Affective Ranking of Movie Scenes Using Physiological Signals and Content Analysis

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

How to cite

SOLEYMANI, Mohammad et al. Affective Ranking of Movie Scenes Using Physiological Signals and Content Analysis. In: Proceedings of the 2nd ACM Workshop on Multimedia semantics, MS´08. Vancouver (Canada). [s.l.] : Association for Computing Machinery (ACM), 2008. doi: 10.1145/1460676.1460684

This publication URL: https://archive-ouverte.unige.ch//unige:47673

Publication DOI: 10.1145/1460676.1460684

Affective Ranking of Movie Scenes Using Physiological Signals and Content Analysis

Mohammad Soleymani Guillaume Chanel Joep J.M. Kierkels Thierry Pun
Computer Science Department
University of Geneva
Battelle campus, Building A, 7 Route de Drize
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 ranking of movie scenes based on the emotions that are actually felt by spectators. Such a ranking can be used for characterizing the affective, or emotional, content of video clips. The ranking can for instance help determine which video clip from a database elicits, for a given user, the most joy. This in turn will permit video indexing and retrieval based on affective criteria corresponding to a personalized user affective profile.
A dataset of 64 different scenes from 8 movies was shown to eight participants. While watching, their physiological responses were recorded; namely, five peripheral physiological signals (GSR - galvanic skin resistance, EMG - electromyograms, blood pressure, respiration pattern, skin temperature) were acquired. After watching each scene, the participants were asked to self-assess their felt arousal and valence for that scene. In addition, movie scenes were analyzed in order to characterize each with various audio- and video-based features capturing the key elements of the events occurring within that scene.
Arousal and valence levels were estimated by a linear combination of features from physiological signals, as well as by a linear combination of content-based audio and video features. We show that a correlation exists between arousal- and valence-based rankings provided by the spectator's self-assessments, and rankings obtained automatically from either physiological signals or audio-video features. This demonstrates the ability of using physiological responses of participants to characterize video scenes and to rank them according to their emotional content. This further shows that audio-visual features, either individually or combined, can fairly reliably be used to predict the spectator's felt emotion for a given scene. The results also confirm that participants exhibit different affective responses to movie scenes, which emphasizes the need for the emotional profiles to be user-dependant.

Categories and Subject Descriptors
H.3.3 **[Information Search and Retrieval]:** Information filtering, Selection process

General Terms
Algorithms, Measurement, Experimentation, Human Factors.

MS'08, October 31, 2008, Vancouver, British Columbia, Canada.

Keywords
Multimedia indexing and retrieval, affective personalization and ranking, emotion recognition and assessment, affective computing, physiological signals.

1. INTRODUCTION
Due to the ever increasing amount of digital multimedia content and searching relevant content within the existing content is becoming more and more difficult. Interactive and novel multimedia indexing and retrieval solutions become more essential to manage multimedia databases. 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 affect can be used for ranking movie scenes, we compared self-assessment of the emotional content of scenes with emotion that is estimated from physiological responses and with emotion that is estimated from multimedia content analysis.
The affective and emotional preferences of a user play an important role in multimedia content selection. Imagine that you missed a part of your favorite TV show and you want to take a brief look at what happened in the missed episode, or 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" [14] system which considers one's emotional preferences should answer these needs. This paper introduces an affective ranking method that can operate at the core of such a system.
Measurement and comparison between affective states based on numerical values is impracticable, because no precise quantitative scale allowing to measure affect exists. Meanwhile, it is possible to qualitatively compare affect by expressing the fact that one emotional state was more exciting or more pleasant than another. Affective ranking is thus proposed as a criterion to be used for similarity measurements and for affective indexing.
To assess emotion, 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 [26]. Finally, while self reports are unable to represent dynamic changes, physiological measurements give the ability of measuring the user responses dynamically [3].Self-assessment is however necessary as ground truth, to show that

the physiological measurements are valid and also to train the affect representation system.

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 [6][14][33]. Wang and Cheong [33] 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 treated them separately to shape an affective vector. They used the audio affective vector with video-based features such as key lighting, visual excitement to generate a scene affective vector. The scene affect vector was classified and labeled with emotions.

Hanjalic et al [14] 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 [30]; 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, which is 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 depended upon since users are not always speaking or turning their head towards the front of a camera. With the advancement of wearable systems for recording peripheral physiological signals, it has become practically feasible to employ these signals in an easy-to-use human computer interface [2][15]. We therefore concentrated on the use of peripheral physiological signals for assessing emotion. We used galvanic skin resistance (GSR), blood pressure which provided heart rate, respiration pattern, and skin temperature. In order to record facial expressions we also used electromyograms (EMG) from the Zygomaticus major and Frontalis muscles[1]. At this stage of the study, we opted for not using EEG - electroencephalograms due to the cumbersomeness of the apparatus and acquisition protocols, although EEG's have been shown to be very useful for assessing emotions [1][7][8][19][31] .

This paper demonstrates a first step towards benefiting from physiological responses to determine personalized emotional profiles and subsequently to permit affect based video indexing. Peripheral physiological signals were recorded for monitoring the arousal/valence level 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 levels. These grades assessed while watching movie scenes can be used as a new dimension of information in the 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. Many of the 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 had the highest correlation 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 rest 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 how features are combined for ranking. The experimental results are given in Section 4 and finally conclusions are presented in Section 5.

2. AFFECTIVE REPRESENTATION AND EMOTION MODELS

In order to better analyze emotions, one *s*hould know the processes that lead to emotional activation, how to model emotions and what are the different expressions of emotions. Three of the emotions viewpoints that Cornelius [10] cites are the Darwinian, cognitive and Jamesian ones. The Darwinian theory suggests that emotions are selected by nature in term of their survival value, e.g. fear exists because it helps avoid danger. The cognitive theory states that the brain is the centre of emotions. It particularly focuses on the "direct and non reflective" process, called appraisal [5], by which the brain judges a situation or an event as good or bad. Finally the Jamesian theory stipulates that emotions are only the perception of bodily changes such as heart rate or dermal responses ("I am afraid because I shiver"). Although controversial, this later approach emphasizes the important role of physiological responses in the study of emotions.

These different theories lead to different models. Inspired by the Darwinian theory, Ekman demonstrates the universality of six facial expressions [13]: happiness, surprise, anger, disgust, sadness and fear. Emotions however 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, originating 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 excitement felt by people, from calm to exciting (Figure 1).

Cowie used the valence/activation space to model and assess emotions from speech, which is similar to the valence/arousal space, [6][15][30]. Although both spaces do not provide any verbal description, it is possible to map a point in this space to a categorical feeling label. In the present study it was chosen to model emotions in the valence/arousal space, because this representation seems closer to real feelings, and gives the

[1] The Zygomaticus major muscles extend from each zygomatic arch (cheekbones) to the corners of the mouth, while the Frontalis muscles are located above the eyebrows on the forehead.

possibility to extract emotion labels from a continuous representation.

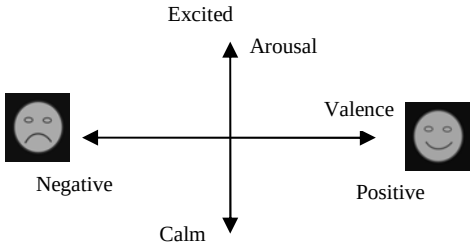

Figure 1 The arousal and valence two dimensional space.

In order to rank expected affects from movie scenes, participants were asked to to grade each movie scene by arousal and valence grades using self-assessment Manikins (SAM)[24]. These grades were used for ranking video scenes from the most exciting to the calmest, in correspondence to the arousal dimension. In a similar way, they were asked to rank the scenes from the happiest (the most pleasant) to the saddest (the most unpleasant), corresponding to the valence dimension. These rankings based on user self-assessed emotional content will then be compared to similar affective rankings obtained automatically from either physiological signals or multimedia content. The scales of arousal and valence values are not linear, *i.e.* the fact that the arousal value increases by a factor of two does not mean the actual arousal increases by the same factor. For this reason, we ranked the movie scenes instead of using a continuous value for valence and arousal. It should be noted that the two dimensions of the arousal-valence space are not independent, e.g., high valence is generally associated with high arousal. However, we treated arousal and valence independently.

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 a similar ranking of 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 extracted 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 extra operation called relevant-feature selection has been implemented. During the relevant-feature selection for arousal, the Spearman correlation between the single-feature vectors and the self-assessed arousal vector is determined. Only the features with a correlation absolute ρ value 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 video contents in the experiment. They then had a brief training about the self-assessment procedure and concerning the meaning of arousal and valence. In emotional-affective experiments the bias of the emotional state or mood of participants creates problems for researchers. To avoid this problem and record a baseline at the start of each trial we showed one short (approximately 30s) neutral clip randomly selected from clips provided by the Stanford psychophysiology laboratory [29] (available online at http://www-psych.stanford.edu/~psyphy/resources.htm,).

Figure 2 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.

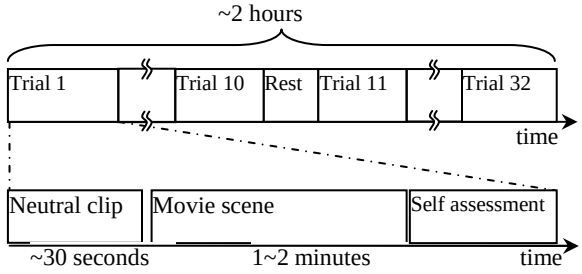

Figure 2. Experimental protocol.

3.3 Data

3.3.1 Movie scenes dataset

To create a video dataset for our research, we extracted video scenes from eight movies (mostly Hollywood movies). The majority of movies were selected either according to similar studies (e.g. [2][14][29][33]), 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).

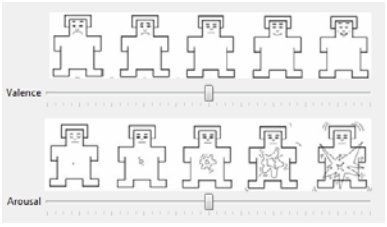

Figure 3. Arousal and valence self-assessment: SAM manikins and sliders.

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 was 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 4. 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.

3.4 Feature Extraction

3.4.1 Audio and video content-based features

3.4.1.1 Audio-based features

Sound has an important impact on user's affect. For example according to the findings of Picard [25], loudness of speech (energy) is related to evoked arousal, while rhythm and average pitch in speech signals 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. 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 [21][22].

Wang et al [33] 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, environment), we implemented a three class audio type classifier using support vector machines (SVM) operating on audio low-level features in a time window of one second. Despite the fact that in various cases the classes were overlapping (e.g. presence of a musical background during a dialogue), the classifier was usually able to recognize the dominant audio type. Before classification, silence could be identified by comparing the audio signal energy of each sound sample (using the averaged square magnitude in a time window) with a pre-defined threshold empirically set at $5 \cdot 10^{-7}$, while the audio signals amplitude range was normalized in the range [-1, 1] at the sampling rate of 48 KHz. After removing silence, the remaining audio signals were classified by the three classes SVM with a polynomial kernel, using the multiclass support vector machine from the OSU SVM toolbox. (http://sourceforge.net /projects/svm/). The SVM was trained utilizing more than 3 hours of audio, extracted from movies and labeled manually. The classification results were used to form a 4 bin (3 audio types and silence) normalized histogram; these histogram values were used as affective features for the affective ranking. MFCC (Mel frequency cepstral coefficients), LPCC (Linear prediction cepstral coefficients) and the pitch of audio signals were extracted using the PRAAT software package [4].

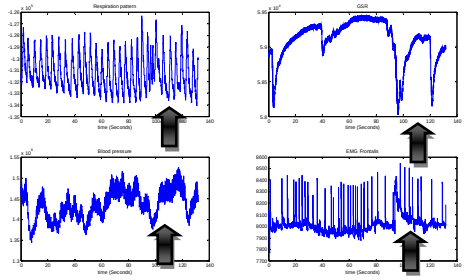

Figure 4. Physiological response (participant 2) to a surprising action scene. The following raw physiological signals are shown: respiration pattern (top-left), GSR (top-right), blood pressure (bottom left), and Frontalis EMG (bottom-right). The surprise moment is indicated by an arrow.

3.4.1.2 Video-based features

Movie scenes have been segmented at the shot level using the OMT shot segmentation software [16]. 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) [23] to capture video frames and extract motion vectors.

From a movie director's point of view, lighting key [25][30] and color variance [28] 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.

Hanjalic et al. [14] showed the relationship between video rhythm and affect. The average shot change rate, and shot length variance were extracted to characterize video rhythm. Fast movement of scene or objects in consecutive frames is also an effective factor for evoking excitement. To measure this factor, the motion component was defined as the amount of motion in consecutive frames. This parameter was computed by accumulating magnitudes of motion vectors for all B and P frames.

Table 1. Low-level features extracted from audio signals.

Feature category	Extracted features	Comments
MFCC - Mel frequency cepstral coefficients	MFCC coefficients	(13 features)
	Derivative of MFCC	(13 features)
	Autocorrelation of MFCC	[21] (13 features)
Energy	Average energy of audio signal	[21]
LPCC - Linear prediction cepstral coefficients	LPCC	[21] (16 features)
	Derivative of LPCC	[21] (16 features)
Time frequency	Spectrum flux	[21]
	Spectral centroid	[21]
	Delta spectrum magnitude	[21]
	Band energy ratio	[22]
Pitch	First pitch frequency	[21]
Zero crossing rate		[21]
Silence ratio	Proportion of silence in a time window	[9]

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 for the 20 bins were used for the 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 [33]. 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 [33]) are assumed to be dark and in shadow in the frame.

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 very small current through the hand. This resistance decreases due to an increase of sudation, which usually occurs when one is experiencing emotions such as stress or surprise. Moreover, Lang et al. discovered that the mean value of the GSR is related to the level of arousal [20]. (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 are variables that correlate with emotions, since stress can increase blood pressure [15]. Pleasantness of stimuli can increase peak heart rate response [20], and heart rate variability decreases with fear, sadness, and happiness [27]. Several of the features in Table 2 are therefore derived from the plethysmograph's recording.

Skin temperature changes in different emotional states [20]. The following features were therefore extracted from the skin temperature signal: minimum, maximum, average, and standard deviation of the temperature.

The respiration pattern is 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[17][27]. Laughing is known to affect the respiration pattern by introducing high-frequency fluctuations to the recorded signal. Features from both the frequency and time domain are therefore used. The energy ratio, between energies in a lower band (0.05 to 0.25 Hz) and a higher band (0.25 to 5Hz) was extracted from the respiration patterns. The spectral centroid was computed to represent the dominant rhythm of breathing.

Table 2. Features extracted 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

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 extracted, since this muscle is active when the user is laughing or smiling [12]. 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 feature was obtained from the energy of EMG signals in this frequency range.

The rate of eye blinking is another feature, which is correlated with anxiety. Eye-blinking affects the EMG signal that is recorded over the Frontalis muscle and results in easily detectable peaks in that signal. By counting these peaks in the Frontalis EMG, the eye blinking rate of a participant can also be determined.

3.5 Relevant features selection

The relevance of features was determined using Spearman ranking correlation between each extracted feature and the users' self-assessment, as motivated in Section 3.1. The ranking correlation coefficient, ρ, can be determined for any two ranked vectors of equal length n, by:

$$\rho = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2-1)}, -1 < \rho < 1, \quad (1)$$

where d_i indicates the absolute difference in position of the i^{th} largest value between the two vectors. For instance for $i = 1$, assuming the highest ranked element in the first vector is the third entry and the highest ranked element in the second vector is the twelfth entry, then d_1 is nine, the absolute value of (3-12). The maximum value of ρ, $\rho = 1$, occurs only when rankings of both vectors are identical. When $\rho = 0$ the two ranked vectors are not correlated, and when $\rho = -1$ the rankings are completely reversed in the two vectors.

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 randomly selected rankings would lead a ρ value that is at least as large as the one observed.. Each multimedia content- or physiological feature vector contains 64 values corresponding to the 64 movie scenes. Each feature's correlations with self assessed arousal/valence were computed and the features which were significantly correlated to arousal/valence self assessments were selected.

3.6 Regression and combination

It will now be presented how actual 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, for the user's arousal and valence, all significantly correlated features are weighted and summed' as is indicated in equation (2), where $\hat{y}(j)$ is the estimate of arousal/valence level, 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 is the total number of significant features for this participant, and w_i is the weight that corresponds to the i-th feature.

$$\hat{y}(j) = \sum_{i=1}^{N_s} w_i x_i(j) + w_0 \qquad (2)$$

In order not to have more features than movie scenes during the estimation of the weights, only the 41 highest correlated features were retained whenever the number of selected features exceeded 41.

In order to determine the optimum $\hat{y}$, the weights in equation (2) were computed by means of a linear relevance vector machine (RVM) from the Tipping RVM toolbox [32]. This procedure was applied on the user self assessed arousal/valence, $y(j)$, and 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 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.

After computation of weights by the train set in the first step, in the second step, the obtained weights were applied to the test set, and the Spearman correlation between the resulting estimated arousal/valence levels 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

Audio and video signals from the movie scenes contain valuable information about the emotion that we expect to see in the spectator. Content-based features were thus extracted from the audio and video signals of our dataset and those features that were significantly correlated with the self-assessment were retained in a first set. In a similar way a second set of significantly correlated features obtained from physiological responses was formed. As explained above, self-assessed arousal and valence levels were also recorded. To illustrate the fact that almost all different levels in the 2D valence/arousal space are reported.

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

Participant	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

Significantly correlated features (ρ> 0.25, p<0.05) have been selected from physiological responses and multimedia features for each participant. For each of the eight participants, Table 3 and Table 4 show the physiological and multimedia feature that had the highest correlation with that participant's self-assessments of perceived arousal and valence.

The large variations between participants regarding which multimedia features correlate most with their self assessments, indicate the variance in individual preferences to different audio or video features. For physiological signals, the variation of correlated features over different subjects illustrates the difference between participants' responses. While Average derivative of GSR signal was more informative regarding the valence of participant 6, EMG signals and thus facial expressions are more important to measure arousal in other of participants.

Table 4. Multimedia features with the highest absolute correlation with self assessments for participants 1 to 8.

Participant	Arousal	ρ	Valence	ρ
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

Table 5 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 5. The linear correlation ρ values between 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

The Spearman correlation between the self-assessed valence/arousal and the estimated valence/arousal, discussed in Section 3.6, is determined in each of the 1000 iterations over the test-set. The proportion of significantly correlated rankings (ρ>0.25, p<0.05) out of all 1000 iterations provides information on how consistent the effect of the correlated features on the arousal/valence is. This proportion is shown in Figure 5. It should be noted that a 100% result in this Figure does not imply that the rankings from features and self assessments were identical; it implies significant correlation between the two in all 1000 iterations. For example, in figure 6.a, participant 3 multimedia content features' estimated arousal levels ranking are correlated with self-assessed arousal ranking in 60.4 percent of the iterations (880 times over 1000 iterations). Results obtained for arousal rankings were inferior to those for valence ranking. This might be due to inaccurate self-assessment of arousal. Also, the multimedia content which is used as well as the experimental paradigm makes the evaluation and comparison harder for arousal grades. Results show more consistency between grades obtained from multimedia and physiological signals, and valence grades.

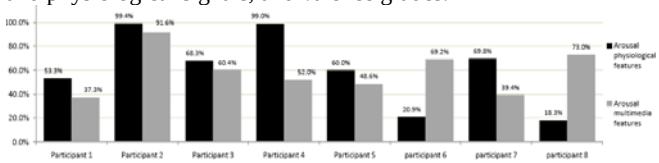

Figure 5.a. Probability of having significant Spearman correlation between estimated- and self assessed arousal.

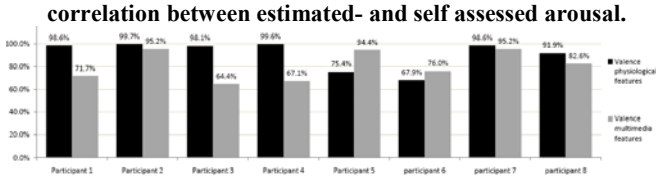

Figure 5.b. Probability of having significant Spearman correlation between estimated- and self assessed valence.

5. CONCLUSIONS

In this paper, an affective ranking method for movie scenes is proposed based on the emotions that are actually felt by spectators. Content based multimedia features were extracted from a movie scene dataset and their correlation with users' self-assessment of arousal and valence was shown to be significant. Furthermore, physiological responses of participants were recorded and key features were extracted from these responses. The correlation between these key physiological features and the users' self-assessment of arousal and valence was also verified. By computing correlations between these key physiological features and the user's self-assessment of arousal and valence, it was identified which physiological features are essential for affective ranking.

Promising results were thus obtained for affective ranking using both multimedia and physiological features. This ranking can facilitate video indexing and retrieval based on truly personal preferences; it can also contribute to understand emotional preferences of spectators. In addition, physiological responses can be used to predict what would be the self-assessment of valence and arousal levels. Currently we used the self assessments to serve as the ground truth but it is expected that in future physiological signals can be used for this with equal or superior reliability. Finally, the effects that specific multimedia features have on arousal and valence can also be predicted, and saved in a personal profile for personal affect profiling.

Participants exhibit markedly different emotional reactions to movie scenes. These differences can be explained by different factors, e.g., personalities, emotional bias or mood during experiments, or varying personal standards for reporting self-assessed true feelings. This shows the need for affect profiling to be, at least in part, user-dependant.

The exact nature of emotional processes is still under debate. We do not pretend in this work to explain affective mechanisms, but rather to employ the widely accepted valence and arousal measures as an additional feature for multimodal human-computer interaction in general, and for affective video indexing and retrieval as a case study. In the future we plan first to refine affect labeling of movie scenes by the means of classification algorithms. We aim at more precisely assessing which are the most important content-based characteristics able to elicit given emotions. Another important aspect to be investigated, through studies involving more users, concerns the determination of which emotional responses are common to all users and which are really user-dependent. The ranking correlation threshold of 0.25 for detecting relevant rankings was selected based on empirical results. Instead, in future studies the users' satisfaction feedbacks can be employed to optimize this threshold.

6. ACKNOWLEDGMENTS

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

7. REFERENCES

[1] K. Ansari-Asl, G. Chanel, and T. Pun, A channel selection method for EEG classification in emotion assessment based on synchronization likelihood, In *Eusipco 2007, 15th Eur. Signal Proc. Conf.*, Poznan, Poland, September 2007.

[2] 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, In *3rd IFIP Conference on Artificial Intelligence Applications & Innovations (AIAI),* Athens, Greece, June 2006.

[3] K. Boehner, R. DePaula, P. Dourish, and P. Sengers, "How emotion is made and measured," *International Journal. of Human-Computer Studies*, 65(4): 275-291, April 2007.

[4] P. Boersma and D. Weenink, Praat: doing phonetics by computer (Version 5.0.05) [Computer program]. Retrieved January 19, 2008, from http://www.praat.org/.

[5] A. Brave, C. Nass, and K. Hutchinson, Computers that care: investigating the effects of orientation of emotion exhibited by an embodied computer agent, *International Journal of Human Computer Studies*, 62(2):161-178, 2005.

[6] C.H. Chan and G.J.F. Jones, Affect based indexing and retrieval of films, In *Multimedia 2005*, Singapore, November 2005.

[7] G. Chanel G., K. Ansari-Asl, and T. Pun, Valence-arousal evaluation using physiological signals in an emotion recall paradigm, In *2007 IEEE SMC Int. Conf. on Systems, Man and Cybernetics, Smart cooperative systems and cybernetics: advancing knowledge and security for humanity*, Montreal, Canada, October 2007.

[8] G. Chanel, J. Kronegg, D. Grandjean, and T. Pun, Emotion assessment: Arousal evaluation using EEG's and peripheral physiological signals , *Proc. Int. Workshop Multimedia Content Representation, Classification and Security (MRCS), , B. Gunsel, A. K. Jain, A. M. Tekalp, B. Sankur, Eds., Lecture Notes in Computer Science, Vol. 4105, Springer,* pages 530-537, Istanbul, Turkey, September 2006.

[9] L. Chen, S. Gunduz, and M.T. Ozsu, Mixed type audio classification with support vector machine, In *Int. Conf. on Multimedia and Expo*, Toronto, Canada, July 2006.

[10] R.R Cornelius, Theoretical approaches to emotion, *Proc. Int. Speech Communication Association (ISCA) Workshop on Speech and Emotion*, pages 3-10, Belfast, Northern Ireland, September 2000.

[11] R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis, G. Votsis, S. Kollias, W. Fellenz, and J. G. Taylor, Emotion recognition in human computer interaction, *IEEE Signal Processing Magazine*, 18(1):32-80, 2001.

[12] G.B. Duchenne and R.A. Cuthbertson, *The mechanism of human facial expression*, Cambridge University Press, Cambridge, 1990.

[13] P. Ekman, et al., Universals and cultural differences in the judgments of facial expressions of emotion, *Journal of Personality and Social Psychology*, 53(4):712-717, 1987.

[14] A. Hanjalic and L.Q. Xu, Affective video content representation and modeling, *IEEE Trans. Multimedia*, 7(1):143–154, February 2005.

[15] J.A. Healey*, Wearable and Automotive Systems for Affect Recognition from Physiology*, PhD thesis, Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, May 2000.

[16] B. Janvier, É. Bruno, S. Marchand-Maillet, and T. Pun, Information-theoretic temporal segmentation of video and applications: multiscale keyframes selection and shot boundaries detection, *Multimedia Tools and Applications*, 30(3):273-288, 2006.

[17] F. H. Kanfer, "Verbal Rate, Eyeblink, and Content in Structured Psychiatric Interviews," *Journal of Abnormal and Social Psychology*, 61(3): 341-347, 1960.

[18] J. Kim, Emotion recognition from physiological measurement, In *Humaine European Network of Excellence Workshop*, Santorini, Greece, September 2004.

[19] J. Kronegg, G. Chanel, S. Voloshynovskiy, and T. Pun, EEG-based synchronized brain-computer interfaces: A model for optimizing the number of mental tasks, *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 15(1):50-58, March 2007.

[20] P.J. Lang, M. K. Greenwald, M. M. Bradley, and A.O. Hamm, Looking at pictures: affective, facial, visceral, and behavioral reactions, *Psychophysiology*, 30(3):261-273, 1993.

[21] D. Li, I.K. Sethi, N. Dimitrova, and T. McGee, Classification of general audio data for content-based retrieval, *Pattern Recognition Letters*, 22(5):533-544, 2001.

[22] L. Lu, H. Jiang, H.J. Zhang, , A robust audio classification and segmentation method, In *Multimedia 2001*, 2001.

[23] N. Moenne-Loccoz , OVAL: an object-based video access library to facilitate the development of content-based video retrieval systems. Technical report, Viper group - University of Geneva, 2004.

[24] J.D. Morris, SAM:the self-assessment manikin, an efficient cross-cultural measurement of emotional response*, Journal of Advertising Research,*35(6):63-68 ,1995.

[25] R.W. Picard, *Affective Computing*, MIT Press, Cambridge, MA, 1997.

[26] 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: April 2005.

[27] P. Rainville, A. Bechara, N. Naqvi, and A.R. Damasio, Basic emotions are associated with distinct patterns of cardiorespiratory activity, *International Journal of Psychophysiology*, 61(1): 5-18, 2006.

[28] Z. Rasheed, Y. Sheikh, and M. Shah, On the use of computable features for film classification, *IEEE Transactions on Circuit and Systems for Video Technology*, (11):52-64, 2005.

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

[30] J. Russell, A. Mehrabian, Evidence for a 3-factor theory of emotions, *Journal of Research in Personality*, 11(3):273-294, 1977.

[31] K. Takahashi, Remarks on Emotion Recognition from Bio-Potential Signals, In *2nd International Conference on Autonomous Robots and Agents,* Palmerston North, New Zealand, December 2004.

[32] M. Tipping, Sparse bayesian learning and the relevance vector machine, *Journal of Machine Learning Research*, 1(3):211–244, 2001.

[33] H.L. Wang, L.F. Cheong, Affective understanding in film, *IEEE Transactions on Circuits and Systems for Video Technology,* 16(6),:689 - 704, 2006.