UTILIZING AFFECTIVE ANALYSIS FOR EFFICIENT MOVIE BROWSING

Shiliang Zhang[1,2]*, Qi Tian*[1]*, Qingming Huang*[3]*, Wen Gao*[2]*, Shipeng Li*[1]

[1]Microsoft Research Asia, No. 49, Zhichun Road, Beijing, 100190, China
[2]Key Lab of Intelli. Info. Process., Inst. of Comput. Tech., CAS, Beijing 100190, China
[3]Graduate University of Chinese Academy of Sciences, Beijing, 100049, China
{slzhang, qmhuang, wgao}@jdl.ac.cn, {qitian, spli}@microsoft.com

ABSTRACT

Because of the fast increasing number of movies and long time span each movie lasts, novel methods should be developed to help users browse movies and find their desired clips effectively. Affective information in movies is closely related with users' experiences and preferences. Therefore, in this paper, we analyze the affective states of movies and propose affective information based movie browsing. Affective movie content analysis is challenging due to the great variety of movie contents and styles. To address this challenge, we first extract rich audio-visual features. Then, feature selection and affective modeling are carried out to select and map effective features into corresponding affective states. Finally, we propose novel Affective Visualization techniques which intuitively visualize affective states to achieve efficient and user-friendly movie browsing. Experiments on representative movie dataset demonstrate the effectiveness of our proposed methods.

Index Terms— Affective Video Content Analysis, Affective Visualization, Dimensional Affective Model

1. INTRODUCTION

Statistics show that more than 4500 new movies are produced around the world each year[1]. The fast increasing speed of movies and their important roles in human daily life have posed new challenges for video content analysis research. Affective information in videos is closely related with audiences' experiences, feelings, and habits. Thus, as a new technique, affective video content analysis has great potential to be a more friendly and intelligent technique. In recent years, works have been reported on music, MTV, and movie affective analysis [2-10]. Different from music and MTV affective analysis, movie affective analysis is more challenging due to the great variances of audio-visual contents and complicated shooting styles in movies. However, affective information such as happy, excited, *etc.* is not only closely related with audiences' experiences but also an important factor considered during film-making. Hence, affective information in movies is significant and movie affective analysis is a valuable research topic.

Due to the fast increasing number of movies as well as their long time span, fast and efficient movie browsing methods have become urgent demands. This paper addresses this problem by proposing a novel affective information based movie browsing. As shown in Fig. 1, because of the varying affective states, a movie is first segmented into video clips with length of 14 seconds and each with 2 seconds overlap between the previous and next clips. After that, rich affective features are extracted from each segment. Effective features are then selected to describe the two components of affective state: Arousal and Valence [11], respectively. Based on the effective features, as well as the affective models, the proposed method is capable of identifying different emotions in movies. Finally, novel Affective Visualization techniques are proposed to convert abstract affective states into intuitive forms. Therefore, the affective states of each movie could be intuitively visualized, and users could browse and navigate to their interested movie clips conveniently.

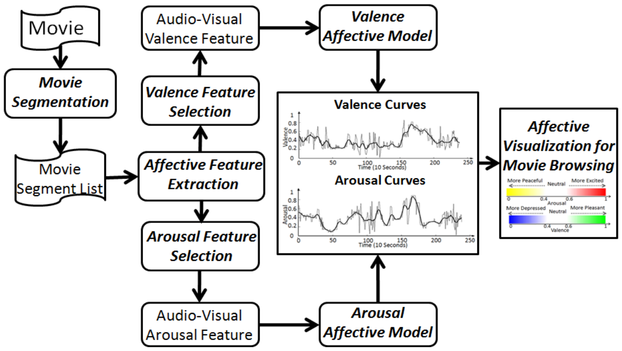


Fig.1. The framework of proposed method

The rest of the paper is organized as follows. Section 2 introduces the related work. Section 3 describes our proposed framework. Experiments and results are presented in Section 4. Section 5 concludes this paper.

2. RELATED WORK

The existing methods on affective video content analysis can be summarized into two categories: categorical affective content analysis and dimensional affective content analysis.

In categorical affective content analysis, emotions are commonly discrete and belong to one of a few basic groups such as "fear", "anger", *etc.* and classifiers are commonly applied for affective analysis. For example, Kang [3] trains two Hidden Markov Models (HMM) to detect affective states including “fear”, “sadness” and “joy” in movies. The recent work of Xu [8] utilizes fuzzy clustering and HMM to classify video shots in movies into 5 emotion types including fear, anger, sad, happy, and neutral.

Dimensional affective content analysis commonly employs the Dimensional Affective Model [11] for affective state representation and modeling. In this model, human affective responses are represented using two components: Arousal (A) and Valence (V). Because different A-V combinations stand for different affective states, this model can represent rich affective states. One representative work is reported by Hanjalic and Xu [2, 4]. Modeling A and V using linear feature combinations, they can obtain the affective sates of different parts of the video. Another representative work is reported by Arifin [7]. Their work employs more complicated affective models and recognizes 6 affective states from the computed affective components.

Because the predefined affective categories are fixed, the flexibility and descriptive power of categorical affective analysis are limited. Thus, categorical affective analysis is not the optimal solution for movie affective modeling. Due to the promising advantages of the A-V model, it is utilized in our work. Section 3 will introduce our affective modeling and Affective Visualization in detail.

3. THE PROPOSED FRAMEWORK

3.1. Affective Feature Extraction

Table 1. Extracted affective features

Audio Intensity Features	Zero Crossing Rate (ZCR), Short Time Energy (STE), ZCR Standard Deviation(STD), STE STD
Audio Timbre Features	Pitch, Bandwidth, Brightness, Roll Off, Spectral Flux, Sub-band Peak, Sub-band Valley, Sub-band Contrast, Pitch STD, Bandwidth STD, Brightness STD
Audio Rhythm Features	Average Tempo, Rhythm Strength, Rhythm Contrast, Rhythm Regularity, Average Onset Frequency, Average Drum Amplitude, Tempo STD
Visual Features	Motion Intensity, Shot Switch Rate, Frame Brightness, Frame Saturation, Color Energy

According to the Table 1, 27 audio and visual features are extracted from each movie segment. The *Audio Intensity Features* which typically describe the intensity of the audio in the movie clip are closely related with the Arousal component and has been frequently utilized in audio affective analysis. The detailed implementations of intensity feature extraction can be found in [12]. The *Audio Timbre Features* are frequently used to distinguish different types of sound production, such as voices or musical instruments and have been proved important in distinguishing different moods [13]. Besides the 7 timber features in [13], we also extract *Pitch*, *Pitch STD*, *Bandwidth STD*, and *Brightness STD* to increase the descriptive power of timber features. Rhythm is an important tool used by artists to express their emotions. Therefore, 7 *Audio Rhythm Features* are utilized. Rhythm Feature extraction is based on Onset detection and Drum detection. The detailed implementations of rhythm feature can be found in [13]. Because cinematography commonly utilizes visual effects to render the emotions as well as to evoke certain moods in audience, the visual clues in movies are important in describing the affective states. The 5 visual features listed in Table 1 are frequently used in related work and have been proved effective [2, 4, 6-10]. After the feature extraction, Gaussian Normalization is employed to normalize these features into [0, 1].

3.2. Affective Modeling

3.2.1. Affective Feature Selection

Because of noises, complicated contents and various shooting styles, the commonly used affective features might be invalid in describing the movie emotions. Thus, as an important contribution of our work, affective feature selection is carried out to compare the performance of different features and select reliable ones for movie affective computation. This is done by running affective analysis on each individual feature, and selecting the most effective ones. The details are given in Algorithm 1.

Algorithm1: Affective feature selection

```
Input: feature array $(Fea_1^{(j)} ... Fea_N^{(j)})$ and ground truth array $G^{(j)}$,
$j=1...M$ of $M$ movie segments; Threshold $T$ for effective
feature selection.
Output: effective feature array $FeaArray$
Suppose: Model_n( $Fea_n^{(j)}$ ) is a model that maps the value of
feature $Fea_n^{(j)}$ to the corresponding Arousal or Valence values.
For $i = 1 : N$   do   $Correct_i=0$
    For $j = 1 : M$   do
        If $\left| G^{(j)} - Model_i \left( Fea_i^{(j)} \right) \right| \leq 0.125$   do
            $Correct_i$ ++
        End
    End
    $Validity_i = Correct_i / M$
    If $Validity_i \geq T$ do Add $Fea_i$ to $FeaArray$
    End
End
```

During our feature selection, Support Vector Regression Model (SVR) with RBF kernel is used as $\mathrm{Model}_n(Fea_n^{(j)})$. The threshold T is experimentally set as 0.35. Ground truths of the affective states are divided into two parts for affective model training and feature selection, respectively. The results of feature selection and details of ground truth acquisition will be presented in the Section 4.

3.2.2. Affective State Computation

After feature selection, the selected features will be mapped into the corresponding affective states. This could be solved as a regression problem, in which input is a multi-dimensional vector and output is a value between 0 and 1. Various regression methods could be applied for this. In our experiment, SVR model is utilized because besides its fast implementation, it provides better prediction on unseen data and better solution for training problems [14]. Two SVR models (Arousal model and Valence model) are trained independently with the selected features, the ground truths, as well as the cross validation parameter selection. The emotions of movie clips can be computed by feeding their features into the trained affective models. The computed affective state sequences are then convolved by a smoothing Gaussian window. Fig. 2 illustrates the computed Arousal curve of an example movie before and after smoothing. The Valence curve is not given due to the page limit.

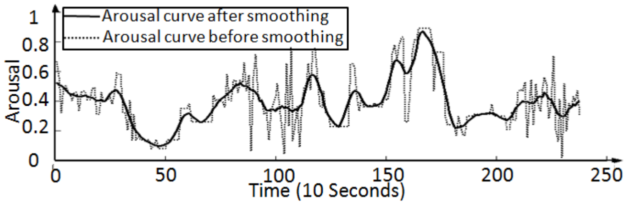
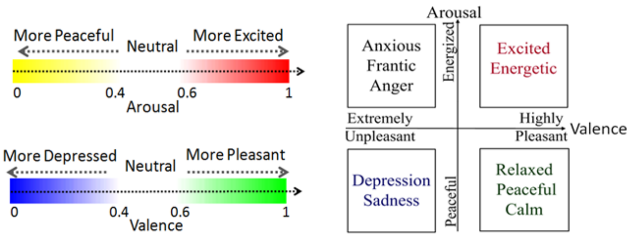


Fig. 2. The illustration of the computed Arousal values

3.3. Affective Visualization



(a) Affective Color-bar (b) Affective Space

Fig. 3. Affective color-bar and affective space

The computed affective state of each movie segment is denoted as a 2-D vector (A, V) which is in general too abstract to understand for the users. Therefore, two novel methods are developed to visualize the affective states.

Firstly, in order to let users browse the movie content as well as navigate to the interested segments conveniently, we visualize the affective states of movies on time axis. Color, the commonly used sign of emotion is employed to represent affective states. A color-bar (Fig. 3-(a)) is shown to explain the affective states of different colors.

Secondly, as different affective regions represent different affective states (Fig. 3-(b)), the emotions of movie clips can be intuitively represented in the affective space. Meanwhile, movie clip retrieval can be simplified by choosing the corresponding regions. Hence, another visualization method is proposed to map the movie clips onto the affective space. Instances and illustrations of Affective Visualization will be presented in Section 4.

4. EXPERIMETNS AND RESULTS

4.1. Movie Database and Ground Truth

Representativeness is one of the primary considerations when we collect the movie database. Our database contains 13 movies (*Godfather, Titanic, etc.*) with total duration of about 30 hours and different film types including action, drama, *etc.* As far as we know, our database is currently one of the most representative movie databases [2, 4, 6-8].

We collect affective ground truths from user study. Several users are invited to accomplish this task. They are required to choose A and V values (between [0, 1]) to describe the movie segments' affective states which they feel confident of. In this way, we manage to obtain the ground truths of about 4000 movie segments.

4.2. Results of Feature Selection

Each feature's *Validity* values in the feature selection are presented in Fig. 4. Accordingly, 13 Arousal features (*ZCR, STE, Sub-band Peak, Sub-band Valley, Sub-band Contrast, Tempo, Rhythm Strength, Rhythm Contrast, Rhythm Regularity, Drum Amplitude, Motion Intensity, Short Switch Rate, Frame Brightness*) and 9 Valence features (*Pitch, Sub-band Peak, Sub-band Valley, Sub-band Contrast, Pitch STD, Rhythm Regularity, Frame Brightness, Saturation, Color Energy*) are selected out of 27 features. It is also necessary to point out that our result and selected features can be important references for further researchers.

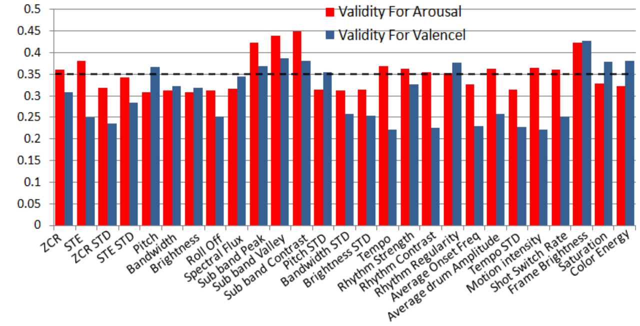


Fig. 4. The result of feature selection

4.3. Evaluation of Affective Modeling

4.3.1 Evaluation on the Entire Movie Database

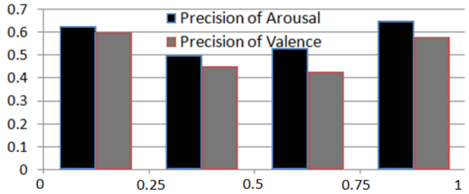


Fig. 5. The performance of affective analysis

To evaluate the performance of our affective analysis on different affective states, the A and V *Precision* values are computed with Eq. (1) on 4 ground truth sets, respectively (four movie segment sets with A or V ground truth values between [0, 0.25], [0.25, 0.5], [0.5, 0.75], and [0.75, 1]).

$$Precision^{(K)} = \frac{\sum_{i=1}^{M^{(K)}} \mathrm{F}(|GTruth_i^{(K)} - Val_i|)}{M^{(K)}}, \quad \begin{cases} \mathrm{F}(v) = 0 & if: v > 0.125 \\ \mathrm{F}(v) = 1 & if: v \le 0.125 \end{cases} \quad (1)$$

where, $M^{(k)}$ is the number of movie segments in ground truth set K, *GTruth* and *Val* denote the ground truth value and corresponding computed value, respectively.

It is clear from Fig. 5 that, the A and V precisions in the ground truth sets [0.25, 0.5] and [0.5, 0.75] are worse than the ones in the other two sets. This is because the affective states in these two sets are relatively obscure, which makes manual annotations and computer models less accurate. The precisions of Arousal are generally better than the ones of Valence. Besides the different performance of A and V features, this result is also related to the fact that users have more consistent opinions on Arousal than Valence. Considering the representativeness of our database and the precision of worse case (which is 0.25), we can conclude that the performance of our algorithm is still reasonably good.

4.3.2 Evaluation on Different Movie Types

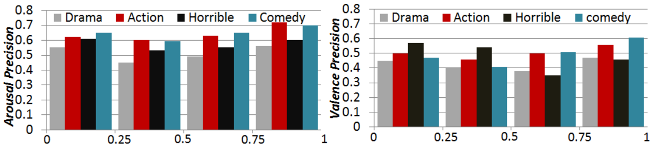
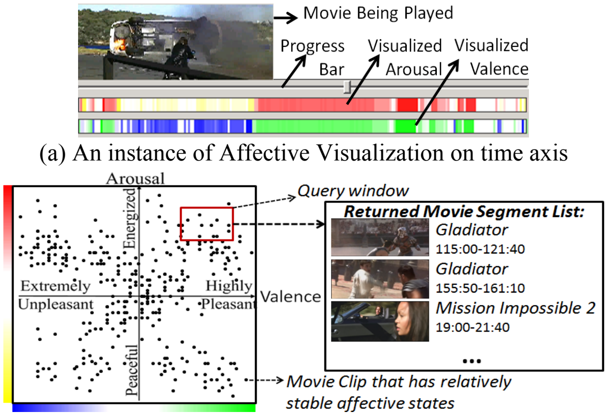


Fig. 6. The comparisons of A-V precisions

In order to test our affective analysis on different kinds of movies, four representative movie types are selected: Drama movies (*Godfather* Ⅰ, *American Beauty*), Horror movies (*Mirrors*), Comedy movies (*Angus Thongs and Perfect Snogging*), and Action movies (*Mission Impossible* Ⅰ *and* Ⅱ). The experimental results are presented in Fig. 6.

From Fig. 6, it is noticeable that the A and V precisions of Drama movies are lower than the others. This is because the emotions in drama movies are more likely expressed by conversations between characters and are difficult to be captured by the extracted features. While for the other three kinds, because their emotions are commonly expressed by music or certain audio-visual effects, our method performs better. Still, we believe further work is needed to make movie affective analysis accurate for different film types.

4.4 The Result of Affective Visualization



(a) An instance of Affective Visualization on time axis

(b) An instance of Affective Visualization in Affective Space

Fig. 7. Illustrations of Affective Visualization

Some instances of Affective Visualization are presented in Fig.7. In Fig. 7-(a) the affective information of a movie is visualized on time axis with different colors. Thus, users can intuitively sense the affective states of movie segments and conveniently move to their interested parts. In Fig. 7-(b), movie clips in our database with relatively stable emotions are visualized as points in the affective space. User can effectively browse and retrieve movies by selecting different regions. This visualization, although loses the temporal information in each movie, is convenient for movie database browsing and movie clip retrieval. Consequently, the proposed Affective Visualization techniques are valid and efficient for movie browsing.

5. CONCLUSIONS

In this paper, we propose a framework for movie affective analysis and browsing. From 27 commonly used affective features, we have identified 22 effective ones for movie affective computation. SVR based models are built to map the affective features into affective states. Two novel Affective Visualization techniques are proposed to achieve efficient and convenient movie browsing. Numerical experiments on representative movie dataset illustrate the validity of our proposed methods. In the future work, we will finish a movie browsing prototype system based on the proposed techniques. Meanwhile, further work will be done to improve the accuracy of movie affective content analysis.

6. REFERENCES

[1] C. Weng, W. Chu, and J. Wu, "RoleNet: Movie Analysis from the Perspective of Social Networks," *IEEE Transactions on MM*, 11(2):256-271, Feb. 2009.

[2] A. Hanjalic, "Extracting Moods from Pictures and Sounds: Towards Truly Personalized TV," *IEEE Signal Processing Magazine*, 23(2):90-100, Mar. 2006.

[3] H. Kang, "Affective Content Detection Using HMMs," *ACM MM*, pp. 259-262, 2003.

[4] A. Hanjalic and L. Xu, "Affective Video Content Representation and Modeling," *IEEE Transactions on MM*, 7(1):143-154, Feb. 2005.

[5] L. Lu, D. Liu and H. Zhang, "Automatic Mood Detection and Tracking of Music Audio Signals," *IEEE Transactions on Audio and Language Processing*, 14(1):5-18, Jan. 2006.

[6] H. Wang and L. Cheong, "Affective Understanding in Film," *IEEE Transactions on CSVT*, 16(6):689-704, 2006.

[7] S. Arifin and P. Y. K. Cheung, "A Computation Method for Video Segmentation Utilizing the Pleasure-Arousal-Dominance Emotional Information," *ACM MM*, pp. 68-77, 2007.

[8] M. Xu, J. Jin, and S. Luo, "Hierarchical Movie Affective Content Analysis Based on Arousal and Valence Features," *ACM MM*, pp.677-680, 2008.

[9] S. Zhang, Q. Tian, S. Jiang, Q. Huang, and W. Gao, "Affective MTV Analysis Based on Arousal and Valence Features," *IEEE ICME*, pp.1369-1372, 2008.

[10] S. Zhang, Q. Huang, Q. Tian, S. Jiang, W. Gao, "Personalized MTV Affective Analysis Based on User Profile", *Pacific-Rim Conference on Multimedia*, pp. 327-337, 2008.

[11] H. Schlosberg, "Three Dimensions of Emotion," *Psychol. Rev*, 61(2):81-88, Mar. 1954.

[12] R. Cai, L. Lu, A. Hanjalic, H. Zhang, and L. Cai, "A Flexible Framework for Key Audio Effects Detection and Auditory Context Inference," *IEEE Transactions on Audio and Language Processing*, 14(3):1026-1039, May. 2006

[13] L. Lu, H. Zhang, and S. Li, "Content-Based Audio Classification and Segmentation by Using Support Vector Machines," *ACM Multimedia Systems Journal*, 8(6):482-492, Mar. 2003.

[14] A. J. Smola and B. Scholkopf, "A Tutorial on Support Vector Regression," *Statistics and Computing*, 14(3):199-222, 2004.