Continuous Emotion Detection in Response to Music Videos

Mohammad Soleymani, Sander Koelstra, Ioannis Patras and Thierry Pun

***Abstract*— Viewers' preference for multimedia selection depends highly on their emotional experience. In this paper, we present an emotion detection method for music videos using central and peripheral nervous system physiological signals as well as multimedia content analysis. A set of 40 music clips eliciting a broad range of emotions were first selected. After extracting the one minute long emotional highlight of each video, they were shown to 32 participants while their physiological responses were recorded. Participants self-reported their felt emotions after watching each clip by means of arousal, valence, dominance, and liking ratings. The physiological signals included electroencephalogram, galvanic skin response, respiration pattern, skin temperature, electromyograms and blood volume pulse using plethysmograph. Emotional features were extracted from the signals and the multimedia content. The emotional features were used to train a linear ridge regressor to detect emotions for each participant using a leave-one-out cross-validation strategy. The performance of the personalized emotion detection is shown to be significantly superior to a random regressor.**

I. Introduction

Emotional preference is one of the most important factors in multimedia content selection and consumption. Knowing a viewer's emotion while watching videos helps recommendation systems to better understand his/her preferences. A retrieval or recommendation system can use multimedia content analysis and a user's previous emotional feedback to estimate the emotion which is likely to be elicited in response to a new video. An alternative to explicitly receiving the emotional feedback can be using non-verbal behavior cues to detect emotions. Non-verbal cues are valued for not interrupting users for explicit feedback or self-reporting phases. Peripheral and central physiological responses are among the cues that have been used for emotion detection [11], [9], [13], [2]. Moreover, self-reporting emotions is not always an easy task for an ordinary viewer. The self-reporting becomes even more complicated in the case of using dimensional models of emotion.

Arousal, valence, dominance and liking ratings were used in this study for emotional representation of music videos. 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-lingual: they do not necessarily exist or have the same meaning in different languages. e.g. "disgust" does not have an exact equivalent in Polish [20]. The emotional labels can also be misinterpreted in a single culture. 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 3D valence-arousal-dominance space [21]. The valence scale ranges from unpleasant to pleasant. The arousal scale ranges from passive to active or excited. The dominance scale ranges from submissive (or "without control") to dominant (or "in control, empowered"). Fontaine et al. [5] added predictability to these three dimensions. In this study, predictability self-reporting in response to music videos was found too complex and hence not used.

There have been a large number of published works in the domain of emotion recognition from physiological signals [11], [9], [13], [2]. Of these studies, only a few achieved notable results using video stimuli. Lisetti and Nasoz used physiological responses to recognize emotions in response to movie scenes [13]. The movie scenes were selected to elicit six emotions, namely sadness, amusement, fear, anger, frustration and surprise. They achieved a high recognition rate of 84% for the recognition of these six emotions. However, the classification was based on the analysis of the signals in response to pre-selected segments in the shown video known to be related to highly emotional events.

Some efforts have been made towards implicit affective tagging of multimedia content. Kierkels et al. [8] proposed a method for personalized affective tagging of multimedia using peripheral physiological signals. Valence and arousal levels of participants' emotions when watching videos were computed from physiological responses using linear regression [24]. Quantized arousal and valence levels for a clip were then mapped to emotion labels. This mapping enabled the retrieval of video clips based on keyword queries. So far this novel method achieved low precision.

In a more recent study, Koelstra e al. [10] recorded EEG and peripheral physiological signals from six participants in response to 20 music videos. Participants reported their emotions using arousal, valence, and like/disliking rating. The responses of each participant was classified into two classes of low/high arousal, low/high liking rating, and low/high valence. The average classification rates varied between 55% and 58% which is slightly above random level. The low classification rates were caused by the the low number of the samples per participant. We hence increased the number of samples to 40 videos per participant in the current study.

Mohammad Soleymani and Thierry Pun are with the Computer Vision and Multimedia Laboratory, University of Geneva, Switzerland. mohammad.soleymani@unige.ch

Sander Koelstra and Ioannis Patras are with the School of Computer Science and Electronic Engineering, Queen Mary University of London. sander.koelstra@eecs.qmul.ac.uk

In this paper, we used music videos as stimuli and recorded users' emotional responses to a selected set of music videos. First, a relatively large music video dataset was gathered. For each video, a one-minute highlight was then extracted automatically for use in the experiment. The selection of videos was narrowed down to 40 clips to be shown during the experiments. 32 participants volunteered to participate in the experiment and their physiological signals (EEG and peripheral physiological signals) were recorded as they watched the 40 selected music videos. Participants rated each video in terms of arousal, valence, liking and dominance. For each video, arousal, valence dominance and liking ratings were estimated from peripheral and central physiological as well as multimedia content features using linear ridge regression and a leave-one-out cross validation strategy.

获取video的标签和emotion detection

The rest of the paper is organized as follows. The experiment and apparatus are explained in Section II. The methodology including features and regression method are presented in Section III. Experimental results are then given in Section IV. The paper is finally concluded in Section V.

II. Experimental protocol and apparatus

A. Music clips

120 music videos were initially selected with the goal of having videos with emotions that are uniformly covering the arousal-valence space. 60 of these were selected manually and 60 were selected using the last.fm website for music recommendation by searching on a list of emotional keywords. The music videos were then segmented into one minute segments with 55 seconds overlap between segments. Arousal and valence of the minute long segments were computed using the method proposed by Soleymani et al. [25] which is trained on movie scenes. In this method a linear regression was used to compute arousal for each shot in movies. Informative features for arousal estimation include loudness and energy of the audio signals, motion component, visual excitement and shot duration. The same approach was used to compute valence. Other content features such as color variance and key lighting that have been shown to be correlated with valence [29] were utilized for valence estimation. The emotional highlight score of the i-th segment e_i was computed using the following equation:

本文如何计算movie中的arousal valence

emotional highlight score的计算

$$e_i = \sqrt{a_i^2 + v_i^2} \quad (1)$$

The arousal, a_i, and valence, v_i, ranged between -10 and 10. Therefore, a smaller emotional highlight score (e_i) is closer to the neutral state. For each video, the one minute long segment with the highest emotional highlight score was chosen to be extracted for the experiment. For a few clips, the automatic affective highlight detection was manually overridden. This was done only for songs with segments that are particularly characteristic of the song, well-known to the public, and most likely to elicit emotional reactions. Given the 120 one-minute music video segments, the final selection of 40 videos used in the experiment was made on the basis of subjective ratings. Each video was rated by 14-16 volunteers using an online self-assesment tool. Valence and arousal was rated on a 9-point discrete scale. To maximize the strength of elicited emotions, we selected those videos that had the strongest volunteer ratings and at the same time a small variation. For each video we calculated a normalized arousal and valence score by taking the mean rating divided by the standard deviation. Then, for each quadrant in the normalized valence-arousal space, we selected the 10 videos that lie closest to the extreme corner of the quadrant.

video clips的选择

B. Apparatus

The experiments were performed in the laboratory environment with controlled illumination. EEG and peripheral physiological signals were recorded using a Biosemi ActiveTwo system[1]. Stimuli were presented using a dedicated stimulus PC that sent synchronization markers directly to the recording PC. For presentation of the stimuli and recording the users' ratings, the "Presentation" software by Neurobehavioral systems[2] was used.

Physiological signals were recorded at a sampling rate of 512 Hz using 32 active AgCl electrodes (placed according to the international 10-20 system). Thirteen channels of peripheral physiological signals were also recorded. 32 Healthy participants (50% female), aged between 19 and 37 (mean age 26.9), participated in the experiment. Prior to the experiment, each participant signed a consent form and filled out a questionnaire. Next, they were given a set of instructions to read informing them of the experiment protocol and the meaning of the different scales used for self-assessment. An experimenter was also present there to answer any questions. When the instructions were clear to the participant, he/she was led into the experiment room. The experiment started with a 2 minute rest period. Then the 40 videos were presented in 40 trials, each consisting of the following steps. First, the current trial number was shown for two seconds to inform the participants of their progress. then, a five seconds baseline recording (display of a fixation cross) was shown. This was followed by a one minute long music video.

Fig. 1. A participant shortly before the experiment.

At the end of each trial, participants self-assessed their emotions reporting the level of arousal, valence, dominance

[1]http://www.biosemi.com
[2]http://www.neurobs.com

and liking rating. The liking scale asks for participants' personal appreciation of the video (that is, how much they liked it). It should not be confused with the valence scale. This measure inquires about the participants' tastes, not their feelings. For example, it is possible to like videos that make one feel sad or angry. Self-assessment manikins (SAM) [16] were used to visualize the scales. For the liking scale, thumbs down/thumbs up symbols were used. The manikins were displayed in the center of the screen with the numbers 1-9 printed below. Participants moved the mouse horizontally and clicked to indicate their self-assessment level. Fig. 1 shows a participant shortly before the start of the experiment.

试验者的 self-assessment

III. Methodology

A. Physiological features

Most of the current theories of emotion [3], [22] agree that physiological activity is an important component of emotional experience. For instance several studies have demonstrated the existence of specific physiological patterns associated with basic emotions [4].

The following peripheral nervous system signals were recorded: GSR, respiration amplitude, skin temperature, electrocardiogram, blood volume pulse (BVP) by plethysmograph, electromyograms of Zygomaticus and Trapezius muscles, and electrooculogram (EOG). GSR provides a measure of the resistance of the skin by positioning two electrodes on the distal phalanges of the middle and index fingers. This resistance decreases due to an increase of perspiration, 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 [11].

生理信号有哪些

GSR 通过在中指和食指的远端指骨上放置两个电极来提供对皮肤电阻的测量

A plethysmograph measures blood volume in the participant's thumb. This measurement can also be used to compute the heart rate (HR) by identification of local maxima (i.e. heart beats), inter-beat periods, and heart rate variability (HRV). Blood pressure and heart rate variability correlate with emotions, since stress can increase blood pressure. Pleasantness of stimuli can increase peak heart rate response [11]. In addition to the HR and HRV features, spectral features derived from HRV were shown to be a useful feature in emotion assessment [15]. Electrocardiogram activities were also detectable from the electrode recording EMG signal from Trapezius muscle. The electrocardiogram (ECG) was extracted and HR and HR related features were extracted. The ECG features are in part overlapping with plethysmograph features.

Skin temperature was also recorded since it varies with different emotions. The respiration amplitude was measured by a respiration belt around the abdomen of the participant. Slow respiration is linked to relaxation while irregular rhythm, quick variations, and cessation of respiration correspond to higher arousal levels.

Regarding the EMG signals, the Trapezius muscle (neck) activity was recorded to investigate possible head movements during music listening. The activity of the Zygomaticus major was also monitored, since this muscle is activated when the participant laughs or smiles. Most of the power in the spectrum of an EMG during muscle contraction is in the frequency range between 4 to 40 Hz. Thus, the muscle activity features were obtained from the energy of EMG signals in this frequency range for the different muscles. The rate of eye blinking is another feature, which is correlated with anxiety. Eye-blinking affects the EOG signal and results in easily detectable peaks in that signal.

TABLE I
THE FEATURES EXTRACTED FROM EEG AND PHYSIOLOGICAL SIGNALS.

Physiological signal	Extracted features
GSR	average skin resistance, average of derivative, average 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, 10 spectral power in the bands from 0 to 2.4Hz, zero crossing rate of Skin conductance slow response (SCSR) (0 to 0.2 Hz), zero crossing rate of Skin conductance very slow response (SCVSR) (0 to 0.08 Hz), SCSR and SCVSR mean of peaks magnitude
BVP	average blood volume pulse, average HR, average and standard deviation of inter beat intervals (HRV), energy ratio between the frequency bands [0.04, 0.15]Hz and [0.15, 0.5]Hz, spectral power in the bands ([0.1-0.2]Hz, [0.2-0.3]Hz, [0.3-0.4]Hz), low frequency [0.01,0.08]Hz, medium frequency [0.08,0.15]Hz and high frequency [0.15,0.5]Hz components of HRV power spectrum.
ECG	average and standard deviation of HR and its derivative, HRV, average of inter beat intervals, energy ratio between the frequency bands [0.04, 0.15]Hz and [0.15, 0.5]Hz, low frequency [0.01,0.08]Hz, medium frequency [0.08,0.15]Hz and high frequency [0.15,0.5]Hz components of HRV power spectrum, Poincaré analysis features [9], average beat interval change between per respiratory cycle [18]
Respiration	band energy ratio (difference between the logarithm of energy 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, range or greatest breath, breathing rhythm (spectral centroid), breathing rate, 10 spectral power in the bands from 0 to 2.4Hz, average peak to peak time, median peak to peak time
Skin temp.	average, average of its derivative, spectral power in the bands ([0-0.1]Hz, [0.1-0.2]Hz)
EMG and EOG	eye blinking rate, energy of the signal, mean and variance of the signal
EEG	theta, slow alpha, alpha, beta, and gamma Spectral power for each electrode. The spectral power asymmetry between 14 pairs of electrodes in these frequency bands.

All the physiological responses were recorded at a 512 Hz sampling rate and later down-sampled to 256 Hz to reduce the memory and processing costs. The trend of the ECG and GSR signals was removed by subtracting the temporal low frequency drift. The low frequency drift was computed by smoothing the signals on each ECG and GSR channels with a 256 points moving average.

In total 177 features were extracted from peripheral physiological responses based on the proposed features in the literature [2], [9], [18], [30]. Details are given in Table I.

B. EEG features

EEG Signals were acquired from 32 electrodes placed on the subjects' scalps according to the international 10-20 system. Signals were recorded at 512Hz and then downsampled to 128Hz to simplify further processing. EOG was recorded from 4 electrodes placed to the sides and above/below the eyes, and eye artefacts were suppressed using the algorithm proposed in [23]. Next, a band pass filter of 4-45Hz was used to further reduce signal artefacts and remove the 50Hz power line interference. Finally, the data was referenced to the common average.

眼动的抑制

Power spectral density (PSD) in different frequency bands was estimated using Welch's method for each 60 second trial. The frequency bands were: theta (4-8Hz), slow alpha (8-10Hz), alpha (8-12Hz), beta (12-30Hz) and gamma (30-45Hz). We also computed the lateralization for 14 left-right pairs of electrodes by computing the difference in PSD for each frequency band. The EEG feature vector is composed of 230 features ($5 \times$ (32 channels $+$ 14 asymmetry pairs)).

C. MCA features

Music videos were encoded into the MPEG-1 format to extract motion vectors and I-frames for further feature extraction. The video stream of the music clips has been segmented at the shot level using the method proposed in [7]. From a movie director's point of view, lighting key [19], [29] and color variance [29] are important tools to evoke emotions. We therefore extracted lighting key from frames in the HSV space by multiplying the average V-value (in HSV) by the standard deviation of the V-values. Color variance was obtained in the CIE LUV color space by computing the determinant of the covariance matrix of L, U, and V.

因为是music video，所以有rhythm

Hanjalic and Xu [6] showed the relationship between video rhythm and affect. The average shot change rate, and shot length variance were extracted to characterize video rhythm. Fast movements in scenes are also an effective factor for evoking excitement. To measure this factor, the motion component was computed by accumulating magnitudes of motion vectors for all B- and P-frames.

颜色激发的情绪

Colors and their proportions are important parameters to elicit emotions [28]. In order to extract color characteristics, a 20 bin color histogram of hue and lightness values in the HSV space was computed for each I-frame. The median of the L value in HSL space was computed to obtain the median lightness of a frame. Finally, visual cues representing shadow proportion, visual excitement, grayness and details were also determined according to the definition given in [29].

声音对情绪的影响

The second information stream, namely sound, also has an important impact on affect. For example, loudness of speech (energy) is related to evoked arousal, while rhythm and average pitch in speech signals are related to valence [17]. The audio channels of the videos were extracted and encoded into monophonic information (MPEG layer 3 format) at a sampling rate of 44.1 kHz. 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 are commonly used in audio and speech processing and audio classification [12], [14]. MFCC, formants and the pitch of audio signals were extracted using the PRAAT software package [1]. Visual features were extracted from key frames and their average, standard deviation, skewness, and kurtosis over each clip was computed as that clip's visual content features. Audio features were extracted from non-overlapping 200ms long segments and their average and standard deviation over each clip was computed as that clip's audio content features. An extensive list of the utilized content features can be found in [25].

D. Regression

After feature extraction, the results for each of the modalities was obtained using the same methodology. The goal here is to train a regression function to map the features to the rating given by the subjects in the experiments. We used a basic linear ridge regression algorithm with $\alpha = 10$ as implemented in the mlpy package[3]. Other, more complicated methods such as Guassian Process Regression [31] and Relevance Vector Machine regression [26] were also tried. Those regression methods were not found to improve the results, possibly due to the noisy nature of the physiological signals, leading to overfitting.

本文使用的回归函数

For each subject, a regressor is trained and tested using leave-one-trial-out cross validation strategy, for which the results are reported below. For comparison, we also present results obtained by a random regressor and a Π-regressor. The Π-regressor uses an estimate of the probability density function of the ground truth ratings to generate its output. This Π-regressor is included for comparison as the ground truth is not uniformly distributed (see Fig. 2). These random and Π-regressors were each run 10000 times and their error were averaged.

IV. Results

A. Rating analysis

After watching each video, participants reported their emotion by means of continuous ratings ranging from 1 to 9. Although they were able to choose any point on continuous scale participants tended to click under displayed numbers (see the red bars on Fig. 2). The blue bars on Fig. 2 show the ratings' histograms quantized in nine levels. From the blue bars, we can see that the distribution of the ratings are skewed towards higher scores.

The average ratings of the videos are shown in Fig. 3. According to the average ratings, the videos are well covering the whole arousal and valence plane on four quadrants, namely, low arousal high valence (LAHV), low arousal low valence (LALV), high arousal low valence (HALV) and high arousal high valence (HAHV). The orientation of the triangles represents the emotional quadrant which was expected to be felt given the ratings submitted by volunteers using the online self-assessment tool. The results show that the expected emotions are in strong agreement with reported

[3] https://mlpy.fbk.eu/

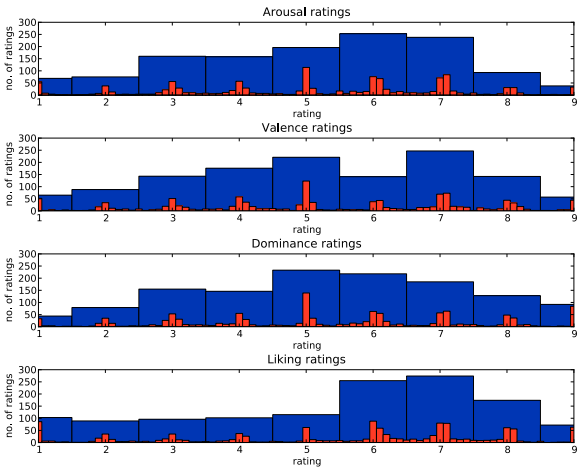

Fig. 2. Histogram of arousal, valence, dominance and liking ratings given to all videos by the 32 participants. The blue bars are the histogram of the quantized ratings in nine levels.The red bars are showing the ratings quantized in 80 levels (The quantization step is equal to 0.1).

emotions (i.e. online volunteers usually place the video in the same quadrant as the participants in the experiment). The average ratings for dominance ratings are also visible in Fig. 3. The liking ratings which are encoded in colors are visually shown to be correlated with valence.

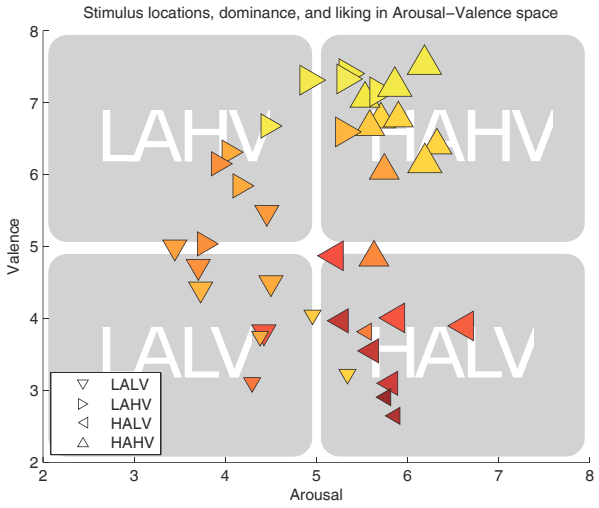

Fig. 3. The mean video ratings from experiment participants on the arousal-valence plane . The orientation of the triangle indicates the quadrant for which the video was selected by the online ratings. Liking is encoded by color: dark red is low liking and bright yellow is high liking. Dominance is encoded by symbol size: small symbols stand for low dominance and big for high dominance.

In order to measure inter-annotation agreement between different participants, we computed the pair-wise Cohen's kappa between self reports after quantizing the ratings into nine levels. A very weak agreement was found on emotional feedbacks with $mean(\kappa) = 0.02 \pm 0.06$ for arousal, $mean(\kappa) = 0.08 \pm 0.08$ for valence, and $mean(\kappa) = 0.05 \pm 0.08$ for liking ratings. A paired t-test was performed on the κ values of valence ratings in comparison to liking and arousal. The t-test results showed that on average the agreement on valence ratings is significantly higher than agreement on arousal ($p = 2.0 \times 10^{-20}$) and liking rating ($p = 4.5 \times 10^{-7}$).

B. Emotion detection results

TABLE II

mae (MEAN ABSOLUTE ERROR) AND ITS STANDARD DEVIATION OVER PARTICIPANTS. mae IS THE MEAN DIFFERENCE BETWEEN THE TRUE AND PREDICTED RATING (RATINGS ON A SCALE OF 1-9). STARS INDICATE SIGNIFICANCE OF RESULT COMPARED TO THE Π-REGRESSOR (**= $p < .01$), (*= $p < .05$). FOR COMPARISON, RESULTS FROM THE RANDOM AND Π REGRESSOR ARE ALSO PRESENTED.

	Arousal	**Valence**	**Dominance**	**Liking**
EEG	1.53(0.40)**	**1.59(0.39)****	1.53(0.49)**	1.78(0.51)**
Peripheral	1.70(0.51)*	1.81(0.41)	1.64(0.49)	1.96(0.64)
MCA	**1.50(0.45)****	1.65(0.35)**	**1.47(0.46)****	**1.68(0.45)****
EEG/Per/MCA	1.49(0.42)**	1.56(0.36)**	1.51(0.49)**	1.66(0.46)**
EEG/MCA	**1.47(0.42)****	**1.55(0.39)****	**1.46(0.48)****	**1.62(0.45)****
EEG/Per	1.58(0.43)**	1.63(0.39)**	1.57(0.50)**	1.83(0.53)*
Per/MCA	1.53(0.52)**	1.66(0.38)**	1.50(0.46)**	1.72(0.52)**
Random regr.	2.51(0.05)	2.58(0.05)	2.57(0.05)	2.69(0.05)
Π-regressor	2.05(0.04)	2.30(0.05)	1.99(0.04)	2.33(0.05)

Table II shows the regression mae (Mean absolute error) for the different modalities and rating scales, with the ratings on a continuous scale between 1 and 9. The regression results for each modality and rating scale are better than the random and Π-regressors. We performed a two-sided repeated samples t-test on the mae-scores per subject between each modality and the Π-regressor. The mae-scores are always significantly higher than Π-regression for the EEG and MCA modalities. For the peripheral physiological signals, the results are only significant for the arousal modality.

This indicates information regarding a subject's emotional state exists in the physiological measurements. It also seems evidence to predict the user's emotional reaction to a video is available in the content features extracted from the videos. The differences between modalities are less clear, but overall regression from MCA performs best, followed by EEG.

We have also performed feature-level fusion of the different modalities. Although for valence and liking, better results are attained, they are not significantly better.

Emotion estimation from MCA features is advantageous since there will not be any need to to attach electrodes and sensors to the users. Emotion detection from physiological modalities are valued not to interrupt users for self-reports. However, if the accuracy of emotion detection using physiological signals is not going to be superior to MCA or high enough to replace the self reports their usage will be under question. Therefore, the current emotion detection method from peripheral and central nervous physiological signals needs improvement to be used in a real application.

V. CONCLUSION AND PERSPECTIVES

A video data set consisting of music videos spanning the whole spectrum of emotions was collected. The physiolog-

ical responses of 32 participants were then recorded while watching the collected emotional music videos. Participants continuously rated video clips by means of arousal, valence, dominance, and liking ratings. The inter annotation agreement measures show that there was slightly more agreement on valence ratings in comparison to arousal and like/dislike ratings. Features that have correlation with emotions were then extracted from both central and peripheral physiological signals as well as multimedia content. A regression method for continuous emotional characterization of music videos using central and peripheral physiological responses as well as content analysis was applied on the feature set. The performance of the emotion detection was evaluated in a leave-one-out cross-validation strategy for each participant. The continuous emotion detection performance was shown to be significantly superior to random estimation. This method can be used in the heart of music video recommendation systems to improve viewers' experience.

Tkalčič et al. [27] showed how using valence, arousal, and dominance scores of images improved the performance of their image recommender. In our future work, we will take a similar approach for music videos to study the effect of using emotional scores in a music video recommendation system. Another topic of interest is to investigate whether it is possible to train a generalized, inter-subject regressor.

VI. ACKNOWLEDGMENTS

The research leading to these results has been performed in the frameworks of European Community's Seventh Framework Program (FP7/2007-2011) under grant agreement no. 216444 (PetaMedia). Furthermore, the authors gratefully acknowledge the support of the Swiss National Foundation. The collection of the dataset in this paper was not possible without the help of Christian Mühl, University of Twente, Ashkan Yazdani, and Jong-Seok Lee, École Polytechnique Fédérale de Lausanne (EPFL). The authors would like to acknowledge their contributions.

REFERENCES

[1] P. Boersma. Praat, a system for doing phonetics by computer. *Glot International*, 5(9/10):341–345, 2001.

[2] G. Chanel, J. J. M. Kierkels, M. Soleymani, and T. Pun. Short-term emotion assessment in a recall paradigm. *International Journal of Human-Computer Studies*, 67(8):607–627, Aug. 2009.

[3] R. R. Cornelius. *The Science of Emotion. Research and Tradition in the Psychology of Emotion*. Prentice-Hall, Upper Saddle River, NJ, 1996.

[4] P. Ekman, W. V. Friesen, M. O'Sullivan, A. Chan, I. Diacoyanni-Tarlatzis, K. Heider, R. Krause, W. A. LeCompte, T. Pitcairn, and P. E. Ricci-Bitti. Universals and cultural differences in the judgments of facial expressions of emotion. *Journal of Personality and Social Psychology*, 53(4):712–717, Oct. 1987.

[5] J. R. J. Fontaine, K. R. Scherer, E. B. Roesch, and P. C. Ellsworth. The World of Emotions is not Two-Dimensional. *Psychological Science*, 18(12):1050–1057, 2007.

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

[7] P. Kelm, S. Schmiedeke, and T. Sikora. Feature-based video key frame extraction for low quality video sequences. In *Proc. Int. Workshop on Image Analysis for Multimedia Interactive Services*, pages 25 –28, May 2009.

[8] J. Kierkels, M. Soleymani, and T. Pun. Queries and tags in affect-based multimedia retrieval. In *Proc. Int. Conf. Multimedia and Expo, Special Session on Implicit Tagging*, pages 1436 – 1439, New York, USA, Jun. 2009.

[9] J. Kim and E. André. Emotion recognition based on physiological changes in music listening. *IEEE Trans. Pattern Anal. Mach. Intell.*, 30(12):2067–2083, 2008.

[10] S. Koelstra, A. Yazdani, M. Soleymani, C. Mühl, J.-S. Lee, A. Nijholt, T. Pun, T. Ebrahimi, and I. Patras. Single trial classification of EEG and peripheral physiological signals for recognition of emotions induced by music videos. In Y. Yao et al., editor, *Brain Informatics*, volume 6334 of *Lecture Notes in Computer Science*, chapter 9, pages 89–100. Springer, Berlin, Heidelberg, 2010.

[11] P. Lang, M. Greenwald, M. Bradely, and A. Hamm. Looking at pictures - affective, facial, visceral, and behavioral reactions. *Psychophysiology*, 30(3):261–273, May 1993.

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

[13] C. L. Lisetti and F. Nasoz. Using noninvasive wearable computers to recognize human emotions from physiological signals. *EURASIP J. Appl. Signal Process.*, 2004(1):1672–1687, Jan. 2004.

[14] L. Lu, H. Jiang, and H. Zhang. A robust audio classification and segmentation method. In *Proc. ACM Int. Conf. Multimedia*, pages 203–211, Ottawa, Canada, 2001.

[15] R. McCraty, M. Atkinson, W. Tiller, G. Rein, and A. Watkins. The effects of emotions on short-term power spectrum analysis of heart rate variability. *The American Journal of Cardiology*, 76(14):1089 – 1093, 1995.

[16] J. D. Morris. SAM: the self-assessment manikin. An efficient cross-cultural measurement of emotional response. *Journal of Advertising Research*, 35(8):63–68, 1995.

[17] R. W. Picard. *Affective Computing*. MIT Press, Sep. 1997.

[18] 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, Jul. 2006.

[19] Z. Rasheed, Y. Sheikh, and M. Shah. On the use of computable features for film classification. *IEEE Trans. Circuits Syst. Video Technol.*, 15(1):52–64, 2005.

[20] J. A. Russell. Culture and the Categorization of Emotions. *Psychological Bulletin*, 110(3):426–450, 1991.

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

[22] D. Sander, D. Grandjean, and K. R. Scherer. A systems approach to appraisal mechanisms in emotion. *Neural Networks*, 18(4):317–352, 2005.

[23] A. Schlögl, C. Keinrath, D. Zimmermann, R. Scherer, R. Leeb, and G. Pfurtscheller. A fully automated correction method of EOG artifacts in EEG recordings. *Clinical Neurophysiology*, 118(1):98–104, 2007.

[24] M. Soleymani, G. Chanel, J. J. M. Kierkels, and T. Pun. Affective characterization of movie scenes based on content analysis and physiological changes. *International Journal of Semantic Computing*, 3(2):235–254, Jun. 2009.

[25] M. Soleymani, J. J. M. Kierkels, G. Chanel, and T. Pun. A Bayesian framework for video affective representation. In *Proc. Int. Conf. Affective Computing and Intelligent interaction*, pages 1–7, Amsterdam, Netherlands, Sep. 2009.

[26] M. E. Tipping. Sparse bayesian learning and the relevance vector machine. *Journal of Machine Learning Research*, 1:211–244, Jun. 2001.

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

[28] P. Valdez and A. Mehrabian. Effects of color on emotions. *J. Exp. Psychol. Gen.*, 123(4):394–409, Dec. 1994.

[29] H. L. Wang and L.-F. Cheong. Affective understanding in film. *IEEE Trans. Circuits Syst. Video Technol.*, 16(6):689 – 704, Jun. 2006.

[30] J. Wang and Y. Gong. Recognition of multiple drivers' emotional state. In *Proc. Int. Conf. Pattern Recognition*, pages 1 –4, 2008.

[31] C. Williams and C. Rasmussen. *Gaussian Processes for Machine Learning*. MIT Press, 2007.