International Journal of Semantic Computing
Vol. 3, No. 2 (2009) 235–254

World Scientific
www.worldscientific.com

AFFECTIVE CHARACTERIZATION OF MOVIE SCENES BASED ON CONTENT ANALYSIS AND PHYSIOLOGICAL CHANGES

MOHAMMAD SOLEYMANI*, GUILLAUME CHANEL†,
JOEP J. M. KIERKELS‡ and THIERRY PUN§

CVML Laboratory, Computer Science Department
University of Geneva, Route de Drize 7
Carouge (Geneva) CH–1227, Switzerland
**mohammad.soleymani@unige.ch*
†*guillaume.chanel@unige.ch*
‡*joep.kierkels@unige.ch*
§*thierry.pun@unige.ch*
http://cvml.unige.ch

In this paper, we propose an approach for affective characterization of movie scenes based on the emotions that are actually felt by spectators. Such a representation can be used to characterize the emotional content of video clips in application areas such as affective video indexing and retrieval, and neuromarketing studies. A dataset of 64 different scenes from eight movies was shown to eight participants. While watching these scenes, their physiological responses were recorded. The participants were 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 scene. 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 valence-arousal provided by the spectator's self-assessments, and affective grades obtained automatically from either physiological responses or from audio-video features. By means of an analysis of variance (ANOVA), the variation of different participants' self assessments and different gender groups self assessments for both valence and arousal were shown to be significant (p-values lower than 0.005). These affective characterization results demonstrate 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

1.1. *Overview*

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 determined from physiological responses and multimedia content analysis. The affect determination from multimedia content and physiological responses uses computing methods to bring the human preferences and intentions into a semantic multimedia management system.

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. Still rather few publications exist in the field of 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 audio and video features to classify basic emotions elicited by movie scenes. In [4], audio was classified into music, speech and environment signals and these were treated separately to shape an audio affective feature vector. The audio affective vector of each scene was fused with video-based features such as key lighting and visual excitement to form a scene feature vector. Finally, using the scene feature vectors, movie scenes were classified and labeled with emotions.

Hanjalic *et al.* [1] introduced "personalized content delivery" as a valuable tool in affective indexing and retrieval systems. In order to represent affect in video, they first selected video- and audio- content features based on their relation to the

valence-arousal space (for the definition of valence-arousal space, see Sec. 1.2) [1]. Then, arising emotions were estimated in this space by combining these features. While valence-arousal could be used separately for indexing, they combined these values by following their temporal pattern in the valence-arousal space. This allowed for determining an affect curve, shown to be useful for extracting video highlights in a movie or sports video. Although the detected highlights were shown to be correct, there was no ground truth for evaluating the system.

A hierarchical movie content analysis method based on arousal and valence related features was presented by M. Xu *et al.* [5]. In this hierarchical content analysis method the affect of each shot was first classified into the three arousal categories of calm, average and exciting. Using the arousal correlated features and fuzzy clustering, the audio short time energy and the first four MFCC - Mel frequency cepstral coefficients (as a representation of energy features), shot length, and the motion component of consecutive frames were used for arousal classification. Next, color energy, lighting and brightness were used as valence related features in a HMM-based valence classification of the shots which were previously classified in one of the three arousal categories. A drawback of the proposed approach is that a shot can however last less than few seconds; it is thus not realistic to form a ground-truth with assigning an emotion label to each shot.

A regression based arousal and valence representation of MTV (Music-TV) clips using content features was presented in [6]. The arousal and valence values were separated into 8 clusters by an affinity propagation method. Two different feature sets were used for arousal and valence determination which was evaluated using a ground truth. The ground truth was based on the average assessments of 11 users.

T. Wu *et al.* [7] presented an interactive content representation based on expressed emotion and physiological feedback. In this interactive content representation, peripheral physiological signals were used to recognize the preference of the user when the content (picture, text, and music video) were showed to the user. This was one of the first papers to demonstrate physiological feedback in a real-time interactive content representation system.

An affect representation suited to one particular person or a particular group of users cannot be generalized to an entire population as has been presented in [4, 5]. The affective responses vary from person to person and from culture to culture. Focusing on stimuli (video and audio content) and ignoring the participant's preferences and their variety does not lead to an efficient affect analysis. The personalized affective characterization which is introduced in [1] is a possible answer to this problem.

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 [9, 10]. 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 [9, 11–14].

This paper demonstrates a first step towards benefiting from actual physiological responses for creating affect-based tools. Although emotions can induce similar physiological responses amongst people (i.e. abrupt changes in the GSR in case of surprise or fear), the shape and the magnitude of these responses vary from person to person. This requires personalized emotional profiles to be determined, that can subsequently be used for affect based video indexing. Peripheral physiological signals were first recorded for monitoring the valence-arousal 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 valence-arousal 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 valence-arousal grades and those computed from physiological features was determined, as well as the correlation between these self-assessed valence-arousal grades 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 valence-arousal 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 1.2 presents some background on representation of affect and on the valence-arousal model to represent emotions. Section 2 elaborates on data acquisition, feature extraction and selection, and how features are combined for representation. The experimental results are given in Sec. 3 and finally conclusions and perspectives are presented in Sec. 4.

1.2. *Affect and affective representation*

Although the most straightforward way to represent an emotion is to use discrete labels such as fear, anxiety and joy, label-based representations suffer from several disadvantages. The main disadvantage is that labels are not cross-cultural: they do not have the same meaning in different cultures. The emotional labels can also be misinterpreted in a single culture (for instance the difference between joy and happiness is problematic). In addition, emotions are continuous phenomena rather than discrete ones and labels are unable to define the strength of an emotion. 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 (see Fig. 1). 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 [10, 15, 16]. Although such spaces do not provide any verbal description, a point in such space can be mapped to a categorical feeling label.

In a dimensional approach for affect representation, the affect of video scene can be represented by its coordinates in the valence-arousal space. Valence and arousal can be determined by self reporting. In order to record their felt emotions, participants were asked to grade each movie scene by valence-arousal grades using self-assessment Manikins (SAM) [17]. The arousal grade represented the level of arousal or excitement felt when watching the scene while the valence grade represents the felt pleasantness.

2. Material and Methods

2.1. *Overview*

A video dataset of 64 movie scenes was created (see Sec. 2.3) from which content-based low-level features were extracted. Experiments were conducted during which

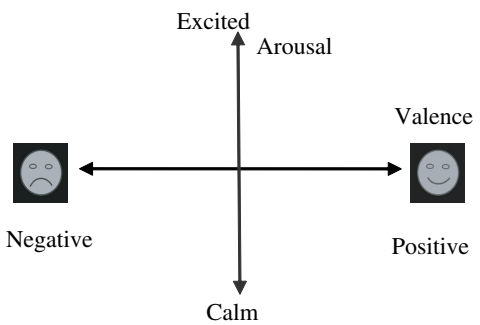


Fig. 1. Valence-arousal 2 dimensional space. The horizontal axis represents valence or degree of pleasantness and the vertical axis represents arousal or degree of excitement.

physiological signals were recorded from spectators. After each scene, the spectator self-assessed his/her valence-arousal 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 valence-arousal 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.

2.2. *Feature extraction*

2.2.1. *Audio and video content-based features*

Sound has an important impact on user's affect. For example according to the findings of Picard [18], 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

Table 1. Low-level features from audio signals.

Feature set	Extracted features
MFCC	MFCC coefficients, derivative and autocorrelation of MFCC, each 13 features [19]
Energy	Average energy of the audio signal [19]
Formants	Formants up to 5500 Hz (female voice), 5 features
Time frequency	Spectrum flux, spectral centroid, delta spectrum magnitude, band energy ratio, dominant pitch frequency [19, 20]
ZCR	Average and standard deviation of zero crossing frequency (ZCR) [19]
Silence ratio	Proportion of silence in a time window [22]

$[-1, 1]$. All of the resulting audio signals were normalized to the same amplitude range before further processing. A total of 53 low-level audio features were determined for each of the audio signals. These features, listed in Table 1, are commonly used in audio and speech processing and audio classification [19, 20].

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), Formants and the pitch of audio signals were extracted using the PRAAT software package [21].

Movie scenes have been segmented at the shot level using the OMT shot segmentation software [23]. 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) [24] to capture video frames and extract motion vectors.

From a movie director's point of view, lighting key [4, 25] and color variance [25] 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.

Visual excitement is a measure of the average pixel's color change between two consecutive frames [4]. It is defined as the average change between the CIE Luv histograms of the 20×20 blocks of two consecutive frames. In our case, this visual excitement feature was implemented from the definition given in [4] for each key frame. Two visual cues were also implemented to characterize these key frames. The first one, called visual detail, is used as an indicator of the distance from the camera to the scene and differentiates between close-ups and long-shots. The visual detail was computed by the average gray level co-occurrence matrix (GLCM) [4]. The other visual cue is the grayness which was computed from the proportion of the pixels with saturation below 20%, which is the threshold determined for colors that are perceived as gray [4].

2.2.2. *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 [26]. (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 [10]. Pleasantness of stimuli can increase peak heart rate response [26], and heart rate variability decreases with fear, sadness, and happiness [27]. The placement of electrodes on

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, Number of local minima in the GSR signal, average rising time of the GSR signal, kurtosis, skewness, spectral power in the bands ([0–0.1] Hz, [0.1–0.2] Hz, [0.3–0.4] Hz)
Blood flow and ECG (Plethysmograph)	Average blood pressure, heat rate, heart rate derivative, heart rate variability, standard deviation of heart rate, ECG multiscale entropy (4 levels), finger pulse transit time, kurtosis and skewness of the heart rate, energy ratio between the frequency bands [0, 0.08] Hz and [0.15, 5] Hz, spectral power in the bands ([0–0.1] Hz, [0.1–0.2] Hz, [0.3–0.4] Hz)
Respiration	Band energy ratio (energy ratio between the lower (0.05–0.25 Hz) and the higher (0.25–5 Hz) bands), average respiration signal, mean of derivative (variation of the respiration signal), standard deviation, dynamic range or greatest breath, breathing rhythm (spectral centroid), breathing rate, spectral power in the bands ([0–0.1] Hz, [0.1–0.2] Hz, [0.3–0.4] Hz)
EMG Zygomaticus	Energy, average, standard deviation of energy, variance
EMG Frontalis	Energy, average, standard deviation of energy, variance
Eye blinking rate	Rate of eye blinkings per second, extracted from the Frontalis EMG
Skin Temperature	Range, average, minimum, maximum, standard deviation, kurtosis, skewness, spectral power in the bands ([0–0.1] Hz, [0.1–0.2] Hz, [0.3–0.4] Hz)

the face (for EMG) and ground electrodes on the left hand enabled us to record an Electrocardiogram (ECG) signal. Using the electrocardiogram the pulse transit time was computed as a feature. In addition to the heart rate and heart rate variability features, the multi-scale entropy (MSE) of the heart rate variability was computed from ECG signals. The MSE of the heart rate was shown to be a useful feature in emotion assessment [28].

Skin temperature was also recorded since it changes in different emotional states [29]. 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 [27, 28]. 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 [30]. Most of the power in the spectrum of an EMG during muscle contraction is in the frequency range between 20 to 400 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 [31]. Eye-blinking affects the EMG signal that is recorded over the Frontalis muscle and results in easily detectable peaks in that signal.

2.3. *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 Sec. 2.1 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 63 scenes was formed by selecting 63 of the 64 movie scenes and the corresponding feature values. The remaining scene served as a test set.

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

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

In order to determine the optimum y, the weights in Eq. (1) were computed by means of a linear relevance vector machine (RVM) from the Tipping RVM toolbox [33].

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 absolute error between the resulting estimated valence-arousal grades and self assessed valence-arousal was examined. These two steps were repeated 64 times. Each time the 63 movie scenes of the training set were selected from the total of 64 scenes while the single remaining scene served as the test set. The results from this cross-validation will be presented in the next section.

3. Experimental Results

3.1. *Experimental protocol*

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 30 s. neutral clip randomly selected from clips provided by the Stanford psychophysiology laboratory [32].

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.

In the self-assessment step for evaluating arousal and valence, the SAM Manikin pictures with a slider to facilitate self-assessment of valence and arousal were used (see Fig. 3). The sliders correspond to a numerical range of [0, 1] while the numerical scale was not shown to the participants.

3.2. *Data*

3.2.1. *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, 32]), 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),

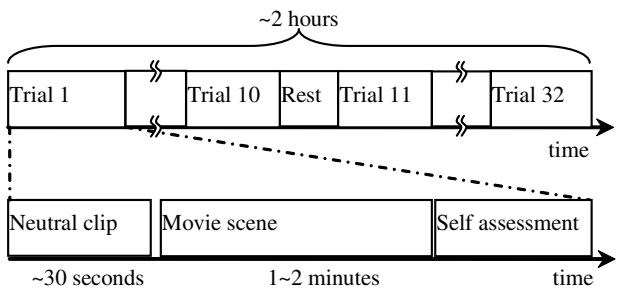


Fig. 2. Experimental protocol.

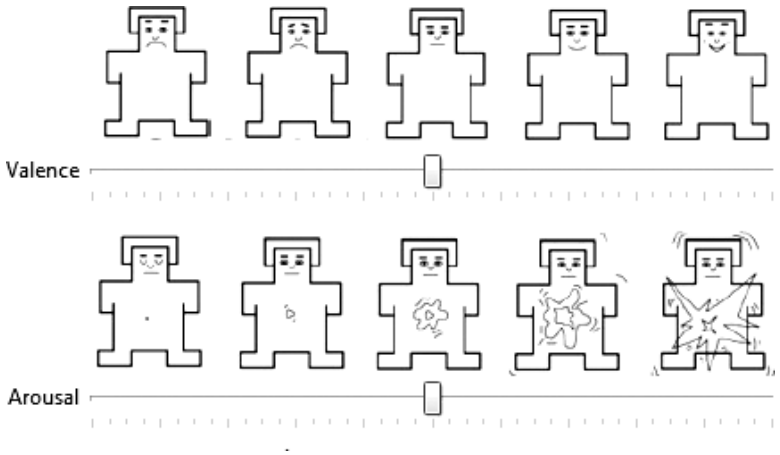


Fig. 3. Arousal and valence self-assessment: SAM manikins and sliders.

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). The complete list of the scenes with their timings and descriptions is available online (http://cvml.unige.ch/doku.php/mmi/movieaffectivecharacterization).

3.2.2. *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 Fig. 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 20 Hz and a higher cutoff frequency of 400 Hz.

3.3. *Results*

The correlations between multimedia features, physiological features and self assessments were determined. Table 3 shows, for each participant, the features which had

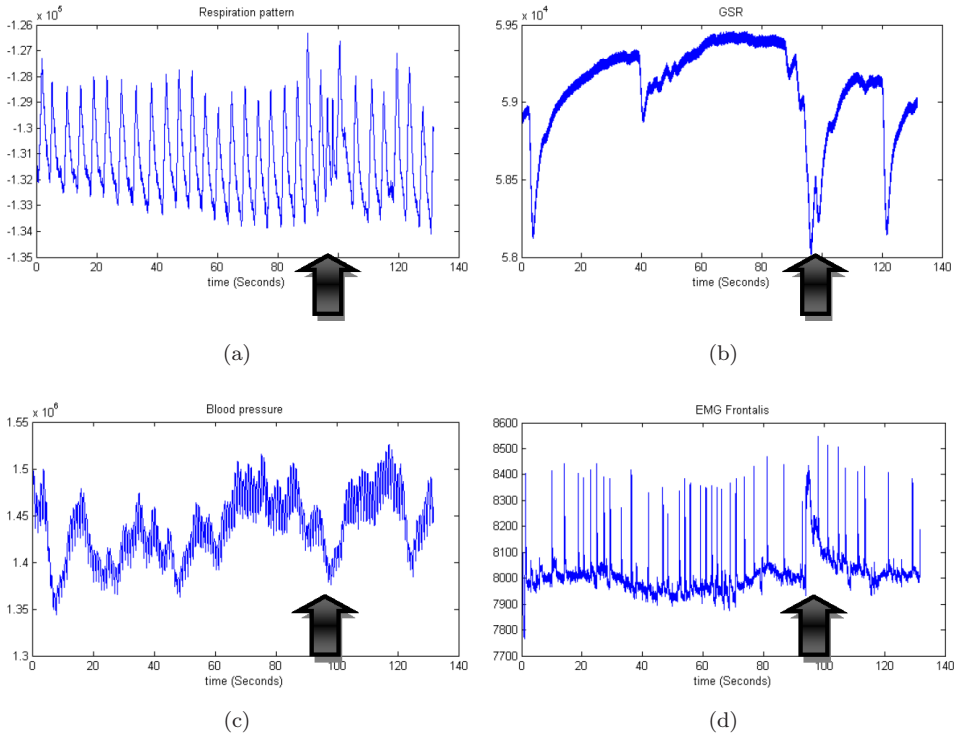


Fig. 4. Physiological response (participant 2) to a surprising action scene. The following raw physiological signals are shown: respiration pattern (a), GSR (b), blood pressure (c), and Frontalis EMG (d). The surprise moment is indicated by an arrow.

the highest absolute correlations with that participant's self-assessments of valence and arousal. Table 3a shows results for physiological features whereas Table 3b shows results for multimedia features.

For physiological signals, the variation of correlated features over different subjects illustrates the difference between participants' responses. While GSR features are more informative regarding the arousal level of participants 1, 3, and 7, EMG signals were more important to estimate arousal in participants 2, 4, 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 an increase in motion component leads to higher arousal for participant 8. For the same feature, increase in motion component resulted in lower valence 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

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

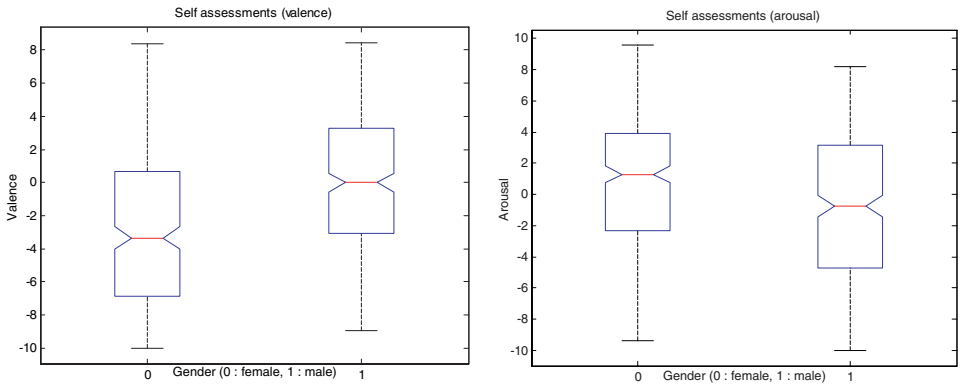
Participant	Arousal	ρ	Valence	ρ
	(a) *Physiological features*			
1	GSR Skewness	0.43	EMG Zygomaticus (sum of absolute)	0.74
2	EMG Frontalis (sum of absolute)	0.66	EMG Frontalis (sum of absolute)	−0.73
3	GSR power spectral density 0.1–0.2 Hz band	0.48	EMG Zygomaticus (sum of absolute)	0.53
4	EMG Zygomaticus average	0.32	EMG Zygomaticus (sum of absolute)	0.49
5	EMG Frontalis (sum of absolute)	0.38	EMG Frontalis (sum of absolute)	−0.49
6	Plethysmograph multi-scale entropy (2nd)	0.42	EMG Zygomaticus (sum of absolute)	0.56
7	GSR standard deviation	0.55	EMG Zygomaticus (sum of absolute)	0.71
8	Blood pressure (volume)	−0.33	EMG Zygomaticus (sum of absolute)	0.64
	(b) *Multimedia Features*			
1	6th MFCC coefficient	0.44	15th bin of the Hue histogram (purplish)	−0.47
2	19th bin of the Hue histogram (purplish)	−0.47	Shadow proportion standard deviation	−0.51
3	8th MFCC coefficient	0.45	Last autocorrelation MFCC coefficient (standard deviation)	0.53
4	First autocorrelation MFCC coefficient (standard deviation)	0.44	3rd autocorrelation MFCC coefficient (standard deviation)	0.39
5	4th Derivative MFCC	0.35	Motion component	−0.47
6	11th autocorrelation MFCC coefficient	−0.37	5th bin of Luminance histogram	−0.39
7	12th MFCC coefficient	0.43	Color variance standard deviation	0.48
8	Motion component	0.40	Visual cue, detail	0.52

assessments and thus are more importance for affect characterization. The correlations show that the indicated physiological responses are significantly correlated with changes in multimedia content. This is for instance the case with the positive correlation between EMG Zygomaticus energy and key lighting of the video content: lighter scenes have a direct positive effect on the Zygomaticus activity.

The difference between the self assessments of male and female participants was investigated by means of a one way ANOVA test of variance applied on these assessments. The difference between the two genders group self assessments was found to be significant for gender groups' valence ($F = 50.6$, $p < 0.005$) and arousal ($F = 11.9$, $p < 0.005$), and for participants' valence ($F = 10.3$, $p < 0.005$) and arousal ($F = 20.3$, $p < 0.005$). The female participants reported lower valence and higher arousal in average. Comparison with assessed valence showed that this gender

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

	EMG zygomaticus (sum of absolute values)/key lighting	GSR power spectral density 0–0.1 Hz band/standard deviation of the first autocorrelation of MFCC	Blood volume spectral density 0.1–0.2 Hz band/Shot length variation	EMG zygomaticus (sum of absolute values)/visual cue, details
1	—	0.73	0.54	—
2	0.71	0.72	0.84	0.53
3	0.35	0.71	0.89	0.33
4	0.51	0.50	0.78	0.43
5	0.39	0.63	0.88	0.36
6	0.41	0.63	0.91	0.30
7	0.46	0.76	0.86	0.40
8	0.64	0.55	0.82	0.56



(a) (b)

Fig. 5. Results of the one way ANOVA test on the self assessments showing significant differences between the average assessments levels of the two gender groups; (a) valence, p = 3.8E-12, F = 50.6; (b) arousal, p = 6E-4, F = 11.9.

difference comes from the fact that female participants had more intense unpleasant feelings about unpleasant scenes and reported more excitement from the exciting scenes. J. Rottenberg *et al.* [32] showed that female participants reported more intense emotions in response to emotional movie scenes. The female participant's emotional responses in our study were also stronger compared with those from male participants. Figure 5 shows the results of this one way ANOVA test on the two gender groups' valence self assessments.

The accuracy of the estimated valence and arousal is evaluated by computing the mean absolute error between the estimates and the self assessments of either valence or arousal (Table 5). The mean absolute error (E_{MAE}) was calculated from

Table 5. Mean absolute error (E_{MAE}), and Euclidean distance (E_{ED}) between estimated valence-arousal grades and self assessments (participants 1 to 8).

	E_{MAE} Arousal estimated from physiological features	E_{MAE} Arousal estimated from multimedia features	E_{MAE} Valence estimated from physiological features	E_{MAE} Valence estimated from multimedia features	E_{ED} physiological signals	E_{ED} multimedia features
1	0.17	0.18	0.10	0.13	0.21	0.24
2	0.12	0.17	0.09	0.15	0.16	0.23
3	0.15	0.15	0.11	0.04	0.21	0.21
4	0.15	0.15	0.12	0.13	0.20	0.21
5	0.15	0.15	0.14	0.15	0.22	0.24
6	0.18	0.15	0.18	0.12	0.27	0.22
7	0.16	0.12	0.11	0.12	0.21	0.18
8	0.16	0.13	0.07	0.07	0.18	0.16
Average	**0.15**	**0.15**	**0.11**	**0.11**	**0.21**	**0.21**
Random level			**∼0.4**			**∼0.5**

a leave-one-out cross validation on 64 video clips for each participant.

$$E_{MAE} = \frac{1}{N_{test}} \sum_{j=1}^{N_{test}=64} |\hat{y}_j - y_j| \tag{2}$$

E_{MAE} was computed from Eq. (2) where N_{test} is the number of test samples (64) and $\hat{y}_{ij}$ is the estimated valence-arousal in ith iteration for jth 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 determination. E_{MAE} values are shown in Table 5; all E_{MAE} values are considerably smaller than a random level determination of E_{MAE} (which is around 0.4, and is estimated by generating random measurements of valence and arousal).

While E_{MAE} separately considers valence and arousal determinations, a more global performance measure can be defined. Considering valence and arousal as coordinates in the 2-D valence-arousal space, the overall accuracy of the estimated, joint valence-arousal grades was evaluated by computing the Euclidean distance (E_{ED}) between the estimated points and the self assessments (ground truth). This Euclidean distance is a useful indicator of the system's performance for affect representation and affects similarity measurement, when using valence and arousal as indicators. With valence and arousal being expressed in normalized ranges [0–1], E_{ED} is computed as follows:

$$E_{ED} = \frac{1}{N_{test}} \sum_{j=1}^{N_{test}=64} \sqrt{(\hat{y}_j^{arousal} - y_j^{arousal})^2 + (\hat{y}_j^{valence} - y_j^{valence})^2} \tag{3}$$

E_{ED} values are shown in Table 5. It can in particular be observed that the average Euclidean distance results are all below random level (which is around 0.5).

The E_{MAE} represents the distance of the determined emotion from the self assessed emotion in the dimensions of arousal or valence. E_{MAE} is thus useful to compare each dimension's results. The E_{MAE} of arousal and valence shows that valence determination was more precise than arousal determination. The superior valence results might have been caused by the easier valence assessment and therefore a more precise self assessment on valence. The Euclidean distance, E_{ED}, represents the error caused by arousal and valence errors. E_{ED} between the estimated points and the self assessments was almost the same when estimated from either physiological signals or content analysis showing that none has any significant advantage over the other.

4. Conclusions and Perspectives

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 determination of valence-arousal. Such accurate determinations provide us with a continuous assessment of affect which can serve as a ground truth for affect determination. 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 valence-arousal 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 either from physiological signals or multimedia content. 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 valence-arousal values. It would however be possible to use the genre of movies (e.g., drama, comedy, etc.) or the temporal sequences of the emotional

events as prior knowledge for better affect determination. The temporal prior or the probability of changing the felt emotions from one emotion to another is a constraint that will be useful in emotion assessment. For example, the users' felt happiness cannot drastically changes to deep sadness or frustration.

Participants exhibit markedly different emotional reactions to movie scenes, as it was shown specifically for male/female groups. The difference between the self assessments of the participants and gender groups was verified by statistical tests. The differences between emotional responses can be explained by different factors, e.g., personalities, general mood during experiments, or varying personal standards for self-assessment of true feelings. Therefore, the personal user-dependant or groupwise affect profiles will help the emotion characterization methods. The groupwise profiles are useful to estimate the affect of any user by investigating his/her social background, gender, or age group without any other prior personal assessments.

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.

Acknowledgements

The authors gratefully acknowledge the support of the Swiss National Science Foundation. 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. The research leading to these results has received funding from the European Community's Seventh Framework Programme [FP7/2007-2011] under grant agreement Petamedia no. 216444.

References

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

[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, Oregon, April 2005.

[3] K. Boehner, R. DePaula, P. Dourish and P. Sengers, How emotion is made and measured, *Int. Jounal of Human-Computer Studies* **65**(4) (2007) 275–291.

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

[5] M. Xu, J. S. Jin, S. Luo and L. Duan, Hierarchical movie affective content analysis based on arousal and valence features, *Proceedings of the 16th ACM International Conference on Multimedia ACM*, Vancouver, Canada, October 2008, pp. 677–680.

[6] S. Zhang, Q. Tian, S. Jiang, Q. Huang and W. Gao, Affective MTV analysis based on arousal and valence features, *IEEE International Conference on Multimedia and Expo*, Hannover, Germany, 2008, pp. 1369–1372.

[7] T. L. Wu, H. K. Wang, C. C. Ho, Y. P. Lin, T. T. Hu, M. F. Weng, L. W Chan, C. H. Yang, Y. H. Yang, Y. P. Hung, Y. Y. Chuang, H. H. Chen, H. H. Chen, J. H. Chen and S. K. Jeng, Interactive content presentation based on expressed emotion and physiological feedback, *Proceedings of the 16th ACM International Conference on Multimedia ACM*, Vancouver, Canada, October 2008, pp. 1009–1010.

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

[9] 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 Conference on Artificial Intelligence Applications & Innovations*, Athens, Greece, 2006.

[10] J. A. Healey, Wearable and Automotive Systems for Affect Recognition from Physiology, MIT, May 2000.

[11] G. Chanel, J. Kronegg, D. Grandjean and T. Pun, Emotion assessment: Arousal evaluation using EEG's and peripheral physiological signals, LNCS 4105, September 2006, pp. 530–537.

[12] G. Chanel, K. Ansari-Asl and T. Pun, Valence-arousal evaluation using physiological signals in an emotion recall paradigm, *IEEE Conference on System, Man, and Cybernetics*, Montreal, October 2007.

[13] K. Ansari-Asl, G. Chanel and T. Pun, A channel selection method for EEG classification in emotion assessment based on synchronization likelihood, *Proceedings of the European Signal Processing Conference*, Poznan, Poland, Sept. 2007.

[14] K. Takahashi, Remarks on emotion recognition from bio-potential signals, *2nd Conference on Autonomous Robots and Agents*, New Zealand, Dec. 2004.

[15] J. A. Russel, M. Lewicka and T. Niit, A cross-cultural study of a circumplex model of affect, *Journal of Personality and Social Psychology* **57**(5) (1989).

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

[17] J. D. Morris, Observations: SAM: The self-assessment manikin — An efficient cross-cultural measurement of emotional response, *Journal of Advertising Research* **35**(6) (1995) 63–68.

[18] R. W. Picard, *Affective Computing* (MIT Press, 1997).

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

[20] L. Lu, H. Jiang and H. Zhang, A robust audio classification and segmentation method, *Proceeding of the Ninth ACM International Conference on Multimedia*, September 2001, pp. 203–211.

[21] P. Boersma and D. Weenink, Praat: Doing Phonetics by Computer, Computer Program, 2008.

[22] C. Lei, S. Gunduz and M. T. Ozsu, Mixed type audio classification with support vector machine, *IEEE Multimedia and Expo*, July 2006, pp. 781–784.

[23] B. Janvier, E. Bruno, T. Pun and S. Marchand-Maillet, Information-theoretic temporal segmentation of videos and applications: Multiscale keyframe selection and transition detection, *Multimedia Tools and Applications* **30** (2006) 273–288.

[24] 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, April 2004.

[25] Z. Rasheed, Y. Sheikh and M. Shah, On the use of computable features for film classification, *IEEE Transactions on Circuits and Systems for Video Technology* **15**(1) (2005) 52–64.
[26] 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) (1993) 261–273.
[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) (2006) 5–18.
[28] J. Kim and E. André, Emotion recognition based on physiological changes in music listening, *IEEE Transactions on Pattern Analysis and Machine Vision* **30**(12) (2008) 2067–2083.
[29] R. A. Mcfarland, Relationship of skin temperature-changes to the emotions accompanying music, *Biofeedback and Self-Regulation* **10**(3) (1985) 255–267.
[30] G.-B. Duchenne de Boulogne and R. Andrew Cuthbertson, *The Mechanism of Human Facial Expression* (Cambridge University Press, 1990).
[31] F. H. Kanfer, Verbal rate, eyeblink, and content in structured psychiatric interviews, *Journal of Abnormal and Social Psychology* **61**(3) (1960) 341–347.
[32] 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. (Oxford University Press, London, 2007).
[33] M. E. Tipping, Sparse Bayesian learning and the relevance vector machine, *Journal of Machine Learning Research* **1**(3) (2001) 211–244.