A Bayesian Framework for Video Affective Representation

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

Abstract

Emotions that are elicited in response to a video scene contain valuable information for multimedia tagging and indexing. The novelty of this paper is to introduce a Bayesian classification framework for affective video tagging that allows taking contextual information into account. A set of 21 full length movies was first segmented and informative content-based features were extracted from each shot and scene. Shots were then emotionally annotated, providing ground truth affect. The arousal of shots was computed using a linear regression on the content-based features. Bayesian classification based on the shots arousal and content-based features allowed tagging these scenes into three affective classes, namely calm, positive excited and negative excited. To improve classification accuracy, two contextual priors have been proposed: the movie genre prior, and the temporal dimension prior consisting of the probability of transition between emotions in consecutive scenes. The f1 classification measure of 54.9% that was obtained on three emotional classes with a naïve Bayes classifier was improved to 63.4% after utilizing all the priors.

1. Introduction

1.1. Overview

Video and audio on-demand systems are getting more and more popular and are likely to replace traditional TVs. Online video content has been growing rapidly in the last five years. For example the open access online video database, YouTube, had a watching rate of more than 100 millions videos per day in 2006 [1]. The enormous mass of digital multimedia content with its huge variety requires more efficient multimedia management methods. Many studies have been conducted in the last decade to increase the accuracy of current multimedia retrieval systems. These studies were mostly based on content analysis and textual tags [2]. Although the emotional preferences of a user play an important role in multimedia content selection, few publications exist in the field of affective indexing which consider emotional preferences of users [2-7].

The present study is focused on movies because they represent one of the most common and popular types of multimedia content. An affective representation of scenes will be useful for tagging, indexing and highlighting of important parts in a movie. We believe that using the existing online metadata can improve the affective representation and classification of movies. Such metadata, like movie genre, is available on internet (e.g. internet movie database http://www.imdb.com). Movie genre can be exploited to improve an affect representation system's inference about the possible emotion which is going to be elicited in the audience. For example, the probability of a happy scene in a comedy certainly differs from that in a drama. Moreover, the temporal order of the evoked emotions, which can be modeled by the probability of emotion transition in consecutive scenes, is also expected to be useful for the improvement of an affective representation system.

It is shown here how to benefit from the proposed priors in a Bayesian classification framework. Affect classification was done for a three labels scene classification problem, where the labels are “calm”, “positive excited”, and “negative excited”. Ground truth was obtained through manual annotation with a FEELTRACE-like [8] annotation tool with the self-assessments serving as the classification ground-truth. The usefulness of priors is shown by comparing classification results with or without using them.

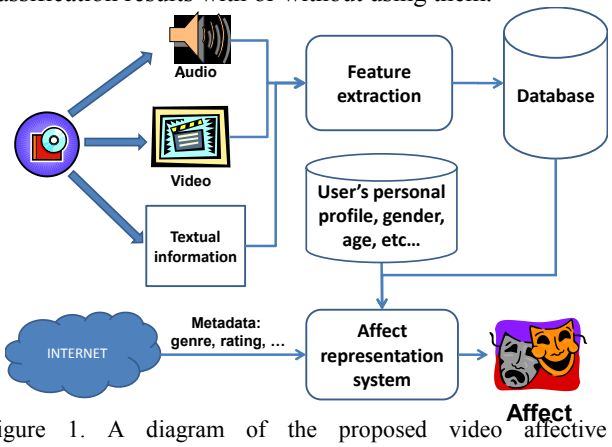

Figure 1. A diagram of the proposed video affective representation.

In our proposed affective indexing and retrieval system, different modalities, such as video, audio, and textual data (subtitles) of a movie will be used for feature extraction. Figure 1 shows the diagram of such a system. The feature extraction block extracts features

from the three modalities and stores them in a database. Then, the affect representation system fuses the extracted features, the stored personal information, and the metadata to represent the evoked emotion. For a personalized retrieval, a personal profile of a user (with his/her gender, age, location, social network) will help the affective retrieval process.

The paper is organized as follows. A review of the current state of the art and an explanation on affect and affective representation are given in the following subsections of the first Section. Methods used including the arousal representation at the shot level and affect classification at the scene level are given in Section 2. Section 3 details the movie dataset used and the features that have been extracted. The obtained classification results at the scene level, the comparisons with and without using genre and temporal priors are discussed in Section 4. Section 5 concludes the article and offers perspectives for future work.

1.2. State of the art

Video affect representation requires understanding of the intensity and type of user's affect while watching a video. There are only a limited number of studies on content-based affective representation of movies. Wang and Cheong [2] used content audio and video features to classify basic emotions elicited by movie scenes. In [2] audio was classified into music, speech and environment signals and had been treated separately to shape an affective feature vector. The audio affective vector was used with video-based features such as key lighting and visual excitement to form a scene feature vector. Finally, the scene feature vector was classified and labeled with emotions.

Hanjalic et al. [4] 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 based features based on their relation to the valence-arousal space that was defined as an affect model (for the definition of affect model, see Section 1.3) [4]. Then, arising emotions were estimated in this space by combining these features. While arousal and valence could be used separately for indexing, they combined these values by following their temporal pattern in the arousal and valence space. This allowed determining an affect curve, shown to be useful for extracting video highlights in a movie or sports video.

A hierarchical movie content analysis method based on arousal and valence related features was presented by M. Xu et al. [6]. In this method the affect of each shot was first classified in the arousal domain using the arousal correlated features and fuzzy clustering. The audio short time energy and the first four Mel frequency cepstral coefficients, MFCC (as a representation of energy features), shot length, and the motion component of consecutive frames were used to classify shots in three arousal classes. Next, they used color energy, lighting and brightness as valence related features to be used for a HMM-based valence classification of the previously arousal-categorized shots.

A personalized affect representation method based on a regression approach for estimating user-felt arousal and valence from multimedia content features and/or from physiological responses was presented by Soleymani et al. [7]. A relevance vector machine was used to find linear regression weights. This allowed predicting valence and arousal from the measured multimedia and/or physiological data. During the experiments, 64 video clips were shown to 8 participants while their physiological responses were recorded; user's self-assessments of valence and arousal served as ground truth. A comparison was made on the arousal and valence values obtained by different modalities which were the physiological signals, the video- and audio-based features, and the self-assessments. In [7] An experiment with multiple participants has been conducted for personalized emotion assessment based on content analysis.

1.3. Affect and Affective representation

Russell [10] proposed a 3D continuous space called the valence-arousal-dominance space which was based on a self-representation of emotions from multiple subjects. In this paper we use a valence-arousal dimensional approach for affect representation and annotation. The third dimensional axis, namely dominance / control, is not used in our study. In the valence-arousal space it is possible to represent almost any emotion. The valence axis represents the pleasantness of a situation, from unpleasant to pleasant; the arousal axis expresses the degree of felt excitement, from calm to exciting. Russel demonstrated that this space has the advantages of being cross-cultural and that it is possible to map labels on this space. Although, the most straightforward way to represent an emotion is to use discrete labels such as fear, anxiety and joy, label-based representations have several disadvantages. The main one is that despite the universality of basic emotions, the labels themselves are not universal. They can be misinterpreted from one language (or culture) to another. In addition, emotions are continuous phenomena rather than discrete ones and labels are unable to define the strength of an emotion.

In a dimensional approach for affect representation, the affect of a video scene can be represented by its coordinates in the valence-arousal space. Valence and arousal can be determined by self reporting. The goal of an affective representation system is to estimate user's valence and arousal or emotion categories in response to each movie segment. Emotion categories are defined as regions in the valence-arousal space. Each movie consists of scenes and each scene consists of a sequence of shots which are happening in the same location. A

shot is the part of a movie between two cuts which is typically filmed without interruptions [11].

2. Methods

2.1. Arousal estimation with regression on shots

Informative features for arousal estimation include loudness and energy of the audio signals, motion component, visual excitement and shot duration. Using a method similar to Hanjalic et al. [4] and to the one proposed in [7], the felt arousal from each shot is computed by a regression of the content features (see Section 3 for a detailed description). In order to find the best weights for arousal estimation using regression, a leave one movie out strategy on the whole dataset was used and the linear weights were computed by means of a relevance vector machine (RVM) from the RVM toolbox provided by Tipping [12]. The RVM is able to reject uninformative features during its training hence no further feature selection was used for arousal determination.

Equation (1) shows how N_s audio and video based features z^k_i of the k-th shot are linearly combined by the weights w_i to compute the arousal $\hat{a}_k$ at the shot level.

$$\hat{a}_k = \sum_{i=1}^{N_s} w_i z_i^k + w_0 \quad (1)$$

After computing arousal at the shot level, the average and maximum arousals of the shots of each scene are computed and used as arousal indicator features for the scene affective classification. During an exciting scene the arousal related features do not all remain at their extreme level. In order to represent the highest arousal of each scene, the maximum of the shots' arousal was chosen to be used as a feature for scene classification.

The linear regression weights that were computed from our data set were used to determine the arousal of each movie's shots. This was done in such a way that all movies from the dataset except for the one to which the shot belonged to were used as the training set for the RVM. Any missing affective annotation for a shot was approximated using linear interpolation from the closest affective annotated time points in a movie.

It was observed that arousal has higher linear correlation with multimedia content-based features than valence. Valence estimation from regression is not as accurate as arousal estimation and therefore valence estimation has not been performed at the shot level.

2.2. Bayesian framework and scene classification

For the purpose of categorizing the valence-arousal space into three affect classes, the valence-arousal space was divided into the three areas shown in Figure 2, each corresponding to one class. According to [13] emotions mapped to the lower arousal category are neither extreme pleasant nor unpleasant emotions and are difficult to differentiate. Emotional evaluations are shown to have a heart shaped distribution on valence-arousal space [13]. Hence, we categorized the lower half of the plane into one class. The points with an arousal of zero were counted in class 1 and the points with arousal greater than zero and valence equal to zero were considered in class 2. These classes were used as a simple representation for the emotion categories based on the previous literature on emotion assessment [14].

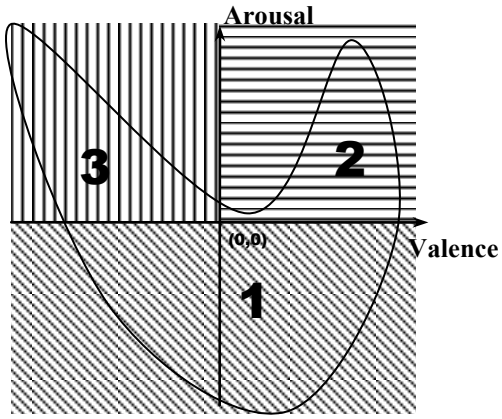

Figure 2. Three classes in the valence-arousal space are shown, namely calm (1), positive excited (2) and negative excited (3). An approximate of the heart shaped distribution of valence and arousal is shown.

In order to characterize movie scenes into these affective categories, the average and maximum arousal of the shots of each scene and the low level extracted audio- and video- based features were used to form a feature vector. This feature vector in turn was used for the classification.

If the content feature vector of the j-th scene is x_j, the problem of finding the emotion class, $\hat{y}_j$, of this scene is formulated as estimating the $\hat{y}_j$ which maximizes the probability $p(y_j|x_j,\theta)$ where θ is the prior information which can include the user's preferences and video clip's metadata. In this paper one of the prior metadata (θ) we used is for instance the genre of the movie. Personal profile parameters can be also added to θ. Since in this paper the whole affect representation is trained by the self report of one participant the model is assumed to be personalized for this participant. When the emotion of the previous scene is used as another prior the scene affect probability formula changes to $p(y_j|y_{j-1},x_j,\theta)$. Assuming for simplification that the emotion of the previous scene is independent from the content features of the current scene this probability can be reformulated as:

$$p(y^j \mid y^{j-1}, x^j, \theta) = \frac{p(y^{j-1} \mid y^j, \theta).p(y^j \mid x^j, \theta)}{p(y^{j-1} \mid \theta)} \quad (2)$$

The classification problem is then be simplified into the determination of the maximum value of the numerator of Equation (2), since the denominator will be the same for all different affect classes y_j. The priors are established based on the empirical probabilities obtained from the training data. For example, the occurrence

probability of having a given emotion followed by any of the emotion categories was computed from the participant's self-assessments and for each genre. This allowed to obtain the $p(y_{j-1}|y_j,\theta)$. Different methods were evaluated to estimate the posterior probability $p(y_j|x_j)$. A naïve Bayesian approach which assumes the conditional probabilities are Gaussian was chosen as providing the best performance on the dataset; the superiority of this method can be attributed to its generalization abilities.

3. Material description

A dataset of movies segmented and affectively annotated by arousal and valence is used as the training set. This training set consists of twenty one full length movies (mostly popular movies). The majority of movies were selected either because they were used in similar studies (e.g. [15]), or because they were recent and popular. The dataset included four genres: drama, horror, action, and comedy. The following three information streams were extracted from the media: video (visual), sound (auditory), and, subtitles (textual).

The video stream of the movies has been segmented at the shot level using the OMT shot segmentation software and manually segmented into scenes [16;17]. Movie videos were encoded into the MPEG-1 format to extract motion vectors and I frames for further feature extraction. We used the OVAL library (Object-based Video Access Library) [18] to capture video frames and extract motion vectors.

The second information stream, namely sound, has an important impact on user's affect. For example according to the findings of Picard [19], loudness of speech (energy) is related to evoked arousal, while rhythm and average pitch in speech signals are related to valence. The audio channels of the movies were extracted and encoded into monophonic information (MPEG layer 3 format) at a sampling rate of 48 kHz. All of the resulting audio signals were normalized to the same amplitude range before further processing.

Textual features were also extracted from the subtitles track of the movies. According to [9] the semantic analysis of the textual information can improve affect classification. As the semantic analysis over the textual data was not the focus of our work we extracted simple features from subtitles by tokenizing the text and counting the number of words. These statistics have been used with the timing of the subtitles to extract the talking rate feature which is the number of words that had been spoken per second on the subtitles show time. The other extracted feature is the number of spoken words in a scene divided by the length of the scene, which can represent the amount or existence of dialogues in a scene. A list of the movies in the dataset and their corresponding genre is given in Table 1.

Table1. List of the movies in the dataset.

Drama Movies	**Comedy Movies**
The pianist, Blood diamond, Hotel Rwanda, Apocalypse now, American history X, Hannibal	Man on the moon, Mr. Bean's holiday, Love actually, Shaun of the dead, Shrek
Horror Movies	**Action Movies**
Silent hill, Ringu (Japanese), 28 days later, The shining	Man on Fire, Kill Bill Vol. 1, Kill Bill Vol. 2, Platoon, The thin red line, Gangs of New York

3.1. Audio features

A total of 53 low-level audio features were determined for each of the audio signals. These features, listed in Table 2, are commonly used in audio and speech processing and audio classification [20;21].

Wang et al. [2] demonstrated the relationship between audio type's proportions (for example, the proportion of music in an audio segment) and affect, where these proportions refer to the respective duration of music, speech, environment, and silence in the audio signal of a video clip. To determine the three important audio types (music, speech, environment), we implemented a three class audio type classifier using support vector machines (SVM) operating on audio low-level features in a one second segment. Before classification, silence had been identified by comparing the average audio signal energy of each sound segment (using the averaged square magnitude in a time window) with a pre-defined threshold empirically extracted from the first seven percent of the audio energy histogram. This audio histogram was computed from a randomly selected 30 minutes segment of each movie's audio stream.

Table 2. Low-level features extracted from audio signals.

Feature category	Extracted features
MFCC	MFCC coefficients (13 features) [20], Derivative of MFCC (13 features), Autocorrelation of MFCC (13 features)
Energy	Average energy of audio signal [20]
Formants	Formants up to 5500Hz (female voice) (five features)
Time frequency	Spectrum flux, Spectral centroid, Delta spectrum magnitude, Band energy ratio, [20;21]
Pitch	First pitch frequency
Zero crossing rate	Average, Standard deviation [20]
Silence ratio	Proportion of silence in a time window [24]

After removing silence, the remaining audio signals were classified by a SVM with a polynomial kernel, using the LIBSVM toolbox. (http://www.csie.ntu.edu.tw/~cjlin/libsvm/). The SVM was trained on about three hours of audio, extracted from movies (not from the dataset of this paper) and labeled manually. Despite the fact that in various cases the audio type classes were overlapping (e.g. presence of a musical background during a dialogue), the classifier was usually able to recognize the dominant audio type with an accuracy of about 80%.

The classification results were used to the ratio of

each audio type in a movie segment. MFCC, formants and the pitch of audio signals were extracted using the PRAAT software package [22].

3.2. Visual features

From a movie director's point of view, lighting key [2;23] and color variance [2] are important tools to evoke emotions. We therefore extracted lighting key from frames in the HSV space by multiplying the average value V (in HSV) by the standard deviation of the values V (in HSV). Color variance was obtained in the CIE LUV color space by computing the determinant of the covariance matrix of L, U, and V.

Hanjalic et al. [4] showed the relationship between video rhythm and affect. The average shot change rate, and shot length variance were extracted to characterize video rhythm. Fast moving scenes or objects' movements in consecutive frames are also an effective factor for evoking 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 are important parameters to elicit emotions [17]. In order to use colors in the list of video features, a 20 bin color histogram of hue and lightness values in the HSV space was computed for each I frame and subsequently averaged over all frames. The resulting averages for the 20 bins were used as video content-based features. The median of 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 [2].

3.3. Affective annotation

FEELTRACE is a self-assessment tool which was proposed by Cowie et al. [8] to assess emotion in the valence-arousal space. In this assessment tool, the coordinates of a pointer manipulated by the user are continuously recorded during the show time of the stimuli (video, image, or external source) and used as the affect indicators. Inspired by this tool designed for psychological studies, an affective self reporting tool has been implemented to assess emotion during the watching of a video. The emotion is recorded as the coordinates of the pointer on the click event.

A set of SAM manikins (Self-Assessment Manikins [25]) are generated for different combinations of arousal and valence to help the user understand the emotions related to the regions of valence-arousal space. *E.g.* the positive excited manikin is generated by combining the positive manikin and the excited manikin. A preview of the annotation software is given in Figure 3.

The participant was asked to annotate the movies so as to indicate at which times his/her felt emotion has changed. Thus, the valence and arousal values received from the participant should occur when there was a change in the participant's emotion. The participant was asked to indicate at least one point during each scene not to leave any scene without assessment. Continuous annotation of the movies is a time consuming process; hence the participant was asked to annotate at most two movies per day of different genres.

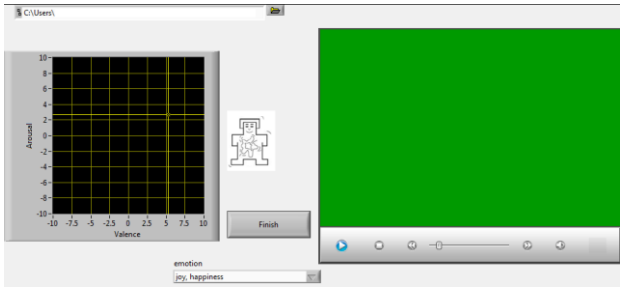

Figure 3. A snapshot of the affective annotation software which is implemented in LABVIEW. The positive excited manikin can be seen in the central part of the display.

4. Results

4.1. Arousal estimation of shots

Figure 4 shows a sample arousal curve from part of the film entitled "Silent Hill". Figure 4 is a typical example of the obtained results on arousal estimation. The estimated affect curve, in the first half, fairly closely follows the self-assessment curve. This moreover shows the correlation between arousal related content features and participant's self-estimated affect. The participant's felt emotion was however not completely in agreement with the estimated curve, as can for instance be observed in the second half of the plot. A possible cause for the discrepancy is the low temporal resolution of the self-assessment. Another possible cause is experimental weariness: after having had exciting stimuli for minutes, a participant's arousal might be decreasing despite strong movements in the video and loud audio. Finally, some emotional feelings might simply not be captured by low-level features; this would for instance be the case for a racist comment in a movie dialogue which evokes disgust for a participant. Without some form of semantic high level analysis of the movie script the content features are unable to detect verbal behavior in movie scenes.

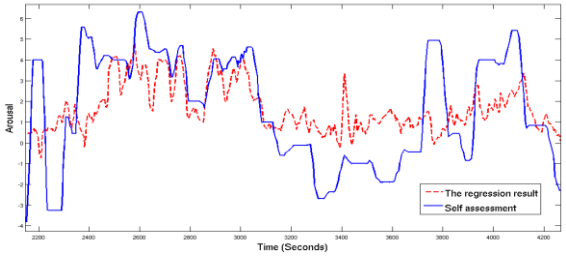

Figure 4. Five-points smoothed shot arousal curve (full line), and corresponding self-assessments (dashed line).

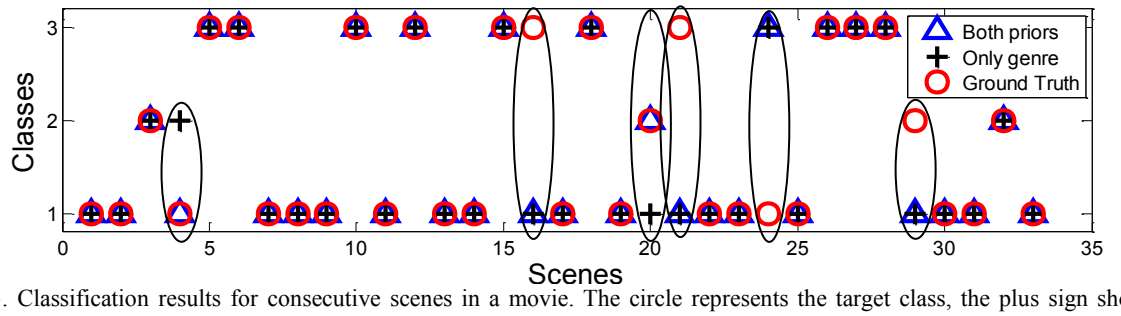

Figure 5. Classification results for consecutive scenes in a movie. The circle represents the target class, the plus sign shows the results of the Naïve Bayesian classifier with genre prior and the triangle shows the results with both genre and time priors. The samples which are misclassified by the Bayesian classifier with genre prior are encircled.

4.2. Classification results

The *f1* measure (Equation 3) that is commonly employed in information retrieval was used to evaluate the performance of emotion classification in a ten-folding cross validation:

$$f1 = \frac{2 \times precision \times recall}{precision + recall} \qquad (3)$$

For the ten-folding cross validation the original samples, movie scenes, were partitioned into 10 subsample sets. At each step of the cross validation, one set was used as the test-set and the rest were used as training-set. A forward feature selection method was used to find the most discriminant set of features [26]. All movie scenes were classified into three classes using a naïve Bayesian classifier alone, or complemented through the proposed Bayesian framework with different combination of priors. The classification accuracy, *f1* measure, and the confusion matrices are reported in Table 3.

The naïve Bayesian classifier results are shown in Table 3-a. A more complex SVM classifier was also evaluated on this dataset but did not demonstrate an improvement in terms of classification results. The superiority of the Bayesian classifier with priors over a more complex classifier such as the SVM shows that the improvement brought in through the use of adequate priors, can be higher than the one provided by a more complex classifier. Results achieved by such a support vector machine classifier with a linear kernel are reported in Table 3-e.

The results obtained using the preceding emotion prior (temporal prior) are given in Table 3-b. The affect class of the previous scene, that is the preceding emotion, was obtained by using the same trained classifier to classify the preceding scene's feature vector. Using the temporal prior, the improvement in the *f1* measure and accuracy of the classifier showed that even in case of misclassification of the previous scenes the Bayesian classifier is robust enough to slightly improve classification.

The genre prior was then included as the only prior; the classification results obtained are shown in Table 3-c. Finally the best results were obtained using both the genre and temporal priors as can be seen in Table 3-d.

Table 3. Affective scene classification accuracies and *f1* measures with different combinations of priors (on the left) and their confusion matrices (on the right). "1", "2", "3" correspond to the 3 classes "calm", "positive excited", and "negative excited".

				1	**2**	**3**
(a) Naive Bayesian	f1	0.549	**1**	0.772	0.165	0.063
	Accuracy	0.559	**2**	0.359	0.364	0.277
			3	0.238	0.218	0.544
(b) Bayesian with time				**1**	**2**	**3**
	f1	0.565	**1**	0.757	0.165	0.078
	Accuracy	0.573	**2**	0.350	0.403	0.247
			3	0.248	0.228	0.524
(c) Bayesian + genre				**1**	**2**	**3**
	f1	0.598	**1**	0.874	0.019	0.107
	Accuracy	0.613	**2**	0.432	0.354	0.214
			3	0.359	0.029	0.612
(d) Bayesian + genre + time				**1**	**2**	**3**
	f1	0.634	**1**	0.830	0.087	0.082
	Accuracy	0.639	**2**	0.315	0.486	0.199
			3	0.345	0.053	0.602
(e) SVM linear kernel				**1**	**2**	**3**
	f1	0.564	**1**	0.762	0.133	0.107
	Accuracy	0.558	**2**	0.325	0.345	0.330
			3	0.208	0.223	0.568

The f1 measure increased about 9 percent utilizing both priors in comparison to naïve Bayesian.

As with the temporal prior, the genre prior leads to better estimate of the emotion class. The classification accuracies of the first class "calm" and the third class, "negative excited" have been improved with this prior. Regarding< the "calm" class, the reason for this improvement is that genre has a clear impact on arousal, thus on "calm" vs. "aroused" classification (horror and action movies have higher arousal than drama and comedy). The determination of the second class, "positive excited", was only improved by utilizing the temporal prior and not by the genre prior. The reason is that positive excited emotions were spread among different genres in this training data. A sample of movie scenes classification along time is shown in Figure 5. The evolution of classification results over consecutive scenes when adding the time prior shows that this prior allows correcting results for some samples that were misclassified using the genre prior only. For example on

the 4th and 20th scene the classifier with time prior was able to find the correct class while the naïve Bayesian with only genre prior missed it. Moreover adding of the time prior did not change any correct classification of the naïve Bayesian classifier.

One of the main drawbacks of the proposed approach is the low temporal resolution of affective annotations. It is impossible to guarantee a perfect continuous assessment and annotation of a movie without the user being distracted at times from the movie events. It is also non-realistic to expect an average user to be able to use psychological terms and consistent words to express his/her emotions. Using physiological signals or audio-visual recordings will help overcome these problems and facilitate this part of the work, by yielding continuous affective annotations without interrupting the user [7].

5. Conclusions and perspectives

An affective representation system for estimating felt emotions at the scene level has been proposed using a Bayesian classification framework that allows taking some form of context into account. Results showed the advantage of using well chosen priors, such as temporal information provided by the previous scene emotion, and movie genre. The *f1* classification measure of 54.9% that was obtained on three emotional classes with a naïve Bayesian classifier was improved to 56.5% and 59.5 using only the time and genre prior. This measure finally improved to 63.4% after utilizing all the priors.

More prior information and semantic analysis of the movie's script (subtitles), as well as higher level features are necessary to further improve affect representation. Priors can be personality related information that help in the definition of personal affective profiles. An example of such pieces of information is social network groups indicating people with the same taste, gender, ethnicity, age, etc. A larger dataset of movies with annotations from multiple participants with different backgrounds will therefore enable us to examine more priors. It will also provide us with a better understanding of the feasibility of using group-wise profiles containing some affective characteristics that are shared between users.

6. Acknowledgment

The research leading to these results has received funding from the Swiss national science foundation and the European Community's Seventh Framework Program [FP7/2007-2011] under grant agreement Petamedia n° 216444.

References

[1]"YouTube serves up 100 million videos a day online," http://www.usatoday.com/tech/news/2006-07-16-youtube-views_x.htm, 2006.

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

[3] C.H. Chan and G.J.F.Jones, "Affect-based indexing and retrieval of films,", MM '05: Proc. of the 13th ACM Mult., pp. 427-430, 2005.

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

[5] A. Hanjalic, "Extracting moods from pictures and sounds," *IEEE Signal Proc. Mag.*, vol. 23, no. 2, pp. 90-100, Mar.2006.

[6] M. Xu, J.S.Jin, S. Luo, and L. Duan, "Hierarchical movie affective content analysis based on arousal and valence features,", MM '08: Proc. of the 16th ACM Mult., pp. 677-680, 2008.

[7] M. Soleymani, G. Chanel, J. J. M. Kierkels, and T. un, "Affective Characterization of Movie Scenes Based on Multimedia Content Analysis and User's Physiological Emotional Responses,", ISM '08: Proc. of the 10th IEEE Int. Symp. on Mult., pp. 228-235, 2008.

[8] R. Cowie, E. Douglas-cowie, S. Savvidou, E. Mcmahon, M. Sawey, and M. Schröder, "feeltrace: an instrument for recording perceived emotion in real time,", Proc. of the ISCA . Northern Ireland.: pp. 19-24, 2000.

[9] M. Xu, Liang-Tien Chia, Haoran Yi, and D. Rajan, "Affective content detection in sitcom using subtitle and audio,", Multi-Media Modelling Conf. Proc., 2006.

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

[11] D. Bordwell and K. Thompson, *Film art, an introduction* Addison-Wesley publishing company, 1980.

[12] M. E. Tipping, "Sparse Bayesian learning and the relevance vector machine," *J. of Machine Learning Research*, vol. 1, no. 3, pp. 211-244, 2001.

[13] P. J. Lang, M. K. Greenwald, M. M. Bradley, and A. O. Hamm, "Looking at Pictures - Affective, Facial, Visceral, and Behavioral Reactions," *Psychophysiology*, vol. 30, no. 3, pp. 261-273, 1993.

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

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

[16] B. Janvier, E. Bruno, T. Pun, and S. Marchand-Maillet, "Information-theoretic temporal segmentation of video and applications: multiscale keyframes selection and shot boundaries detection," *Mult. Tools and Applications*, vol. 30, no. 3, pp. 273-288, Sept.2006.

[17] P. Valdez and A. Mehrabian, "Effects of Color on Emotions," *J. of Experimental Psychology-General*, vol. 123, no. 4, pp. 394-409, Dec.1994.

[18] N. Moënne-Loccoz, "OVAL: an object-based video access library to facilitate the developement of content-based video retrieval systems," Viper group, CVML, University of Geneva,03.04, Oct.2004.

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

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

[21] L. Lu, H. Jiang, and H.J. Zhang, "A Robust Audio Classification and Segmentation Method,", MM '01: Proc. of the 9th ACM Mult., pp. 203-211, 2001.

[22] P. Boersma and D. Weenink, "Praat: doing phonetics by

computer," 2008.
[23] Z. Rasheed, Y. Sheikh, and M. Shah, "On the use of computable features for film classification," *IEEE Trans. on Circuits and Systems for Video Technology*, vol. 15, no. 1, pp. 52-64, Jan.2005.
[24] C. Lei, S. Gunduz, and M. T. Ozsu, "Mixed Type Audio Classification with Support Vector Machine,", IEEE Int. Conf. on Mult. and Expo, pp. 781-784, 2006.
[25] J. D. Morris, "SAM:The Self-Assessment Manikin, An Efficient Cross-Cultural Measurement of Emotional Response," *J. of Advertising Research*, 1995.
[26] R.O. Duda, P.E. Hart, and D.G.Stork, *Pattern Classification* Wiley Interscience, 2001.