Affective Music Video Content Retrieval Features Based on Songs

R.Hemalatha
Department of Computer Science and Engineering,
Mahendra Institute of Technology,
Mahendhirapuri, Mallasamudram West, Tiruchengode, Tamil Nadu 637503.

A.Anbumani
Assistant Professor,
Department of Computer Science and Engineering,
Mahendra Institute of Technology,
Mahendhirapuri, Mallasamudram West, Tiruchengode, Tamil Nadu 637503.

***Abstract* — Nowadays, MTV has become an important favorite pastime to modern people because of its conciseness, convenience to play and the characteristic that can bring both audio and visual experiences to audiences. An affective MTV analysis framework, which realizes MTV affective state extraction, is representation and clustering. Firstly, affective features are extracted from both audio and visual signals. Affective state of each MTV is modeled with 2D dimensional affective model and visualized in the Arousal-Valence space. Finally the MTVs having similar affective states are clustered into same categories. The validity of proposed framework is proved by subjective user study and related work proves that our features improve the performance by a significant margin.**

Index Terms— Categorical affective content analysis, affective visualization, dimensional affective model, Hidden Markov Models(HMMs).

I. INTRODUCTION

MTV (Music TV) is an important favorite pastime to modern people. Especially in recent years, MTV can be played conveniently on mobile sets including cell phones and music players such as iPod and Zune. Consequently, MTV has become more popular and common than before. The increasing amounts of MTV and storage capacity of our digital sets have caused many problems: how to effectively organize, manage and retrieve the desired MTVs. It is true that the traditional MTV classifications based on Artist, Album and Title, could be solutions to this problem. However, these methods have many limitations when people want to manage and retrieve MTVs with semantic and abstract concepts. Affective MTV content analysis which has little been researched before might provide potential This work was supported in part by National Natural Science Foundation of China, in part by National Hi-Tech Development Program (863 Program) of China under Grant solutions to these problems. For example, users will be able to classify MTVs into categories according to MTVs' affective states, so that they can select their desired categories to enjoy. Up till now, researchers have completed some work on music and movie affective content analysis and applications based on these technologies seem promising. In their work, the Arousal-Valence (A-V) model proposed is used to express affective states in movies. Modeling Arousal and Valence using features' linear combinations, the authors can get Arousal and Valence values of different parts of the movie and draw affective curves in the A-V space. Consequently, the affective states of movie can be visualized. Affective classical music content analysis is represented. They extract three types of audio features to represent mood in music: Intensity, Timber and Rhythm. The authors classify music segments into four mood categories: Contentment, Depression, Exuberance and Anxious/Frantic with a hierarchical framework. Furthermore, since the mood in the classical music is usually varying, the authors extend their method to mood tracking for a music piece which contains a constant mood, by dividing the music into several independent segments. Since computers are employed to identify mood and emotion, the design of computer algorithms and models should more or less base on psychological knowledge. The A-V model is adopted in our work to express the affective state of each MTV.
The affective content analysis technique is compared with the traditional methods, affective information based MV access presents advantages in four aspects:

1) Affective states such as happy or sad, which are closely related to user experiences, could be informative and accurate descriptions for each MV;

2) Affective content analysis can be implemented automatically in large-scale MV databases;
3) Users can retrieve MV sintuitively with abstract concepts without converting them into concrete Titles, Artists, or Styles, e.g., with affective information, users could find energetic MVs even if they do not know which Album contains such MVs; and
4) Affective content analysis and traditional metadata describe MVs from different aspects—therefore, they can be complemented to each other.

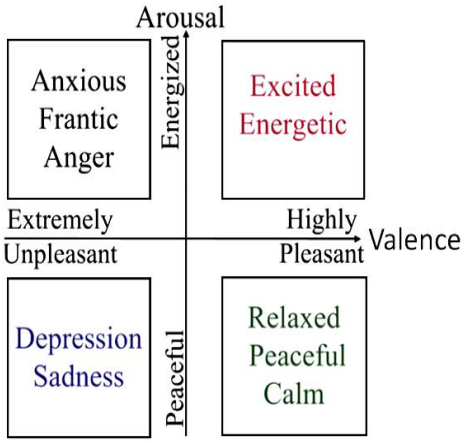


Fig. 1. Illustration of the dimensional affective model.

II. RELATED WORK

As a newly developing research area, a great deal of attention has been paid to affective video and audio content analysis, and many related works have been reported in recent years. Generally, the existing affective content analysis works 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 categories, such as "fear," "anger," or "happy." Consequently, many classifiers are adopted for affective analysis. Precision rates of 63%, 89%, 93%, and 80% are achieved for "startle," "apprehension," "surprise," and "apprehension" to "climax", respectively. Kang trains two Hidden Markov Models(HMMs) to detect affective states including "fear," "sadness," and "joy" in movies. The classification performances of 81.3%, 76.5%, and 78.4% are reported, respectively, for the three affective states. Similarly, a four-state HMM is adopted in the work to classify audio emotional events such as laughing or horror sounds in comedy and horror videos. Hierarchical classification framework is utilized in the work for classical music affective content analysis. They extract three types of audio features to represent the mood in music: intensity, timber, and rhythm. The authors classify music segments into four mood categories: contentment, depression, exuberance, and anxious/frantic with a hierarchical classifier.

Categorical affective content analysis is more suitable to be called as affective classification. Although the flexibility of these methods is limited, they are simple, and easy to build.

Dimensional affective content analysis commonly employs the dimensional affective model for affective state computation. One representative work, the Arousal-Valence (A-V) model is used to express affective states in videos. Modeling Arousal and Valence using linear feature combinations, the authors can obtain the Arousal and Valence values of different video segments and draw affective curves in the A-V space.

III. THE PROPOSED SYSTEM

Since affective features are the important basis for our affective modeling, we proceed to introduce the affective features applied in our work. Generally, the visual contents of MVs are carefully designed by artists to coordinate with the music. Consequently, both the audio and visual contents are used for affective feature extraction. We extract affective features according to the music theory, cinematography studies and related work on affective analysis.

A. Arousal Feature Extraction

Arousal represents the intensity of affective states. The features extracted for Arousal include: motion intensity, short switch rate, zero crossing rate, tempo, and beat strength.

Motion Intensity: Motion intensity which reflects the smoothness of transition between frames is a commonly used Arousal feature. In our work, motion intensity is acquired based on the motion vectors extracted from

MPEG streams. Motion intensity between frames is first computed. Then, the final motion intensity MI is computed and normalized between 0 and 1 with

$$MI' = ([\frac{\overline{MI} - \mu MI}{3\sigma MI}] + 1)/2$$

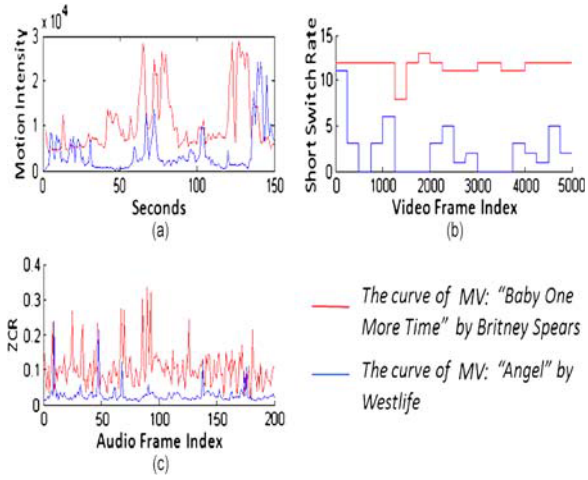


Fig. 2. Comparison of motion intensity, shot switch rate, and ZCR between two MVs.

Where MI indicates the mean of inter-frame motion intensity within each MV before normalization, U_{MI} and σ_{MI} denote the mean and standard deviation of motion intensity computed with all MVs in the database. The same normalization is applied to other features. Equation is the standard form of Gaussian normalization. It is utilized to format the extracted features into more reasonable distributions. Note that U_{MI} and σ_{MI} can be updated when the MV database is changed. The Arousal component is calculated with the linear combinations of these features, which is explained in Fig.3.

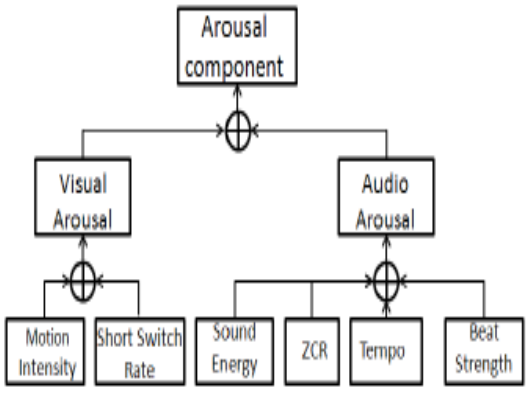


Fig 3. Arousal Components.

Shot Switch Rate: Shot is a powerful tool for directors to control the tempo of videos. The duration of shot is used as an important feature describing Arousal in related work. We take the normalized average shot switch number within each video segment with length of 250 frames as the shot switch rate feature (SSR).

Zero Crossing Rate (ZCR): ZCR is a widely used feature in audio signal analysis. It could be utilized to distinguish harmonic music, speeches and environmental sounds. Our experiments show that intense music generally presents higher ZCR value than the smooth music.

Rhythm Based Features, Tempo and Beat Strength: Rhythm is an important characteristic of music, and artists frequently employ rhythm to express their emotions. In general, three aspects of rhythm are closely related with human affective experience: tempo, beat strength, and rhythm regularity. In this paper, we extract the rhythm-related features based on the music onset detection.

ICA: Independent Component Analysis is a computational method for separating a multivariate signal into additive subcomponents then it's divided into sub categories.
Onset detection: onset in music is caused by the instruments like drums which produce high energy and salient sounds . In our onset detection, the audio signal is first divided into five sub bands. Each is then smoothed with a Gaussian window. Since onsets are usually shown as energy peaks, we set a threshold to filter the signal with low energy to zero. Then, the local maximums of the signal can be detected as onsets. After computing the first derivative of the audio signal, we confirm the positions of onsets by searching the zero crossing points in the falling edges of the signal.

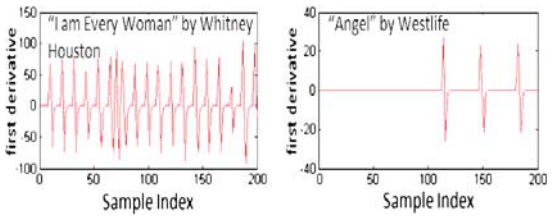


Fig. 4. Comparison between two audio signals' tempo.

Tempo and beat strength calculation: the average onset number within 5 seconds of audio signal is taken as the tempo of the audio.

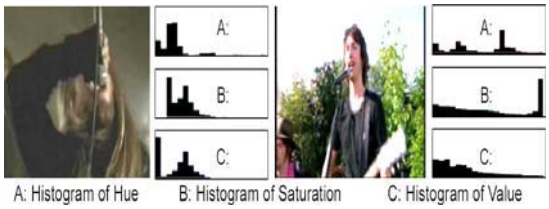


Fig. 5. Comparisons of H, S, and V histograms between two MVs, (a) Histogram of hue, (b) Histogram of saturation, (c) Histogram of value.

Beat strength is obtained by averaging the audio energy at the onsets' positions. Tempo and beat strength on each sub band are fused with linear combinations and then normalized to get the final tempo and beat strength features.

B. Valence Feature Extraction
Valence represents the type of affective state. Valance feature lighting, rhythm regularity, saturation, color energy, and pitch are extracted to describe Valence in fig 6.
Lighting: In cinematography, one of the powerful tools is used specifically for the purpose of affecting viewers' emotions and establishing the mood of a scene. An abundance of bright illumination is frequently used to generate the lighthearted atmosphere while dim illumination is frequently selected for the opposite purposes. With the histogram containing N bins on V component of hue, saturation, and value (HSV) color space, we calculate the lighting feature L with

$$L^k = \sum_{j=M+1}^{N} V_j \cdot (j - M) - \sum_{j=1}^{M} V_j \cdot (M + 1 - j)$$

where j denotes the index of the histogram bin, V_j is the value of the jth bin of the Value histogram, and stands for the frame index. In the lower M bins represent the dark lighting components and the rests denote the bright ones. N is experimentally set to 20. The video frames in MV are commonly dark, which makes the sum of the lower ten bins generally larger than the sum of the upper ten. Consequently, we set M to 7 rather than 10. The final lighting feature L is obtained by computing and normalizing the mean of L^k.

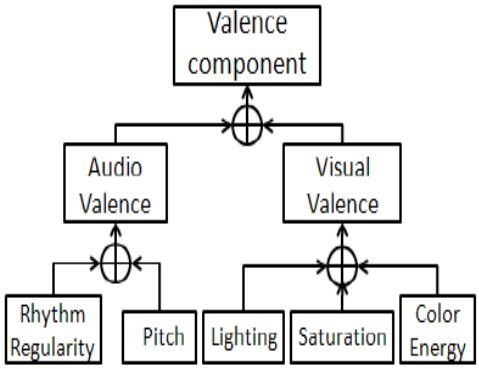


Fig 6. Valence Components

Saturation: It is well known that sad or frightening videos commonly present gray frames which indicate low color saturation. The saturation feature S is calculated in

$$S^k = \sum_{j=M+1}^{N} S_j . (j - M) - \sum_{j=1}^{M} S_j . (M+1-j)$$

Where S_j denotes the value of the jth bin of the Saturation histogram. N and M are experimentally set to 20 and 10, respectively.
The final Saturation feature S is calculated with the same method for computing the Lighting feature.
Color Energy: In most joyous videos, the video frames are colorful and bright, while in sad or frightening videos the video colors are faded and gray. The colorfulness of a video is measured by color energy computed with

$$CEng^k = \sum_{j=1}^{PixelNum} S_j V_j / std(Hist_H). PixelNum$$

Where PixelNum is the number of pixels in the video frame k, S_j and V_j denote the values of the S and V color component of the pixel, respectively. $Std(Hist_H)$ returns the standard deviation of the Hue histogram. The final color energy feature CEng is acquired by computing and normalizing the mean of $CEng^k$. Fig. 6 illustrates the H, S, and V histograms of video frames from two MVs (a frantic MV and a pleasing MV). From the comparisons, it is clear that the above-mentioned three features would be valid for describing the Valence component.
Rhythm Regularity: Regular rhythm is an obvious characteristic of joyous music. The rhythm regularity feature is calculated based on the regularity of onsets' intervals which can be obtained in the onset detection process.
Pitch: Pitch is a popular audio feature for Valence. We compute pitch on each audio frame and then utilize formula similar as to obtain the final pitch feature.
Texture features: Texture Feature analysis of textures requires the definition for a local neighborhood corresponding to the basic texture pattern. Some features are used in texture feature are

1) Entropy
2) Contrast
3) Inverse Difference

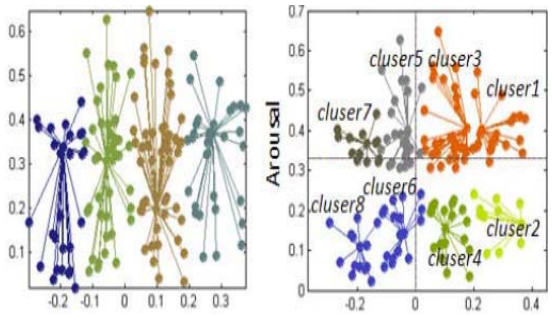
IV. EXPERIMENTS AND RESULTS

4.1. The MTV Dataset

We collected 156 English pop MTVs of MPEG format through two ways: downloading from Internet and converting from DVDs. These MTVs are recorded in different periods and have different resolutions and qualities. Consequently, our dataset of MTV is representative. The 156 MTVs do not include the Live MTVs which scarcely include information about shooting styles.

4.2. Results of Affective Clustering

First, the Valence-based clustering (first-step) is carried out and four clusters are generated. Then, the Arousal-based clustering (second-step) is carried out on these four pregenerated clusters respectively.



(a) First-step clustering (b) Second-step clustering
Fig.7. The results of Affective Clustering

Finally, eight categories, within which the MTVs are similar in both Arousal and Valence, are generated. The results of the first-step and second-step clustering are illustrated in Fig. 7. Each category is marked in the *A-V* space and colored, so user can easily identify each category's affective state and select their desired categories to enjoy from MTV database.

V. CONCLUSIONS

In this paper, we present a framework for affective MTV analysis. Thayer's dimensional affective model is adopted. Six Arousal features and five Valence features are extracted. After affective state extraction, Affinity Propagation is utilized to put MTVs with similar affective states into same categories. Finally, 7 affective categories are generated and visualized in the *A-V* space. We conduct subjective user study to obtain the background truth about the affective state of each MTV. The numerical evaluations prove the validity of our framework. The comparisons between our selected features and those in related work verify that our features improve the performance by a significant number. In the future work, a user interface will be finished. Besides that, we will improve our work through researching new ways for affective feature extraction and Arousal Valence modeling.

REFERENCES

[1] M. Xu, L. T. Chia, and J. Jin, "Affective content analysis in comedy and horror videos by audio emotional event detection", in Proc. IEEE ICME, pp. 622–625, 2005.
[2] S. L. Zhang, Q. Tian, S. Q. Jiang, Q. M. Huang, and W. Gao, "Affective MTV analysis based on arousal and valence features", in Proc. IEEE ICME, pp. 1369–1372, 2009.
[3] H. B. Kang, "Emotional event detection using relevance feedback", in Proc. IEEE ICIP, pp. 721–724, 2003.
[4] S. L. Zhang, Q. Huang, Q. Tian, S. Jiang, and W. Gao, "i.MTV—An integrated system for MTV affective analysis", ACM Multimedia, pp. 985–986, 2008.
[5] M. Xu, S. Luo, and J. Jin**, "**Video adaptation based on affective content with MPEG-21 DIA framework", in Proc. IEEE SCIISP, , pp. 386–390, 2007.
[6] X. L. Liu, T. Mei, X. S. Hua, B. Yang, and H. Q. Zhou, "Video collage", ACM Multimedia, pp. 461–462, 2007.
[7] T. Zhang and C. C. J. Kuo, "Audio content analysis for online audiovisual data segmentation and classification", IEEE Trans. Audio, Speech ,Language Process., vol. 9, no. 4, pp. 441–457, 2001.
[8] N. C. Maddage, C. S. Xu, M. S. Kankanhalli, and X. Shao, "Content based music structure analysis with applications to music semantics understanding", ACM Multimedia, pp. 112–119, 2004.
[9] L. Agnihotri, J. Kender,N. Dimitrova, and J. Zimmerman, "Framework for personalized multimedia summarization", in Proc. ACM Int. Workshop MIR, pp. 31–38, 2005.
[10] L. Rabiner and B. H. Juang, "Fundamentals of Speech Recognition. Englewood Cliffs", NJ: Prentice-Hall, 1993.
[11] P. Kelm, S. Schmiedeke, and T. Sikora, "Feature-Based Video Key Frame Extraction for Low Quality Video Sequences", Proc. Int'l Workshop Image Analysis for Multimedia Interactive Services, pp. 25-28, 2009.
[12] Z. Rasheed, Y. Sheikh, and M. Shah, "On the Use of Computable Features for Film Classification", IEEE Trans. Circuits and Systems for Video Technology, vol. 15, no. 1, pp. 52-64, 2005.
[13] P. Valdez and A. Mehrabian, "Effects of Color on Emotions", J. Experimental Psychology, vol. 123, no. 4, pp. 394-409, 1997.
[14] L. Lu, D. Liu, and H. J. Zhang, "Automatic mood detection and tracking of music audio signals", IEEE Trans. Audio, Speech, Language Process., vol. 14, no. 1, pp. 5–18, 2006.
[15] G. P. Nguyen and M. Worring, "Optimization of interactive visual similarity based search", ACM Trans. Multimedia Comput., Commun., Appl., vol. 4, no. 1, 2007.
[16] G. Tzanetakis and P. Cook, "Musical genre classification of audio signals", IEEE Trans. Speech Audio Process., vol. 10, no. 5, pp. 293–302, 2002.