AFFECTIVE MTV ANALYSIS BASED ON AROUSAL AND VALENCE FEATURES

Shiliang Zhang[1, 2], *Qi Tian*[3], *Shuqiang Jiang*[1], *Qingming Huang*[2], *Wen Gao*[1]

[1]Key Lab of Intelligent Information Processing, Chinese Academy of Sciences(CAS) Institute of Computing Technology, CAS, Beijing 100190, China
[2]Graduate University of Chinese Academy of Sciences, Beijing, P.R. China 100080
[3]Department of Computer Science, University of Texas at San Antonio
{slzhang, sqjiang, qmhuang,wgao}@jdl.ac.cn , qitian@cs.utsa.edu

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. In this paper, we propose an affective MTV analysis framework, which realizes MTV affective state extraction, representation and clustering. Firstly, affective features are extracted from both audio and visual signals. Then, the 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. The comparisons between our selected features and those in related work prove that our features improve the performance by a significant margin.

Index Terms— Dimensional Affective Model, Affinity Propagation, Affective Content Analysis

1. 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 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. One representative work about affective movie content analysis is reported by Hanjalic and Xu [1-3]. In their work, the Arousal-Valence (*A-V*) model proposed by Thayer [4-6] 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. In the work of Lu [7], 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. An illustration of this affective model is shown in Fig. 1.

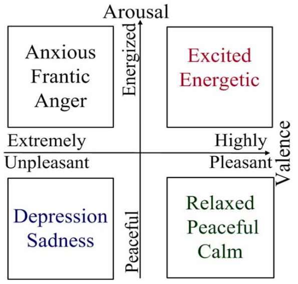


Fig.1. An illustration of the dimensional affective model

This work was supported in part by National Natural Science Foundation of China under Grant 60773136 and 60702035, in part by National Hi-Tech Development Program (863 Program) of China under Grant 2006AA01Z117 and 2006AA01Z315

Compared with related work in affective content analysis field, our contributions can be summarized as follows:

1) This work investigates more *A-V* features than the state-of-the-art work in Hanjalic [1-3] and Lu [7]. Experiments validate that our features are more effective for affective state extraction.
2) We use a novel Affinity Propagation [8] for fast clustering instead of classification. The rationality of our method is proved better.
3) We adopt 2-D *A-V* map for results' visualization. It is convenient for thereafter users' selection and play.
4) This is the first of its kind in affective MTV analysis to the knowledge of the authors. It provides a large potential application for music players.

The rest of paper is organized as follows. Section 2 introduces our framework for affective modeling and affective clustering in detail. The experiments and results are represented in Section 3. Section 4 concludes this paper.

2. THE PROPOSED FRAMEWORK

2.1. Overview of the Framework

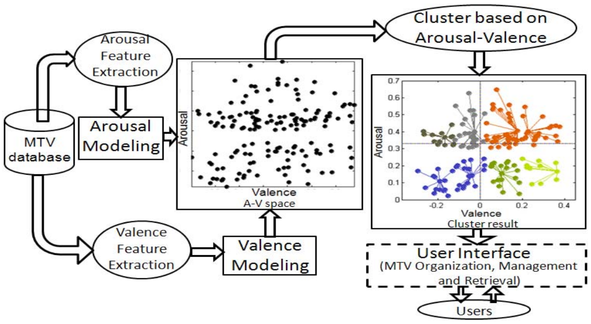


Fig. 2. Our framework

MTV can be simply considered as the combination of music and movie. But this combination is not simple addition or superposition. In most of the cases, the visual content is carefully selected by directors and artists to coordinate with the music. Consequently, both the audio and visual content are employed to extract affective features.

Fig. 2 represents the framework of our affective MTV analysis model. The two affective components Arousal and Valence are modeled and computed respectively. After their computation, the affective states of MTVs can be mapped into the *A-V* space. Then, MTVs similar in affective states are grouped into same categories.

The generated categories are shown in the *A-V* space and marked with different colors. The users can pick their desired categories to enjoy through the user interface.

2.2. Arousal Modeling

Instead of extracting features from the whole segment of each MTV, we utilize a segment length of 50 seconds in the center part of the MTV for feature extraction. Because most of the MTVs express single affective state and the affective climax is usually in the center, our approach is reasonable. Arousal features including *Motion Intensity*, *Shot Switch Rate*, *Sound Energy* [9], *Zero Crossing Rate* [9], *Tempo* [7, 10] and *Beat Strength* [7, 10], are extracted on the selected MTV segment. Then these features are normalized between 0 and 1 and the averages are calculated. The Arousal component is calculated with the linear combinations of these features, which is illustrated in Fig. 3. Due to the space limit, we omit the detailed equations.

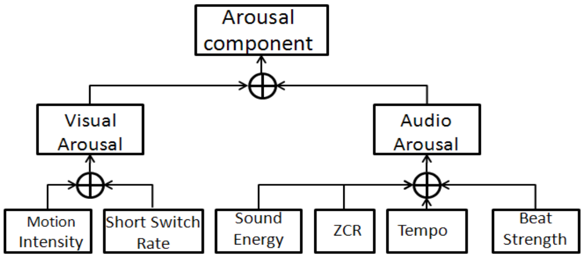


Fig. 3. The framework for Arousal computation

2.3. Valence Modeling

Valence component in the dimensional model represents the type of affective state. The features used for Valence modeling include: *Rhythm Regularity* [10]*, Pitch, Lighting* [11-13]*, Saturation* [12] and *Color Energy* [11].

Similar to the Arousal computation, these features are extracted from the selected segment, then normalized and fused with the framework illustrated in Fig. 4. The final Valence is computed by subtracting "neutral feeling" Valence which serves to map the low (high) values of Valence to the corresponding negative (positive) Valence values [2]. "neutral feeling" Valence is set as 0.45 in our experiment.

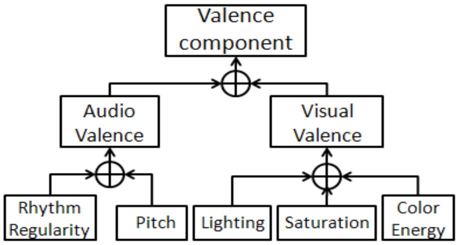


Fig. 4. The framework for Valence computation

2.4. Affective Clustering

In order to put MTVs similar in affective states into same categories, we employ the clustering method: Affinity Propagation (*AP*) [8] rather than traditional classifiers because of the following reasons:

1) It is difficult to know the number of pre-defined affective categories a classifier should output.
2) It is difficult to define the affective state of a MTV, since opinions vary from people to people.
3) The number of clusters produced by *AP* can be adaptively adjusted. This would be useful if we want to control the affective similarity within each cluster.
4) The *AP* presents fast speed for large-scale problems.

AP starts with the construction of a similarity matrix. By viewing each data point as a node in a network, this method recursively transmits real-valued messages along edges of the network until a good set of exemplars and corresponding clusters emerge [8]. Interested reader is referred to [8] for detailed introduction and analysis.

In this work, we utilize a two-step clustering scheme to obtain the final result. This method is illustrated in Fig. 5.

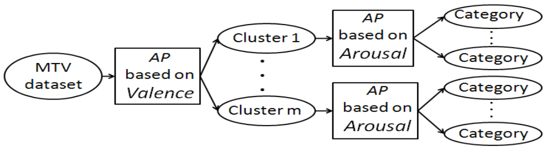


Fig. 5. The process of affective clustering

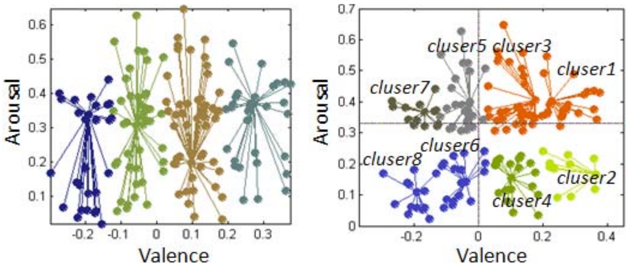
3. EXPERIMENTS AND RESULTS

3.1. 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.

3.2. 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 pre-generated clusters respectively. 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. 6. 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.



(a) First-step clustering (b) Second-step clustering

Fig. 6. The results of Affective Clustering

3.3. Verification of Effectiveness

1) User study for Ground Truth Preparation: There is no objective measure available to evaluate the affective state of the MTV. Consequently, we conduct subjective user study to obtain the background truth. 11 people consisting of 1 female and 10 males aging from 21 to 28 to are invited to accomplish this. All users are naïve to the purpose of the study and the only action is to watch each MTV and give two scores (*A* and *V* scores) from Table 1. Before giving scores to any MTV, they are first asked to have a quick overview of our dataset and informed about the meanings of the options they would choose from.

For each MTV, the Arousal ground truth *ga* and Valence ground truth *gv* are calculated based on the rules represented in Table 2. Since only two clusters are generated for this MTV dataset based on Arousal, the value of *ga* is quantified into two scales rather than four scales.

Table 1. The options and corresponding descriptions

Arousal Level		Valence Level	
score	Description	score	Description
1	very peaceful	-2	sad
2	a little peaceful	-1	a little sad
3	a little intense	1	a little pleasant
4	very intense	2	very pleasant

Table 2. Rules for ground truth computation

Average score of "Arousal Level"	Quantified value of *ga*	Average score of "Valence Level"	Quantified value of *gv*
(3,4]	2	(1,2]	2
[2.5,3]	2	(0,1]	1
[2,2.5)	1	[-1,0]	-1
[1,2)	1	[-2,-1)	-2

2) Evaluation of the result: The affective state of each cluster is first quantified according to Table 3.

Table 3. The quantified affective states of each cluster

Cluster	Arousal Value	Valence Value	Cluster	Arousal Value	Valence Value
Cluster 1	2	2	Cluster 5	2	-1
Cluster 2	1	2	Cluster 6	1	-1
Cluster 3	2	1	Cluster 7	2	-2
Cluster 4	1	1	Cluster 8	1	-2

Then, the numerical results in Table 4 are computed by:

$$P_A(i) = \frac{Num_i(ga = A_i)}{NC_i} \bullet 100\% \quad (1) \qquad P_V(i) = \frac{Num_i(gv = V_i)}{NC_i} \bullet 100\% \quad (2)$$

$$P_{A-V}(i) = \frac{Num_i(gv = V_i \;\&\; ga = A_i)}{NC_i} \bullet 100\% \quad (3)$$

$$R_{A-V}(i) = \frac{Num_i(gv = V_i \;\&\; ga = A_i)}{TotalNum(gv = V_i \;\&\; ga = A_i)} \bullet 100\% \quad (4)$$

Where $P_A(i)$, $P_V(i)$, $P_{A-V}(i)$ and $R_{A-V}(i)$ denote the *Arousal-only Precision* (*A-P*), *Valence-only Precision* (*V-P*), *A-V Precision* (*AV-P*) and *A-V Recall* (*AV-R*) respectively. A_i and V_i denote the Arousal and Valence value of cluster *i*. NC_i denotes the number of MTV in cluster *i*. $Num_i(C_m)$ returns the number of MTV that satisfies C_m in cluster *i*. $TotalNum(C_n)$ returns the number of MTV which satisfies C_n in the total dataset.

In Table 4, it is clear that the *A-P* is higher than the *V-P* in general. This is mainly because Arousal features are more robust and humans' understandings about Arousal are more consistent than their responses about Valence. It also can be indicated that the *V-P, AV-P* and *AV-R* of cluster 3, 4, 5 and 6 are lower than those of the other four clusters. This is because these four clusters contain some MTVs that are obscure in affective state and users' responses about these

MTVs are distinct. The *AV-P* and *AV-R* numbers are smaller than the *A-P* and *V-P* is because finer quantization (thus more strict requirement as shown in Eqs. (3-4)) is required.

Table 4. The numerical evaluation results

Cluster ID (number of MTVs in this cluster)	***A-P***	***V-P***	***AV-P***	***AV-R***
1 (21)	85.7%	76.2%	76.2%	64.0%
2 (11)	81.8%	80.1%	63.6%	77.7%
3 (32)	84.3%	62.5%	53.1%	68.0%
4 (19)	78.9%	63.2%	52.6%	66.7%
5 (27)	85.2%	74.1%	66.7%	72.0%
6 (18)	88.9%	50.0%	50.0%	47.4%
7 (14)	92.9%	85.7%	78.6%	78.6%
8 (14)	90.0%	89.2%	72.5%	54.2%

Table 5. The comparison of *A-V* features

Comparison	**Arousal Features**	**Valence Features**
AV1 [1-3]	Motion Intensity, Shot Switch Rate, Sound Energy	Pitch
AV2 [7]	Sound Energy, Tempo, Beat Strength	Pitch, Rhythm Regularity
AV3	Motion Intensity, Short Switch Rate, Sound Energy, ZCR, Tempo, Beat Strength	Rhythm Regularity, Pitch, Lighting, Saturation, Color Energy

3) Comparisons of results produced by different A-V features: Compared with previous work [1-3, 7], we investigate several more affective features in our work. To prove the validity of our features, the experimental results of different *A-V* feature are compared in this experiment. The *A-V* features compared are shown in Table 5.

AV1 includes the features used in the work of Hanjalic, *et al*. [1-3] about affective movie content analysis. *AV2* includes some of the features in the work of Lu, *et al*. [7] about affective music content analysis. *AV3* denotes the features used in our work. The MTVs are clustered into 8 categories in the same way mentioned above. The comparisons of the results are illustrated in Fig. 7. *Y*-axis stands for the precision and *X*-axis stands for three *A-V* features used in Table 5. The precision improvements of our *A-V* features are shown in Table 6. It is clear that our *A-V* features improve the performance by a significant margin than the other two for affective MTV content analysis.

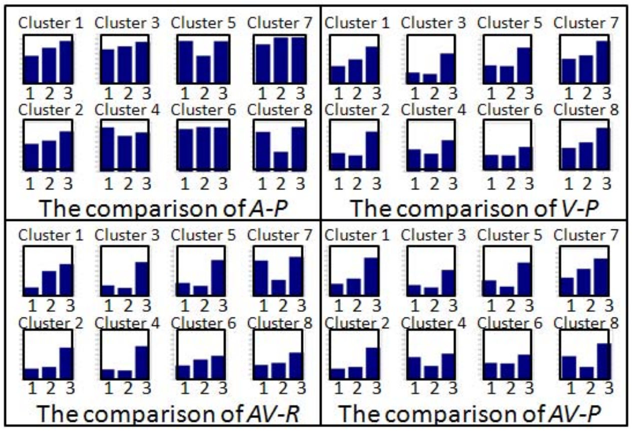


Fig. 7. The comparisons with different *A-V* features

Table 6. The precision improvement

	A-P	***V-P***	***AV-P***	***AV-R***
Improvements over ***AV1***	13%	63%	96%	124%
Improvements over ***AV2***	21%	57%	122%	121%

4. 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, 8 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.

5. ACKNOWLEDGEMENT

The authors gratefully acknowledge the support of K. C. Wong Education Foundation, Hong Kong and Science100 Plan of Chinese Academy of Sciences under Grant 99T3002T03.

6. REFERENCES

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

[2]. A. Hanjalic and L.Q. Xu, "Affective Video Content Representation and Modeling," *IEEE Transactions on Multimedia*, pp.143 – 154, Feb. 2005.

[3]. A. Hanjalic, "Adaptive Extraction of Highlights from a Sport Video Based on Excitement Modeling," *IEEE Transactions on Multimedia*, pp.1114 – 1122, Dec. 2005.

[4]. M. Bradley, "Emotional Memory: A Dimensional Analysis," *Essays on Emotional Theory*, Hillsdale, NJ: LEA, pp. 97–134, 1994.

[5]. J. Russell and A. Mehrabian, *Evidence for a Three-Factor Theory of Emotions*, J. Res. Person, vol. 11:273–294, 1977.

[6]. H. Schlosberg, "Three Dimensions of Emotion," *Psychol. Rev*, vol. 61, no. 2, pp. 81–88, Mar. 1954.

[7]. L. Lu, D. Liu and H.J. Zhang, "Automatic Mood Detection and Tracking of Music Audio Signals," *IEEE Transactions on Audio and Language Processing*, Vol. 14(1), January, 2006.

[8]. B. J. Frey and D. Dueck, "Clustering by Passing Messages Between Data Points," *SECINCE*, Vol.315:972-976,Feb. 2007.

[9]. T. Zhang and C.-C. Jay Kuo, "Audio Content Analysis for Online Audiovisual Data Segmentation and Classification," *IEEE Transactions on Speech And Audio Processing*, Vol. 9(4), pp. 441-457, MAY, 2001.

[10]. N. Maddage, C. S. Xu, M.S. Kankanhalli and X. Shao, "Content-based Music Structure Analysis with Applications to Music Semantics Understanding," *Proc. of ACM Multimedia*, pp. 112-119, Oct. 2004.

[11]. H. L. Wang and L.F. Cheong, "Affective Understanding in Film," *IEEE Transactions on Circuits and Systems for Video Technology*, Vol. 16(6), pp. 689-704, June, 2006.

[12]. D. Bordwell and K. Thompson, *Film Art: An Introduction,* 7rd ed. New York: McGraw-Hill, 2004.

[13]. H. Zettl, *Sight Sound Motion: Applied Media Aesthetics*, 3rd ed. Belmont, CA: Wadsworth, 1998.