the CIE LUV color space by computing the determinant of the covariance matrix of L, U, and V.

The average shot change rate, and shot length variance were extracted to characterize video rhythm. Hanjalic et al. [1] showed the relationship between video rhythm and affect. Fast object movements in successive frames are also an effective factor to evoke excitement. To measure this factor, the motion component was defined as the amount of motion in consecutive frames computed by accumulating magnitudes of motion vectors for all B and P frames.

Colors and their proportions have an effect to elicit emotions. In order to use colors in the list of video features, a 20 bin color histogram of hue and lightness values in the HSV space was computed for each I frame and subsequently averaged over all frames. The resulting averages in the 20 bins were used as video content-based features. The median of L value in HSL space was computed to obtain the median lightness of a frame. Shadow proportion or the proportion of dark area in a video frame is another feature which relates to affect [4]. Shadow proportion is determined by comparing the lightness values in HSL color space with an empirical threshold. Pixels with lightness level below this threshold (0.18 [4]) are assumed to be dark and in shadow in the frame.

Table 1. Low-level features from audio signals.

Feature set	Extracted features
MFCC	MFCC coefficients, derivative and autocorrelation of MFCC, each 13 features [16]
Energy	Average energy of the audio signal [16]
LPCC	LPCC (16 features), derivative of LPCC (16 features), [16]
Time frequency	Spectrum flux, spectral centroid, delta spectrum magnitude, band energy ratio, dominant pitch frequency[16;17]
ZCR	Zero crossing rate [16]
Silence ratio	Proportion of silence in a time window [21]

3.4.2. Physiological features

GSR provides a measure of the resistance of the skin by positioning two electrodes on the tops of two fingers and passing a negligible current through the body. This resistance decreases due to an increase of sudation, which usually occurs when one is experimenting emotions such as stress or surprise. Moreover, Lang et al. discovered that the mean value of the GSR is related to the level of arousal [23]. (See Table 2 which summarizes the list of features extracted from physiological signals.)

A plethysmograph measures blood pressure in the participant's thumb. This measurement can also be used to compute heart rate by identification of local maxima (i.e. heart beats) and inter-beat periods. Blood pressure and heart rate variability correlate with emotions, since stress can increase blood pressure [7]. Pleasantness of stimuli can increase peak heart rate response [23], and heart rate variability decreases with fear, sadness, and happiness [24].

Table 2. Features from peripheral signals.

Peripheral signal	Extracted features
GSR	Average skin resistance, average of derivative, mean of derivative for negative values only(average decrease rate during decay time), proportion of negative samples in the derivative vs. all samples
Blood flow (Plethysmograph)	Average blood pressure, heat rate, heart rate derivative, heart rate variability, standard deviation of heart rate
Respiration	Band energy ratio (energy ratio between the lower (0.05-0.25Hz) and the higher (0.25-5Hz) bands), average respiration signal, mean of derivative (variation of the respiration signal), standard deviation, dynamic range or greatest breath, breathing rhythm (spectral centroid)
EMG Zygomaticus	Energy
EMG Frontalis	Energy
Eye blinking rate	Rate of eye blinking per second, extracted from the Frontalis EMG
Skin Temperature	Range, average, minimum, maximum, standard deviation

Skin temperature was also recorded since it changes in different emotional states [23]. The respiration pattern was measured by tying a respiration belt around the chest of the participant. Slow respiration is linked to relaxation while irregular rhythm, quick variations, and cessation of respiration correspond to more aroused emotions like anger or fear [24;25].

Regarding the EMG signals, the Frontalis muscles activity is a sign of attention or stress in facial expressions. The activity of the Zygomaticus major was also monitored, since this muscle is active when the user is laughing or smiling [26]. Most of the power in the spectrum of an EMG during muscle contraction is in the frequency range between 4 to 40 Hz. Thus, the muscle activity features were obtained from the energy of EMG signals in this frequency range for the different muscles.

The rate of eye blinking is another feature, which is correlated with anxiety [27]. Eye-blinking affects the EMG signal that is recorded over the Frontalis muscle and results in easily detectable peaks in that signal.

3.5. Feature selection and regression

The relevance of features for affect was determined using linear correlation between each extracted feature and the users' self-assessment, as motivated in Section 3.1 In this study, a significant correlation between two vectors was supposed to exist when the absolute correlation exceeded 0.25 ($|\rho| > 0.25$) with p-value below 0.05. The p-value represents the probability that a randomly selected vector would lead to a ρ value that is at least as large as the one observed.

We now demonstrate how user-felt arousal and valence can be estimated, based on the physiological or content-based features which were found to have a significant correlation with the self-assessed valence and arousal. For each participant, a training set of 42 scenes was formed by randomly selecting 42 of the 64 movie scenes and the corresponding feature values. The remaining 22 scenes served as a test set.

In order to obtain an estimate, based on the significantly correlated features, of the user's arousal and valence, all significantly correlated features are weighted and summed as is indicated in Eq. (1), where $\hat{y}(j)$ is the estimate of arousal/valence grade, j is the indexing number of a specific movie scene $\{1,2,..,64\}$, $x_i(j)$ is the feature vector corresponding to the i-th significantly correlated feature, N_s is the total number of significant features for this participant, and w_i is the weight that corresponds to the i-th feature.

$$\hat{y}(j) = \sum_{i=1}^{N_s} w_i x_i(j) + w_0 \qquad (1)$$

In order to determine the optimum $\hat{y}$, the weights in Eq. (1) were computed by means of a linear relevance vector machine (RVM) from the Tipping RVM toolbox [28]. This procedure was applied on the user self assessed arousal/valence, $y(j)$, and on the feature-estimated arousal/valence, $\hat{y}(j)$, over all 42 movie scenes in the test set as can be seen in (2).

This procedure is performed four times for optimizing the weights corresponding to:

- physiological features when estimating valence,
- physiological features when estimating arousal,
- multimedia features when estimating valence,
- multimedia features when estimating arousal.

In a first step weights are computed from the training set. In the second step, the obtained weights were applied to the test set, and the mean squared error between the resulting estimated arousal/valence grades and self assessed arousal/valence was examined. These two steps were repeated 1000 times. Each time the 42 movie scenes of the training set were randomly selected from the total of 64 scenes while the 22 remaining scenes served as the test set. The results from this cross-validation will be presented in next Section.

4. Experimental results

The correlations between multimedia features, physiological features and self assessments were determined. Table 3 shows, for each participant, the features which had the highest absolute correlations with that participant's self-assessments of arousal and valence. Table 3.a shows results for physiological features whereas Table 3.b shows results for multimedia features

Table 3. Physiological and multimedia features with the highest absolute correlation with self assessments for participants 1 to 8.

(a) Physiological features

	Arousal	ρ	Valence	ρ
1	EMG Frontalis	0.39	EMG Zygomaticus	0.66
2	EMG Frontalis	0.57	EMG Frontalis	-0.63
3	Respiration band energy ratio	0.42	EMG Zygomaticus	0.58
4	Blood pressure	-0.29	EMG Zygomaticus.	0.43
5	EMG Zygomaticus	0.46	EMG Frontalis	-0.47
6	Eye blinking rate	-0.32	Average of GSR derivative.	-0.45
7	GSR standard deviation	0.55	EMG Zygomaticus	0.69
8	Blood pressure	-0.33	EMG Zygomaticus	0.56

(b)Multimedia Features

1	13th LPC coefficient	-0.35	Last MFCC coeff.	0.50
2	Last MFCC coefficient	-0.54	14th bin of hue histogram (bluish)	0.43
3	Audio signal energy	-0.4	Last MFCC coefficient	0.5
4	First autocorrelation MFCC coefficient	0.40	3rd autocorrelation MFCC Coefficient	0.35
5	Motion component	0.32	Motion component	-0.47
6	11th autocorrelation MFCC coefficient	-0.43	5th bin of lightness histogram	-0.39
7	12th autocorrelation MFCC coefficient	0.45	Key lighting	0.41
8	Motion component	0.38	15th bin of hue histogram (purplish)	-0.48

For physiological signals, the variation of correlated features over different subjects illustrates the difference between participants' responses. While blood pressure was more informative regarding the arousal level of participants 4 and 8, EMG signals and thus facial expressions were more important to estimate arousal in participants 1, 2, and 5. The large variation between participants regarding which multimedia features have the highest absolute correlation value with their self assessment, indicates the variance in individual preferences to different audio or video features. For instance more motion component leads to more arousal and excitement and less valence and pleasantness for participant 5, which means that the participant had a negative feeling for exciting scenes with large amount of movement in objects or background.

Table 4 shows, for all participants, the correlation coefficients between four different pairs of physiological features and multimedia features. These eight features have been chosen from the features which have significant correlation with self assessments and thus more importance for affect characterization. The correlations show that physiological responses are significantly correlated to changes in multimedia content. As an example, the negative correlation between EMG Zygomaticus energy and the 15th bin of the hue histogram (corresponding to purple) shows that increasing this color in the video content results in less Zygomaticus activity, thus less pleasantness or valence.

Table 4. The linear correlation ρ values btw. multimedia features, and physiological features which are significantly correlated with self assessments (participants 1 to 8).

	EMG Zygomatic. energy/Key lighting	Skin temp. standard deviation /5th MFCC autocorrelation coefficient	Skin temp. range/ Shot length variation	EMG Zygomatic. energy/ hue histogram's 15th bin
1	0.24	-	-	-0.41
2	0.62	0.44	0.42	-0.41
3	0.46	0.40	0.56	-0.34
4	0.40	0.32	0.43	-0.30
5	0.36	0.39	0.58	-
6	0.44	0.31	0.51	-0.32
7	0.47	0.34	0.27	-0.43
8	0.54	0.34	0.42	-0.45

Table 5. Average mean squared error (E_{MSE}), between estimated arousal/valence grades and self assessments (participants 1 to 8).

	Arousal with physiological features	Arousal with Multimedia features	Valence with physiological features	Valence with multimedia features
1	0.044	0.047	0.020	0.031
2	0.030	0.038	0.020	0.032
3	0.034	0.034	0.026	0.043
4	0.037	0.036	0.023	0.023
5	0.028	0.031	0.060	0.047
6	0.043	0.040	0.052	0.037
7	0.025	0.032	0.018	0.026
8	0.031	0.027	0.014	0.017

The accuracy of the estimated arousal/valence is evaluated by computing the mean squared error between the estimates and the self assessments of arousal/valence (Table 5). The mean squared error (MSE) was calculated 1000 times when varying the 22 samples in the test set, using the cross validation technique discussed in section 3.5.

$$E_{MSE} = \frac{1}{1000 \times N_{test}} \sum_{i=1}^{1000} \sum_{j=1}^{N_{test}=22} (\hat{y}_{ij} - y_{ij})^2 \qquad (2)$$

The MSE was computed by Eq. 2 where N_{test} is the number of test samples (here 22) and $\hat{y}_{ij}$ is the estimated arousal/valence in *i*-th iteration for *j*-th sample in test set. The computation used the obtained grades from both physiological features and multimedia content features of each subject. Since it was easier to self assess valence on the video dataset, better results have been obtained for valence estimation. All MSE values are considerably smaller than a random level estimation MSE (around 0.17).

5. Conclusion

In this paper, an affective characterization method for movie scenes is proposed based on emotions that are felt by spectators. Physiological responses of participants were recorded while watching movie scenes and key features were extracted from these responses. By computing correlations between these key physiological features and the users' self-assessment of arousal and valence, it was identified which physiological features are essential for accurate estimation of arousal/valence. Such accurate estimates provide us with a continuous assessment of affect which can serve as a ground truth for affect estimation. For example Zygomaticus EMG signals which represent smile and laughter have high correlation with valence (Table 3).

Furthermore, content based multimedia features were extracted from the movies scenes. Their correlations with both physiological features and users' self-assessment of arousal/valence were shown to be significant. A procedure was proposed to actually estimate user's affect in response to movie scenes based on selected multimedia content features. Predicting user's affect opens the door to many novel applications. One is personalized content delivery systems with configurable emotional-based

preferences. Users will watch a training set of short movie clips; after configuration, the system will be able to predict the users' response to new movie scenes. A similar strategy is applicable to neuromarketing where consumers' reactions to marketing stimuli could be predicted.

The movie scenes did not necessarily correspond to very strong emotions; some of them contained just mild and tranquil scenes. These were intentionally selected because the final application was not only to characterize affect, but also to show the ability to estimate different amplitudes of emotions. The final application will have to index all types of different movie scenes from highly intense ones to calm and fairly neutral.

Felt emotions from the movie scenes where determined without any a priori assumptions on arousal/valence values. It would however be possible to use the genre of movies (e.g., drama, comedy, etc.) as prior knowledge for better affect estimation.

Participants exhibit markedly different emotional reactions to movie scenes. These differences can be explained by different factors, e.g., personalities, general mood during experiments, or varying personal standards for self-assessment of true feelings. This shows the need for affect profiling to be, at least in part, user-dependent. The exact physiology behind emotional processes is still under debate. We do not intend in this work to explain affective mechanisms in the brain, but rather to employ the widely accepted measures of valence and arousal as features for multimodal human-computer interaction and for affective video characterization. In the future we aim at more precisely assessing which are the most important content-based multimedia features able to elicit specific emotions. Studies involving more participants are also needed to determine which emotional responses are individual and which are common to all users.

6. Acknowledgement

The authors gratefully acknowledge the support of the Swiss National Science Foundation and of the EU Network of Excellence Similar. The authors also thank Drs. S. Marchand-Maillet, E. Bruno, and D. Grandjean for their valuable scientific comments, and for enabling us to use their software and datasets during this work.

7. References

[1] A. Hanjalic and L. Q. Xu, "Affective video content representation and modeling," *IEEE Trans. on Multimedia*, vol. 7, no. 1, pp. 143-154, 2005.

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

[3] K. Boehner, R. DePaula, P. Dourish, and P. Sengers, "How emotion is made and measured," *Int. J. of Human-Computer Studies*, vol. 65, no. 4, pp. 275-291, Apr.2007.

[4] H. L. Wang and L. F. Cheong, "Affective understanding in film," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 16, no. 6, pp. 689-704, June2006.

[5] J. A. Russell and A. Mehrabian, "Evidence for A 3-Factor Theory of Emotions," *J. of Research in Personality*, vol. 11, no. 3, pp. 273-294, 1977.

[6] A.Benoit, L.Bonnaud, A.Caplier, P.Ngo, L.Lawson, D.Trevisan, V.Levacik, C.Mancas, and G.Chanel, "Multimodal Focus Attention and Stress Detection and Feedback in an Augmented Driver Simulator,", 3rd IFIP Conf. Artif. Intell. Applicat & Innov., Athens, Greece: 2006.

[7] J.A. Healey, "Wearable and Automotive Systems for Affect Recognition from Physiology.", Massachusetts Institute of Technology, May 2000.

[8] G. Chanel, J. Kronegg, D. Grandjean, and T. Pun, "Emotion assessment: Arousal evaluation using EEG's and peripheral physiological signals,", Springer LNCS, vol. 4105, pp 530-537, Sept. 2006.

[9] G. Chanel, K. Ansari-Asl, and T. Pun, "Valence-arousal evaluation using physiological signals in an emotion recall paradigm,", IEEE SMC Montreal, Oct. 2007.

[10] K. Ansari-Asl, G. Chanel, and T. Pun, "A channel selection method for EEG classification in emotion assessment based on synchronization likelihood,", Eusipco 2007, Poznan, Poland, Sept. 2007.

[11] K. Takahashi, "Remarks on Emotion Recognition from Bio-Potential Signals,", 2nd Conference on Autonomous Robots and Agents, New Zealand, Dec. 2004.

[12] J. A.Russel, M. Lewicka, and T. Niit, "A Cross-Cultural Study of a Circumplex Model of Affect," *J. of Personality and Social Psychology*, vol. 57, no. 5, 1989.

[13] J. D. Morris, "Observations: SAM: The self-assessment manikin - An efficient cross-cultural measurement of emotional response," *J. of Advertising Research*, vol. 35, no. 6, pp. 63-68, 1995.

[14] J.Rottenberg, R.D.Ray, and J.J.Gross, "Emotion elicitation using films," in *The handbook of emotion elicitation and assessment.* A.Coan and J.J.B.Allen, Eds. London: Oxford University Press, 2007.

[15] R. W.Picard, *Affective computing* The MIT press, 1997.

[16] D. G. Li, I. K. Sethi, N. Dimitrova, and T. Mcgee, "Classification of general audio data for content-based retrieval," *Pattern Recognition Letters*, vol. 22, no. 5, pp. 533-544, 2001.

[17] L. Lu, H. Jiang, and H. Zhang, "A Robust Audio Classification and Segmentation Method,", ACM int. conf. on Mult., pp. 203-211, Sept. 2001.

[18] P. Boersma and D. Weenink, "Praat: doing phonetics by computer," computer program, 2008.

[19] B. Janvier, E. Bruno, T. Pun, and S. Marchand-Maillet, "Information-theoretic temporal segmentation of videos and applications: multiscale keyframe selection and transition detection," *Mult. Tools and Ap.*, vol. 30, pp. 273-288, 2006.

[20] N. Moënne-Loccoz, "OVAL: an object-based video access library to facilitate the development of content-based

video retrieval systems," Viper group, Computer Vision and Multimedia Laboratory, Univ. of Geneva, 03.04, 2004.
[21] C. Lei, S. Gunduz, and M. T. Ozsu, "Mixed Type Audio Classification with Support Vector Machine,", IEEE Multimedia and Expo, pp. 781-784, July 2006.
[22] Z. Rasheed, Y. Sheikh, and M. Shah, "On the use of computable features for film classification," *IEEE Trans. on Circuits and Systems for Video Technology*, vol. 15, no. 1, pp. 52-64, Jan. 2005.
[23] 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.
[24] P. Rainville, A. Bechara, N. Naqvi, and A. R. Damasio, "Basic emotions are associated with distinct patterns of cardiorespiratory activity," *Int. J. of Psychophysiology*, vol. 61, no. 1, pp. 5-18, 2006.
[25] J. Kim, "Emotion Recognition from Physiological Measurement,", Humaine European Network of Excellence Workshop Santorini, Greece: Sept. 2004.
[26] G.-B.Duchenne de Boulogne and R.Andrew Cuthbertson, *The Mechanism of Human Facial Expression* Cambridge University Press, 1990.
[27] F. H. Kanfer, "Verbal Rate, Eyeblink, and Content in Structured Psychiatric Interviews," *J. of Abnormal and Social Psychology*, vol. 61, no. 3, pp. 341-347, 1960.
[28] M. E. Tipping, "Sparse Bayesian learning and the relevance vector machine," *J. of Machine Learning Research*, vol. 1, no. 3, pp. 211-244, 2001.