Affective Content Analysis of Music Video Clips

Ashkan Yazdani
Multimedia Signal Processing Group (MMSPG)
Ecole Polytechnique Fédérale de Lausanne (EPFL)
1015 Lausanne, Switzerland
ashkan.yazdani@epfl.ch

Krista Kappeler
Multimedia Signal Processing Group (MMSPG)
Ecole Polytechnique Fédérale de Lausanne (EPFL)
1015 Lausanne, Switzerland
krista.kappeler@epfl.ch

Touradj Ebrahimi
Multimedia Signal Processing Group (MMSPG)
Ecole Polytechnique Fédérale de Lausanne (EPFL)
1015 Lausanne, Switzerland
touradj.ebrahimi@epfl.ch

ABSTRACT

Nowadays, the amount of multimedia contents is explosively increasing and it is often a challenging problem to find a content that will be appealing or matches users' current mood or affective state. In order to achieve this goal, an efficient indexing technique should be developed to annotate multimedia contents such that these annotations can be used in a retrieval process using an appropriate query. One approach to such indexing techniques is to determine the affect (type and intensity), which can be induced in a user while consuming multimedia. In this paper, affective content analysis of music video clips is performed to determine the emotion they can induce in people. To this end, a subjective test was developed, where 32 participants watched different music video clips and assessed their induced emotions. These self assessments were used as ground-truth and the results of classification using audio, visual and audiovisual features extracted from music video clips are presented and compared.

Categories and Subject Descriptors

H.5.2 [[**INFORMATION INTERFACES AND PRESENTATION**]: User Interfaces—*Evaluation/methodology*; I. 5.4 [**PATTERN RECOGNITION**]: Applications—*Signal processing, Waveform analysis*

General Terms

Algorithms, Measurement, Performance, Experimentation,

Keywords

Affect, Emotion, Multimedia content analysis

1. INTRODUCTION

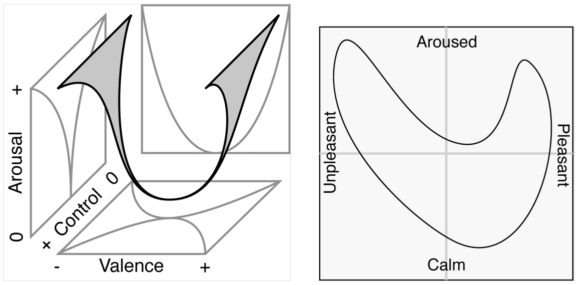
With the increasing popularity of video-on-demand (VOD) and personalized recommendation services, the development of automatic content descriptions for annotation, retrieval and personalization purposes is a key issue. The technology required to achieve this goal is referred to as multimedia content analysis (MCA) and aims at bridging the "semantic gap", that is, to develop models of the relationship between low level features and the semantics conveyed by multimedia contents. To this end, two different approaches have been adopted for MCA so far: cognitive and emotional. The cognitive approach analyzes a piece of multimedia content in terms of the semantics of a scene: location, characters and events. In the past decade, most of the MCA-related research efforts were focused on these methods.

The emotional approach, on the other hand attempts to characterize a given multimedia content by the emotions it may elicit in viewers. This approach is often referred to as affective MCA or affective content analysis and predicts or infers viewers' emotional reactions when perceiving multimedia contents. The emotional approach has been less investigated when compared to cognitive approach, but its importance has been rapidly increasing with the growing awareness of the role that the emotional load of multimedia and viewers reactions to it, play in VOD concepts and personalized multimedia recommendation. Analyzing a multimedia content at affective level reveals information that describes the emotional value of it. This value can be defined as amount and type of the affect (feeling or emotion) of the audience while they consume this multimedia content.

In order to analyze a given multimedia content at affective level, an appropriate modeling for emotion must be developed and used. How to represent and model emotions is, however, a challenging task. Until today, numerous theorists and researchers have conducted research on this subject and consequently a large amount of literature exists today with sometimes very different solutions. Generally, there are two different families of emotion models: the categorical models and the dimensional models. The rational for the categorical models is to have discrete basic categories of emotions from which every other emotion can be built by combining these basic emotions. The most common basic emotions are 'fear', 'anger', 'sadness', 'joy', 'disgust', 'surprise' found by Ekman [4]. The dimensional models describe the components of emotions and are often represented as a two dimensional or three dimensional space where the emotions are presented as points in the coordinate space of these dimensions.

The goal of the underlying dimensional model is not to find a finite set of emotions as in the categorical model but to find a finite set of underlying components of emotions. Many theorists have proposed that emotions can be modeled with three underlying dimensions namely, Valence (V) or Pleasantness (P), Arousal (A), Dominance (D) or

MIRUM'11, November 30, 2011, Scottsdale, Arizona, USA.



(a) (b)

Figure 1: Relevant areas of the (a) 3-D emotion space and (b) 2-D emotion space[3].

Control (C). The dimension 'valence' provides information about the degree of pleasantness of the content and ranges from pleasant (positive) to unpleasant (negative). The dimension 'arousal' represents the inner activation and ranges from energized to calm. The third dimension dominance (control) helps to distinguish between 'grief' and 'rage' and goes from no control to full control [5]. Psycho-physiological experiments revealed that only certain areas of the 3-D P-A-C space are relevant. Affective responses of a large group of subjects are included in their experiments. The stimuli are collected from the International Affective Picture system (IAPS) [8] and the International Affective Digitized Sounds system (IADS) [2]. Figure 1 (a) shows the relevant areas of the 3-D P-A-C space. The P-A model is a simplified 2-D model of the P-A-D model. The underlying dimensions are arousal and pleasure (valence). The emotions are represented as points in the 2-D coordinate space. A large number of emotional states can be represented using this 2-D model and many studies about affective content analysis use this simplified model as for example in [5] or in [7].
Similar to the 3D P-A-D model, only some parts of the 2-D space is relevant as shown by figure 1 (b).

In this paper, the P-A-D model of emotion is used and multimedia content analysis of music video clips is performed to extract their arousal and valence information. The rest of the paper is organized as follows. Section 2 provides the description of the database and the methodology used in this study. Section 3 presents the results obtained and section 4 concludes the paper.

2. AFFECTIVE MULTIMEDIA MODELING FOR MUSIC VIDEOS

The aim of this section is to explain and show how affective multimedia content analysis is performed. In the first part of the section, the database of music video clips used in this study is described. Then, the features which are extracted from the music video clips (audio features and video features) are introduced and finally, in the last part of this section, feature conditioning and classification methods used in this study are explained.

2.1 Database description

The music video clips were taken from the DEAP (Database for Emotion Analysis using Physiological Signals) database [7]. For selection of emotional music video clips, a web-based subjective test was developed and performed where subjects were asked to watch 120 music video clips and rate their perceived emotion. More precisely, the subjects used a discrete 9 point scale to rate: valence, with the range going from unhappy or sad to happy or joyful, arousal, with the range going from calm or bored to stimulated or excited, dominance, with the range going from submissive or 'without control' to dominant or 'in control, empowered', and whether they liked the video or not. Using this subjective data, 40 music video clips were selected so that only the music video clips which induce strong emotions are used. More precisely, 10 music video clips from each quadrant or arousal-valence space, which all had the strongest possible volunteer ratings with a small variation were selected. More information about the selection procedure can be found in [7].

After selecting the test material, 32 participants (50% female), aged between 19 and 37 (mean age 26.9), participated in the experiment to create the DEAP. While they were watching the 40 videos their physiological signals were recorded. After watching each music video clip they were asked to perform the rating of their perceived emotion. In this paper, we are only interested in users' self-assessment ratings and the music video clips. More details about the measurements are available in the paper describing the creation of the DEAP [7].

2.2 Feature extraction

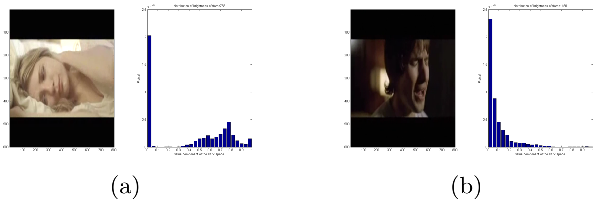
Affective multimedia content analysis and modeling is a complex problem. In general, machine learning approaches are used to deal with this problem and low-level features are extracted from the multimedia content in order to be mapped to different self assessed emotions. This section introduces the features that are extracted from music video clips and provides information on how these low-level features and the emotions relate to each other.

2.3 Related work

Several studies show that color, light, motion and the number of shots in a video influence the emotional perception of a spectator.
In [6] the authors developed a Hidden Markov Model (HMM)-based affective content analysis model to map low level features of video data to high level emotional events. In their study, they discovered a strong relationship between the low level features motion, color and shot cut rate and the emotions fear/anger, joy and sadness. In [15], the authors used sound effects from comedy and horror movies to detect some audio emotional events. They trained a four state HMM to detect the two emotions 'amusement' and 'fear'. Other researchers studied the characteristics of sound energy dynamics to detect emotional video segments [12]. With their algorithm they detected four types of emotion 'fear', 'apprehension or emphasis','surprise or alarm' and 'apprehension up to climax.

Sun [14] used a Video Affective Tree (VAT) and HMM to recognize the basic emotions 'joy', 'anger', 'sadness' and 'fear'. A relationship between these four basic emotions and the features 'color', 'motion', 'shot cut rate' and 'sound information' was found.

Rasheed et al. [13] classified movies into four broad categories: Comedies, Action, Dramas and Horror films by using low-level video features as average shot length, color variance, motion content and lighting key. They combined the features into a framework to provide a mapping of these four



(a) (b)

Figure 2: (a) HIGH Lighting Key and (b) LOW Lighting Key [3].

high-level semantic classes. They consider their work as a step toward high-level semantic film interpretation.
In [7], the authors proposed to extract the low - level features 'average shot change rate', 'shot length variance', 'magnitudes of motion vectors for all B- and P-frames', 'histogram of hue and lightness values in the HSV - space' and some sound characteristics in order to map these low-level features to the two dimensional arousal - valence space.

2.4 Video feature extraction

Most previous work reported that there is a relationship between the low-level features 'motion', 'color', 'shot length' and 'lighting key'. These four different features were extracted from the music videos.

2.4.1 Lighting key

Filmmakers often use lighting as an important agent to evoke emotions. They use multiple light sources to balance the direction and the intensity of light in order to create for example dramatic effects with the contrast of shadow and light or direct the attention of the viewer [13]. A simple computation method is chosen to detect the key of lighting in order to have a measure of lighting key of a frame and to distinguish between high-key lighting and low-key lighting [13].

The algorithm to compute the key lighting is based on the fact, that the proportion of bright pixels in hight-key shots is high whereas the the proportion of bright pixels in low-key shots is low. Figures 2 (a) and (b) show examples of a high-key shot and a low-key shot with the distribution of brightness, respectively. The 25 bin histograms are computed by taking the value component of the HSV space. The mean and variance is low for low-key shots and high for high-key shots, so that the lighting quantity $\zeta_i(\mu, \sigma)$ for a frame i with m x n pixels is defined as

$$\zeta_i = \mu_i * \sigma_i. \tag{1}$$

where μ_i and σ_i represent the mean and the standard deviation values, respectively, computed from the value component of the HSV space of a frame i.

2.4.2 Key frame and shot boundary detection

In [6] and [14], the authors demonstrated that there is a relationship between the average shot length and the affective content. In [10] and [11], the authors proposed a method based on the color histogram. They investigated the differences between color histograms of frames belonging to a video sequence in order to detect hard cuts, fades or dissolves. These methods are computationally expensive and can not be used in real time applications.

In this work, the proposed method in [1] is used. The algorithm is based on singular value decomposition (SVD) and extracts low-cost, multivariate color features to construct 2D feature matrices. It is relatively a difficult task to extract shot boundaries and the results often have many false detections. Finally , three features are taken from the extracted shot boundaries: The average shot length of a video sequence and the highest and the lowest shot length found in video sequences were taken as features.

Table 1: Summary of the extracted video features

feature	description
Lighting Key 3 features	average value of lighting key of frames of a video sequence number of High key and Low Key frames
Shot boundary 3 features	average shot length of a video highest and shortest length of a shot in a video
color 3 features	median of maximum, minimum and median value of hue component of the HSV space
Motion 2 features	mean of median and mean absolute value of the motion vector

2.4.3 color

Many research studies on affective content analysis found a relationship between colors and evoked emotions in viewers. For example [6] shows that the colors 'yellow', 'orange' and 'red' correspond to the emotion 'fear' and 'anger' and the color 'blue', 'violet' and 'green' can be found when the spectator feels 'high valence' and 'low arousal'.

In this work, the color features are extracted in the following manner. From each frame of a video sequence, a color histogram was computed. For a given frame i with n x m pixels, the hue component of the HSV space is computed. Then the maximum, minimum, mean and median hue values of frame i were calculated. The median of these maximum, minimum, mean and median values over all frames of a video sequence were considered as features.

2.4.4 Motion vectors

Some of the related works found a relationship between motion and affective content. Sun [14] shows the relation between the emotions 'joy', 'anger', 'sadness' and 'fear' and the camera motion. A fast Motion Vector is computed for every 4th frame in a video sequences. A block size of 16 is chosen. The median and mean of the absolute value of the motion vector of each frame was computed. The mean over all median and mean values was computed in order to get two features.

The algorithm used in this study is a simplified block matching algorithm (BMA), so that it can be used in a real time problem. The aim of a block matching algorithm is to locate and follow blocks in a sequence of a digital video. In other words, the algorithm finds blocks of a given frame i in an other frame j. The displacement of two blocks is the motion of all pixels of the block.

Table 1 presents an overview of the extracted video features. Basically, the above-mentioned features were extracted from the frames. The final feature vector was computed by taking the mean value of the features of all frames in a sequence. One to twelve samples were computed from a video by varying the sequence size from 60 seconds to 5 seconds.

2.5 Audio feature extraction

As reported in several research studies, sound can have a close relationship with the affective content of a music video clip. In this work some characteristics of sound were selected and used as features. Most of the features are computed by using the matlab MIR toolbox.

2.5.1 Zero-crossing rate

The zero-crossing rate is defined as the number of times the signal crosses the zero line (x - axis) per unit time. In other words, it is the number of times the signal change its sign per unit time.

2.5.2 Energy

The global energy of the signal is computed by simply taking the root-mean-square (RMS).

$$x_{rms} = \sqrt{\frac{1}{n}\sum_{i=1}^{n} x_i^2} \tag{2}$$

2.5.3 Mel-frequency cepstral coefficients (MFCC)

The MFCC can be seen as the description of the spectral shape of the sound. Most of the signal information can be found in low - frequency coefficients. Therefore, only the 13 first coefficients are used as features.

2.5.4 Delta MFCC coefficient (ΔMFCC) and autocorrelation MFCC (aMFCC)

The ΔMFCC and the aMFCC provide quantitative measures of the movement of the MFCC and can be derived as follows:

$$\Delta MFCC_i(\nu) = MFCC_{i+1}(\nu) - MFCC_i(\nu) \tag{3}$$

$$aMFCC_i^{(l)}(\nu) = \frac{1}{L}\sum_{j=i}^{1+l}(MFCC_j(\nu) * MFCC_{j+l}(\nu)) \tag{4}$$

$MFCC_i(\nu)$ represent the νth MFCC of frame i and L denotes the correlation window length. The superscript l denotes the value of correlation lag [9].

2.5.5 Linear prediction coefficient (LPC)

LPC estimates the coefficients of a forward linear predictor by minimizing the prediction error in the least squares sense. The basic idea of linear prediction analysis is that a music sample can be estimated from the combination of the past samples. In [**?**], a unique set of predictor coefficients can be determined, by minimizing the sum of the squared differences.

2.5.6 Delta LPC coefficient (ΔLPC)

These feature can be interpreted as a quantitative measure of the movement of the LPC and is defined as follows [9].

$$\Delta LPC_i(\nu) = LPC_{i+1}(\nu) - LPC_i(\nu) \tag{5}$$

2.5.7 Silence ratio

The average Silence Ratio is computed by calculating fraction between 'very low' energy over the whole signal.

2.5.8 Pitch

The best pitch frequency is computed. First of all the FFT spectrum of the audio signal is computed. Then an autocorrelation is done in order to find the pitch frequencies. Only the best pitch frequency is chosen.

Table 2: Summary of the extracted audio features

description	number of features
Zero Crossing	1 feature
Energy	1 feature
MFCC	13 features
ΔMFCC	13 features
Autocorrelation MFCC	7 features
LPC	13 features
ΔLPC	13 features
Silence Ratio	1 feature
Pitch	1 feature
centroid	1 feature
Band Energy Ratio	1 feature
Delta Spectrum Magnitude	1 feature

2.5.9 Spectral centroid

A centroid is computed on the frequency spectrum using short - time Fourier transform (STFT). The mean μ of the spectrum is taken as feature. Let $F_i = \{f_i(u)\}_{u=0}^{M}$ represent the short-time Fourier transform of the ith frame, and M denotes the index for the highest frequency band [9]. The spectral centroid of frame i is calculated as:

$$c_i = \frac{\sum_{u=0}^{M} u.|f_i(u)|^2}{\sum_{u=0}^{M} |f_i(u)|^2} \tag{6}$$

2.5.10 Band energy ratio (BER)

The bandwidth of the FFT of frame i [9] can be computed as

$$b_i^2 = \frac{\sum_{u=0}^{M}(u - c_i)^2.|f_i(u)|^2}{\sum_{u=0}^{M} |f_i(u)|^2} \tag{7}$$

2.5.11 Delta spectrum magnitude

The Delta Spectrum Magnitude is computed by taking the difference of the spectrum between the current and the previous frame. It can be seen as measure of the movement of the spectrum over time.

These features are extracted from the audio frames. An audio frame is defined as the audio signal of a frame. The length of an audio frame is computed by dividing the length of the audio signal by the number of video frames. The feature vector is computed by taking the mean over all frames of a sequence. Sequences between five seconds and 60 seconds are extracted so that 12 to one samples are computed per video. Table 2 gives an overview of the extracted audio features.

2.6 Classification

In order to detect the emotion induced in different participants while watching music video clips, audio and visual features were extracted from the video as described in the previous section. Then a classification system is needed to be trained so that it can classify the feature vectors correctly with a high success rate. Before training a classifier, the extracted feature vectors were conditioned to remove the linear dependency of different features and to reduce their dimensionality.

The dynamic range of different features were very different. For some features the values were close to 0 and others had an order of magnitude around 100. This difference in dynamic range may cause an unwanted weighting in classification in that the features with higher dynamic range have greater impact on classification. In order to solve this problem, the feature vector was normalized. The normalization is done in the way that after the normalization the mean μ of a feature is zero and the standard deviation σ equals one.

After the audio and video feature vectors extracted for the analysis were normalized.they were further processed to remove their correlation and reduce their dimensionality. Principal component analysis (PCA) can be used as a method to select the most important components with respect to the variance. An orthogonal transformation is used in order to convert a set of observations of possible correlated variables into a set of values of uncorrelated variables.

The k-nearest neighbor classifier (kNN) and cross validation scheme were used in order to perform the classification. Due to the fact that only 40 music video clips exist in database, the leave-one-out cross validation scheme were used.

3. RESULTS AND DISCUSSION

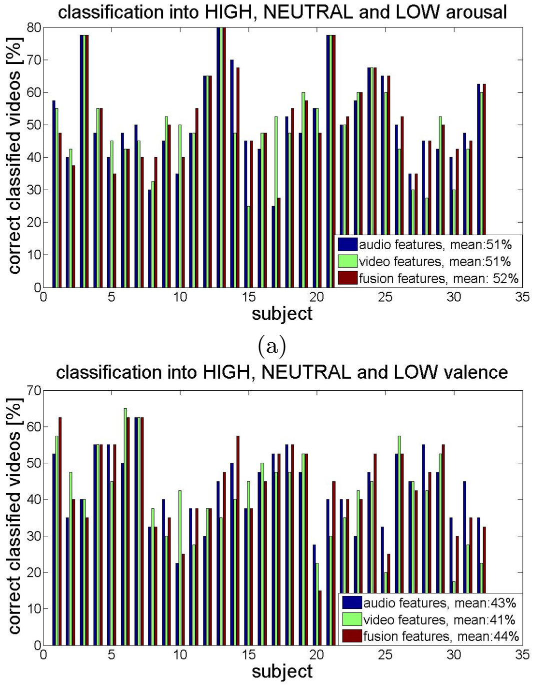
The classification can be performed for each subject separately, (as it is done in this work) or for the average ratings over all subjects. In the first case, we train the classifier for each subject separately. The same features can be mapped to other emotions for another subject and each music video clip is classified into other emotions depending on the subject. In the second case, the average 'arousal' and 'valence' values are considered to classify the video clips. Then the video clips are classified according to the average rating values into the emotion. The second classification protocol assumes that the subjective ratings for a video are all very close together with a small standard deviation. In other words, for this kind of classification, it is favorable that the video clips evoke the same or similar emotions in all subjects.

The 40 music video clips from the DEAP database are classified using two different classification protocols namely, three classes classification (high, neutral, low arousal or positive, neutral, negative valence) and four classes classification (quadrants of arousal-valence space). These quadrants are positive valence high arousal, negative valence high arousal, negative valence low arousal, and positive valence low arousal, which correspond to happiness, anger and fear, sadness and calmness, respectively.

As described in previous section, different features were extracted form the video and audio contents of the 40 music video clips used in this study. In orther to fuse the audio and video features, these features were concatenated together to form audiovisual features.

These features were extracted from the frames and the feature vector was computed by taking the mean value over a certain sequences of frames. The length of a sequence was varied from 60 seconds to five seconds, so that one to 12 sequences per video were extracted. Some related works proposed to take the shot as length of a sequence. These was not done in this work since the number of shots per video varied a lot and it is needed to have a similar number of sequences for all video clips. Furthermore, for some video clips, the algorithm detected only one shot in the whole video .

The feature vectors were then normalized as described in a previous section in order to have a similar range of values of the extracted features and PCA was used to reduce the dimension. More precisely, It was observed that by taking only the 32 most significant components only 1% of information is lost. Therefore, the feature vector of 77 audio and video features is reduced to the 32 most important features.



(a)

(b)

Figure 3: Results of classification into (a) HIGH, LOW and NEUTRAL arousal, (b)POSITIVE, NEUTRAL and NEGATIVE valence.

Figures 3, and 4 illustrate the classification results of the music video clips into the underlying basic emotion dimensions 'arousal' and 'valance' according to the classification protocols introduced in the beginning of this section.

On each figure the classification is done for three different feature vectors: only 'audio' features, only 'video' features and the 'fusion' of both. After the feature extraction the size of the feature vector is reduced by PCA and the kNN - classifier is used in a leave-one-out scheme.

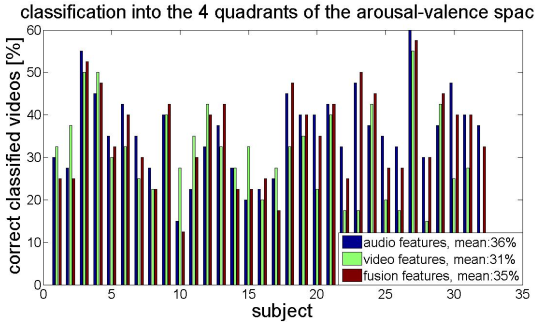


Figure 4: classification into the 4 quadrants of the VA-space

As it can be seen in these figures, the results of affective multimedia content analysis is very subjective and for some subjects, a relatively high accuracy can be achieved, whereas for some other subjects the results of classification is not good. Furthermore, it can be observed that the classification results obtained using audio features often outperform the classification results obtained using only visual features. This is more significant for classification of arousal. It can be seen that fusing the audio and visual features by concatenating these features will not improve the results drastically. The classification accuracy obtained using this feature fusion, is only occasionally better than the results obtained using either of the single modalities. This can be due to curse of dimensionality. In other words, the number of variables will be increased while the number of samples is still low.

4. CONCLUSIONS

This paper investigated affective multimedia content analysis or the analysis of multimedia contents in order to extract cues for determining and predicting users' affective states. To this end, low level features were extracted form music videos (video and audio) with the aim of finding a relationship between them and different emotions. The groundtruth was obtained through a subjective test where 32 participants watched the music video clips used in this study and assessed their induced emotions. The results revealed that the audio features play an important role in determining participant's induced emotion. Further research on this subject would be to explore different feature extraction techniques which also consider semantics involved in self assessment of emotions. In this study, fusion of audio, and visual features is performed by concatenation of features extracted from audio and video contents. however, it was observed that feature fusion is not an appropriate technique due to the fact that it increases the numbers of variables while the number of samples is still very low. Other fusion techniques such as decision fusion can be further studied to explore whether they can improve the results obtained by single modalities.

5. ACKNOWLEDGMENTS

The research leading to these results has been performed in the frameworks of Swiss National Foundation for Scientic Research (FN 200020-132673-1), European Community's Seventh Framework Program (FP7/2007-2011) under grant agreement no. 216444 (PetaMedia), and the NCCR Interactive Multimodal Information Management (IM2).

6. REFERENCES

[1] W. Abd-Almageed. Online, simultaneous shot boundary detection and key frame extraction for sports videos using rank tracing. In *Image Processing, 2008. ICIP 2008. 15th IEEE International Conference on*, pages 3200–3203. IEEE, 2008.

[2] M. Bradley, P. Lang, U. of Florida. Center for the Study of Emotion, Attention, and N. I. of Mental Health. *The International affective digitized sounds (IADS)[: stimuli, instruction manual and affective ratings*. NIMH Center for the Study of Emotion and Attention, 1999.

[3] R. Dietz and A. Lang. Affective agents: Effects of agent affect on arousal, attention, liking and learning. In *Proceedings of the Third International Cognitive Technology Conference, San Francisco*, 1999.

[4] P. Ekman. Basic emotion. *Handbook of Cognition and Emotion*, 1999.

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

[6] H. Kang. Affective content detection using hmms. In *Proceedings of the eleventh ACM international conference on Multimedia*, pages 259–262. ACM, 2003.

[7] S. Koelstra, C. Muehl, M. Soleymani, J. Lee, A. Yazdani, T. Ebrahimi, T. Pun, A. Nijholt, and I. Patras. Deap: A database for emotion analysis using physiological signals. *IEEE Transactions on Affective Computing*, 2011.

[8] P. Lang, M. Bradley, and B. Cuthbert. International affective picture system (iaps): Instruction manual and affective ratings. *The Center for Research in Psychophysiology, University of Florida*, 1999.

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

[10] R. Lienhart. Comparison of automatic shot boundary detection algorithms. In *Proc. SPIE*, volume 3656, pages 290–301. Citeseer, 1999.

[11] J. Mas and G. Fernandez. Video shot boundary detection based on color histogram. *Notebook Papers TRECVID2003, Gaithersburg, Maryland, NIST*, 2003.

[12] S. Moncrieff, C. Dorai, and S. Venkatesh. Affect computing in film through sound energy dynamics. In *Proceedings of the ninth ACM international conference on Multimedia*, pages 525–527. ACM, 2001.

[13] Z. Rasheed, Y. Sheikh, and M. Shah. On the use of computable features for film classification. *Circuits and Systems for Video Technology, IEEE Transactions on*, 15(1):52–64, 2005.

[14] K. Sun and J. Yu. Video affective content representation and recognition using video affective tree and hidden markov models. *Affective Computing and Intelligent Interaction*, pages 594–605, 2007.

[15] M. Xu, L. Chia, and J. Jin. Affective content analysis in comedy and horror videos by audio emotional event detection. In *2005 IEEE International Conference on Multimedia and Expo*, page 4. IEEE, 2005.