Personalized MTV Affective Analysis Using User Profile

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

[1] Key Lab of Intell. Info. Process., Inst. of Comput. Tech., Chinese Academy of Sciences, Beijing 100080, China
[2] Graduate University of Chinese Academy of Sciences, Beijing 100080, China
[3] Department of Computer Science, University of Texas at San Antonio, TX 78249, USA
{slzhang,qmhuang,sqjiang,wgao}@jdl.ac.cn, qitian@cs.utsa.edu

Abstract. At present, MTV has become an important favorite pastime to people. Affective analysis which can extract the affective states contained in MTVs could be a potential and promising solution for efficient and intelligent MTV access. One of the most challenging and insufficiently covered problems of affective analysis is that affective understanding is personal and various among users. Consequently, it is meaningful to develop personalized affective modeling technique. Because user's feedbacks and descriptions about affective sates provide valuable and relatively reliable clues about user's personal affective understanding, it is supposed to be reasonable to conduct personalized affective modeling by analyzing the affective descriptions recorded in user profile. Utilizing the user profile, we propose a novel approach combining support vector regression and psychological affective model to achieve personalized affective analysis. The experimental results including both user study and comparisons between current approaches illustrate the effectiveness and advantages of our proposed method.

Keywords: Affective Content Analysis, Dimensional Affective Model, Support Vector Regression, Personalized Affective Analysis.

1 Introduction

To date, tons of works have been reported on affective music, movie content analysis [1-6]. Similar to music and movie that are important favorite pastimes, MTV (Music TV) which combines the features of both is an important entertaining media form to people. Furthermore, in recent years, the amount of MTV data has been increasing at an astonishing speed and mobile sets such as cell phone and music player like iPod, Zune which can play both video and music have become more and more ubiquitous. Accordingly, MTV has become more popular and widely spread. Meanwhile, the fast increasing number of MTV and storage capacity of our digital sets has generated new requirements for more efficient and intelligent MTV retrieval and access. The traditional methods of MTV classification based on Artist, Album, Title, are not intelligent enough and have many limitations when people want to manage or retrieve MTVs using semantic and abstract concepts. For example, when an individual is sad and tired, he/she might want to enjoy some happy and energetic MTVs on iPod. In this

Y.-M.R. Huang et al. (Eds.): PCM 2008, LNCS 5353, pp. 327–337, 2008.

case, the traditional ways hardly work. Because affective analysis tries to catch human abstract thinking and recognize the affective states, it could bring great surprise and new experiences for MTV lovers. For example, new commercial systems can be built to enable affective analysis based MTV retrieval. Meanwhile, affective analysis based MTV recommender can be implemented to recommend MTVs in a new and interesting way.

In order to identify the affective states contained in MTVs, several problems should be considered:

1. Extraction of valid affective features.
2. Bridging the gap between affective features and affective states.
3. Users' understandings of affective states are various, so the affective model should be properly established to take user's personality into consideration.

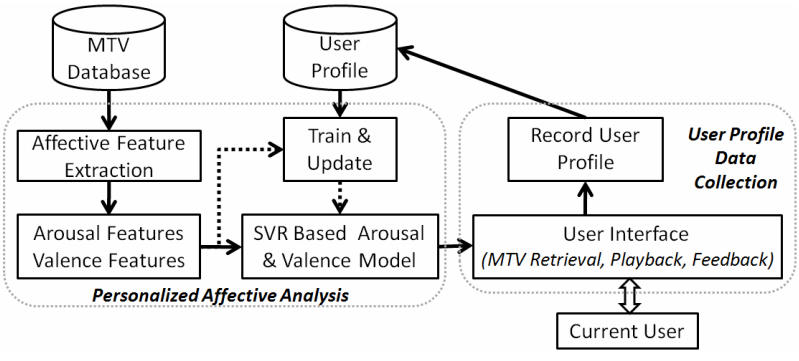

Fig. 1. Illustration of the proposed framework

The first two problems have already been discussed in literature [1-6]. However, the third problem has not been fully covered in existing affective analysis methods. In order to solve the three problems, our framework shown in Fig.1 is proposed. In Personalized Affective Analysis module, we first extract valid and well-documented affective features by referring to related work, musical knowledge and cinematographic study. Then, we apply machine learning method to bridge the affective gap by combing the psychological Dimensional Affective Model [7] and Support Vector Regression (SVR). Dimensional Affective Model which is also called as Arousal-Valence (A-V) model [7] is currently the most popular psychological affective model utilized in affective content analysis. In this model, human affective responses and states are represented using two basic independent components: Arousal which stands for the intensity of affective experience and Valence which typically characterizes the level of "pleasure". Finally, in order to make the affective analysis more personal, we train and update the two SVR based models (Arousal model and Valence model) with the user profile collected. User Profile Data Collection module is proposed to collect the user profile data. A user interface is constructed to enable users retrieve, play MTVs and provide their feedbacks. The user feedbacks will then be captured to update the profile database.

The novel contributions of our work are summarized as follows:

- ✧ The variations of affective understandings are taken into consideration for affective modeling. The effectiveness and rationality are proved better than the existing affective modeling methods.
- ✧ The techniques proposed in this work could be utilized for establishing a MTV affective retrieval and management system, which might provide large potential applications for music players.

The rest of the paper is organized as follows. Section 2 introduces the related work. Section 3 describes the algorithms for capturing and storing user profile. Section 4 presents our affective modeling algorithm. Experiments and analysis are discussed in Section 5. Finally, conclusions and future work are summarized in Section 6.

2 Related Work

As a newly developing research area, Affective Content Analysis has been paid much attention and much work has been reported in recent years. The existing affective analysis methods 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 affective groups. For example, Moncrieff [1] uses four sound energy events to detect "startle", "apprehension", "surprise" and "climax" in videos. Xu [2] trains Hidden Markov Models (HMM) to classify affective states in videos. Lu [3] extracts three types of audio features and classifies music segments into four mood categories with hierarchical framework. *Dimensional affective content analysis* commonly employs the psychological A-V Affective Model for affective state representation and modeling. Some representative works are reported by Hanjalic and Xu [4]. Modeling Arousal and Valence using linear feature combinations, the authors can obtain the Arousal and Valence values of different parts of the video and draw affective curves in the A-V space. Consequently, the affective states expressed in videos can be visualized. Different from the Arousal and Valence modeling methods proposed by Hanjalic, the work of Arifin [5] successfully takes the influences of former emotional events and larger emotional events into consideration by employing Dynamic Bayesian Network for affective modeling. The authors evaluate the proposed method using 23 videos and improvement of 9% over the work of Hanjalic [4] is reported.

The above work mainly focuses on the first two problems mentioned before. The third problem has not been fully investigated and studied. In fact, personalization is gaining more attention nowadays. An overview of current personalization techniques can be found in [8] and [9]. Among current personalization techniques, feedback and profiling are effective and commonly used for capturing user's preferences, intentions, interests, and so on. For example, the personal TV listing system: PTVplus [10] exploit user profile including user's viewing behavior, playback, implicit feedbacks to learn user's preferences and present personalized daily TV guide. In some webpage recommender systems such as WebMate [11] and SiteIF [12], user profile is denoted as a bag of keywords to represent user's preferences and is updated with both explicit feedbacks and implicit feedbacks. Discussions about learning to personalize and user

profile acquirements can be found in [13]. The above mentioned personalization mostly focuses on the various preferences and requirements among different users. Since affective understandings are also various among users, it is necessary to introduce personalization into the affective analysis field. In next sections, we will introduce our approach of personalized affective analysis based on user profile.

3 User Profile Data Collection

3.1 The Design of User Interface

In order to collect user profile data, a user interface is constructed to enable users retrieve MTVs by affective state and provide feedbacks and descriptions. In the psychological points of view, human emotion can be denoted as a two-dimensional continuous affective space and different regions represent different affective states and emotions [7]. Based on this principle, MTVs' affective states which are computed as two-dimensional vectors (i.e., Arousal value and Valence value in the range of 0 to 1 [14]) could be visualized as points in the A-V space (Fig. 2-a). Thus the affective states of MTVs can be intuitively visualized according to their spatial positions and users can retrieve MTVs by selecting different regions in affective space (Fig. 2-b).

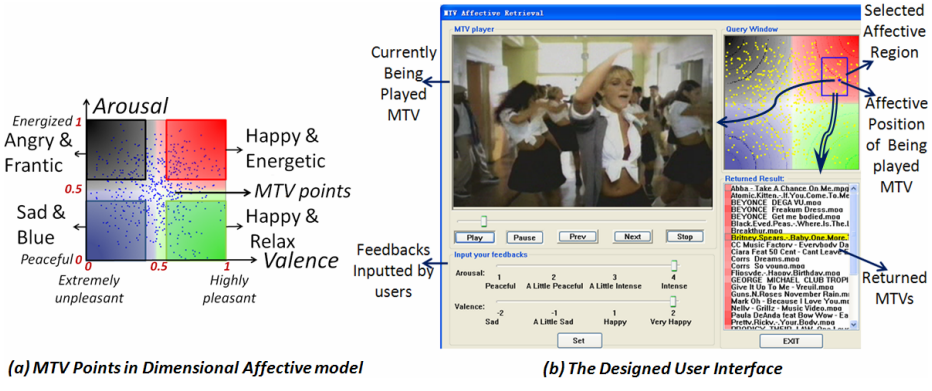

Fig. 2. The design of user interface

3.2 Record User Profile Data

We define the profile data as follows:

$$Profile_m = \{P_1, P_2, ..., P_K\} \qquad P_i = [MTVid, Arousal, Valence] \tag{1}$$

where $Profile_m$ denotes the profile database of user m, P_i is the i-th item in the profile database. Each profile item is denoted as a vector containing three elements: MTV's unique ID in database, user's descriptions about the MTV's Arousal and Valence.

When users retrieve and play MTVs, if they are not satisfied with the computed affective states, they can provide their feedbacks to the system. Users will then be asked to give two scores to describe their own opinions about Arousal (1: peaceful; 2: a little

peaceful; 3: a little intense; 4: very intense) and Valence (-2: unhappy; -1: somewhat unhappy; 1: happy; 2: very happy) of the played MTVs (Fig. 2-b). For example, if a user plays a MTV in the most upper-right part of the affective space and thinks this MTV is not very energetic, which means its actual Arousal is lower than the computed value, the user can set the MTV's Arousal score a lower value (1, 2 or 3). We set the feedbacks in four scales is mainly because of the two considerations:

1. Most users want to provide their feedbacks in a simple form.
2. It is not easy for users to confirm affective states' precise positions in affective space but it is easy for them to confirm the rough regions containing the affective states.

After a user providing feedbacks, new profile items will be generated:

1. Compute the *Arousal* and *Valence* components in the new profile item:

$$Arousal = (Ascore - 1) \times 0.25 + 0.125$$
$$Valence = \begin{cases} (Vscore + 2) \times 0.25 + 0.125 & if: Vscore < 0 \\ (Vscore + 1) \times 0.25 + 0.125 & if: Vscore > 0 \end{cases} \quad (2)$$

   where *Arousal* and *Valence* denote the two elements in profile item, *Vscore* and *Ascore* denote the user's scores for Valence and Arousal.
2. Assign the *MTVid* as the ID of the MTV scored. If this MTV exists in the profile database, then overwrite the old profile item; otherwise, add the new item to the profile database.

Consequently, as a user retrieves, plays MTVs and provides feedbacks, the user's profile data will be updated.

4 Personalized Affective Modeling Based on User Profile

4.1 Extracted Affective Features

In most of the cases, the visual content of MTV is carefully selected by directors and artists to coordinate with the music. Consequently, both the audio and visual content are used to extract affective features. The features employed for Arousal modeling include: *motion intensity* [4, 6], *short switch rate* [4], *zero crossing rate* [15], *music tempo* [3] and *beat strength* [3]. The features used for Valence modeling include: *lighting* [16], *saturation* [16], *color energy* [6], *rhythm regularity* [3] and *pitch* [4]. Most of the features have been utilized and proved effective in related work for music and movie affective analysis. Furthermore, most of them are extracted based on psychological knowledge, musical knowledge or cinematographic study. For example, shot switch rate is a powerful tool used by directors to control the tempo of a video. Rhythm is commonly utilized by artists to express their affection or emotion. Lighting and color are effective and common tools used specifically for the purpose of affecting the emotions of viewers and establishing the mood of a scene in cinematography. Therefore, the extracted features are supposed to be valid and well-documented. Due to the limitation of space, we will skip the details of extractions of these features [3-4, 6, 14-16].

4.2 Profile Based Personalized Arousal and Valence Model

The definition of profile based personalized Arousal and Valence modeling is given as follows:

Definition: *Profile based personalized Arousal and Valence modeling is to look for two functions f_V, f_A for each user that map the Valence features $Feature_V$ and Arousal features $Feature_A$ to real values of Valence and Arousal within 0 and 1 respectively, based on collected user profile.*

$$f_V : Feature_V \mapsto [0,1] \qquad f_A : Feature_A \mapsto [0,1]$$

According to this definition, we want to calculate the Arousal and Valence values of each MTV based on the underlying relationship between user's affective understanding and the extracted features indicated by the recorded profile. Once the f_V, f_A can be properly constructed, the affective state of each MTV can be computed by feeding the extracted features into the two functions.

4.3 Solution to the Problem

To utilize the profile data and learn user's personal affective understanding, we will train a model to give "best fit" of the user's affective descriptions. This is similar to a regression problem. So, we choose a suitable regression model to achieve this.

Support Vector Machine (SVM), enjoys solid theoretical foundations and has demonstrated outstanding performance in many machine learning problems [17]. Support Vector Regression (SVR) [18] is based on the SVM theories and presents several advantages:

- ✧ Providing better prediction on unseen data. SVR employs the Structural Risk Minimization principle, which is superior to Empirical Risk Minimization principle in the aspect of generalization.
- ✧ Providing a better solution for a training problem. This is important for personalized affective modeling. If we have effective affective features and detailed, precise user affective descriptions, it is possible to achieve personalized affective modeling by training SVR models.
- ✧ An important basis for commonly used linear combination based affective modeling is that the relationships between affective features and affective states are linear. This is a loose hypothesis since it still remains to be further studied by psychologists on the mechanism of human's affective responses. Therefore, it is more reasonable to build a non-linear model for affective state computation.

Due to the above considerations, we build two SVR models for each user: a model for Arousal and the other for Valence.

Train the default SVR models: the system does not have any profile data from the new user. So we train default SVR models for each new user. It is critical to choose the training set for the default SVR models. We invited 10 people to conduct a user study, from which, we collected these users' descriptions (scores) for 552 MTVs. The MTVs suitable for default SVR training should be the ones of which different users

give similar descriptions. Consequently, we sorted the 552 MTVs by the times that they are similarly described and selected the top 50 MTVs to generate the default training set. It must be pointed out that, if more users begin to use the system, better default training set can be generated. The process for default Arousal SVR model training is gives as follows:

1. Convert each selected MTV's *Ascore* into *Arousal* with Eq. (2) and generate the training set: $(Arousal, Feature_A)_1 \ldots (Arousal, Feature_A)_{50}$.
2. SVR model with RBF kernel function is selected and 5-fold cross validation is utilized for parameter selection.
3. Train the default Arousal SVR model with the selected parameters.

The default Valence SVR is trained in similar process. After the SVR models being generated, the affective states of new MTVs can be computed.

Update the SVR models: when the system begins to collect more user profile data, the SVR models can be updated. We define several rules to produce new training set by combining the old training set and user's new profile data:

1. Generate new items $(Arousal, Feature_A)$ and $(Valence, Feature_V)$. *Arousal* and *Valence* can be get from the new user profile.
2. If the newly described MTVs exist in the old training set, overwrite the old items with the new ones. If not, add them to the old training set.

Since training new SVR models is time consuming, the models will only be updated when the number of newly updated items in training set exceeds a threshold.

5 Experiments and Evaluations

5.1 MTV Dataset

Our dataset consists of 552 MTVs of MPEG format (about 25 GB in total). MTVs in the dataset are recorded in different languages (e.g., English, Chinese, Japanese, Korean), different periods and different styles (e.g., Rock, Jazz, R&B, country music). Therefore, this MTV dataset is representative enough to conduct experiments.

5.2 User Study for Profile Data Preparation

In order to capture enough profile data and ground truth to finish the experiments, we invite 10 people consisting of 1 female and 9 males aging from 21 to 28 to accomplish a subjective user study. During this, the only action of the participant is to watch MTVs and give two scores (1, 2, 3, 4 for Arousal and -2, -1, 1, 2 for Valence) for each one. Before the user study, they are asked to have a quick overview of our dataset and are informed about the meanings of the options they would choose from. Since it is time consuming to score all MTVs by each person, every participant is required to score 150 evenly selected MTVs rather than the whole. It is necessary to point out that although these scores are not directly captured from feedbacks, they still reflect users' affective descriptions. Therefore, they are valid to generate profile data and to test the effectiveness of the proposed methods.

5.3 Effectiveness of Our Profile Based Personalized Affective Modeling

With the introduced interface, users can retrieve their desired MTVs by selecting affective regions. Therefore, the selected regions can be taken as users' affective queries, while MTVs contained in the selected regions can be considered as query results. It can be inferred that users will be satisfied with MTV affective analysis if the MTVs falling in the selected regions present high precision rates and recall rates. Therefore Regional Precision Rate (*PRreg*) and Regional Recall Rate (*RRreg*) are used to measure the effectiveness of our method. Each user's affective models are first trained with the default training set. Then, 80 of scored MTVs are utilized for generating profile data and updating models; the left 70 MTVs which are not included in the default training set are used for testing. For each user, we calculate the *PRreg* and *RRreg* with Eq. (3) in the 4×4 (16) regions divided by the two 4-scale scores. Then, based on each user's *PRreg* and *RRreg*, Fig. 3 could be computed and drawn.

$$\begin{aligned} RRreg_i &= CorrectNum_i / ScoredNum_i \bullet 100\% \\ PRreg_i &= CorrectNum_i / ComputedNum_i \bullet 100\% \end{aligned} \tag{3}$$

where $ScoredNum_i$, $CorrectNum_i$ and $ComputedNum_i$ denotes the number of scored, correctly computed and totally computed MTVs falling in region *i* respectively.

From Fig. 3-a, after updating the default SVR models for each user, the average *PRreg* and *RRreg* are stable in different affective regions. The overall average *PRreg* and *RRreg* reach 71.6% and 70.7%, respectively. Considering fine quantization (16 regions denoting 16 different affective states are considered), the result is still promising for affective based MTV retrieve. From Fig. 3-b and Fig. 3-c, it can be seen that the performance of the proposed model is stable for different users although they may have different affective understandings. Consequently, the effectiveness and validity of the proposed model can be proved.

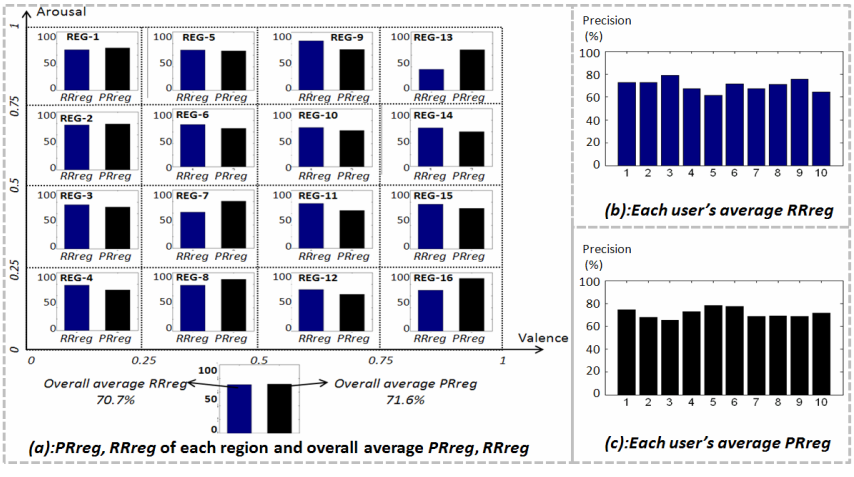

Fig. 3. The effectiveness of our proposed affective model

5.4 Performance Evaluation on SVR Based Affective Model

In this experiment, we will compare our SVR based affective model with other models. The descriptions of compared models are listed as follows:

1. ***Comparison1 (C1): linear feature combination.*** Linear feature combination is a commonly used method for affective modeling [4]. Linear feature combination with equal weights is used as a baseline for comparison.
2. ***Comparison2 (C2): linear feature combination with weight adjustment.*** In order to make the comparisons fair, we introduce a weight adjusting method to update each feature's weight according to the user profile data. The basic idea of this method is to enlarge the weights of more efficient features and decrease the weights of less important features. The initial weight of each features are set identical. Weights are first adjusted based on the default training set and then updated according to each user's profile data.
3. ***Comparison3 (C3): SVR based affective mode,*** the model introduced in the paper.

For each user, the 150 scored MTVs are divided into two groups (Group A containing 80 MTVs for updating the affective models and Group B containing the other 70 MTVs for testing) in similar way mentioned before. Fig. 4 presents the curves of Overall Arousal Precision Rate (*OAPR*), Overall Valence Precision Rate (*OVPR*) and Overall Arousal-Valence Precision Rate (*OAVPR*) computed by *C1, C2* and *C3* when the models and weights are updated.

Fig. 4-a illustrates the *OAPR* curves. After updating models and weights with the default training set, the performances of *C2* and *C3* remain stable and do not show obvious enhancements for Arousal. This might be because most users' opinions about Arousal are similar and the default training set is representative of most users' understandings about Arousal. From Fig. 4-b, it is clear that when more user profile data is utilized for model updating, the performances of both *C2* and *C3* show obvious improvements. This is because users' opinions about Valence are various and as default models learn from user profile, they will become more "personal" and effective. Since both Arousal and Valence are considered for *OAVPR* computation, which poses a stricter requirement, the *OAVPR* curves in Fig. 4-c are lower than *OAPR* and *OVPR* curves. According to the three figures, it is clear that the precision curves of *C3* are all above the ones of *C1* and *C2*. Consequently, we can draw the conclusion that, our model presents better performance for affective analysis than the commonly used linear feature combination scheme.

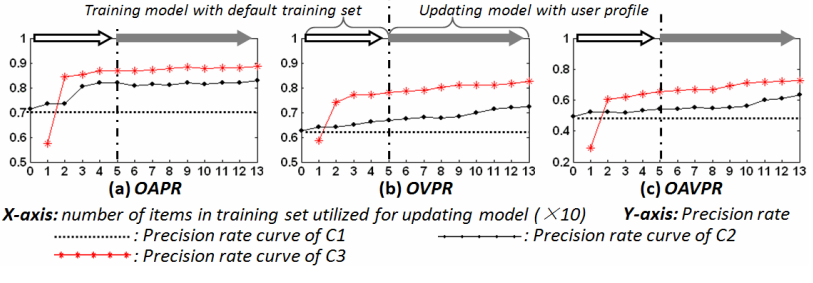

Fig. 4. The performance comparisons between C1, C2 and C3

5.5 Discussions about the Limitations and Solutions

Besides the advantages, we must address the limitations and challenging issues with our schemes, as well as provide feasible directions for solutions in our future work.

The first limitation is the computational complexity, which is mainly due to the property of SVR. In order to update the SVR based affective models, new parameters should be selected by cross validation. Besides that, as the profile database increases, the training set will be augmented, which will also slow the training speed down. To solve this problem, we will investigate incremental machine leaning methods and better training set generation scheme.

The second limitation is the noise in user profile. The presence of noise in user profile is inevitable when feedbacks are collected. In addition to new training set generation scheme, more robust models will be designed.

Third, the feedbacks are provided in only 4 fixed scales, which will limit the precision of affective analysis. We will try to solve this problem by proposing better feedback schemes and employ implicit feedbacks to capture more precise information about user's personal affective understanding.

Despite of the above limitations, personalized affective analysis by capturing and utilizing user's affective descriptions has been proved feasible and effective in our work. Our future work will be conducted to overcome the above limitations.

6 Conclusions

In this paper, we propose a novel profile based approach for personalized affective modeling. Compared with most of the previous methods, the important contribution of our work is that we try to take the user's personal affective understanding into consideration for affective modeling. The basic idea of this paper is to utilize the potential relationship between user's affective understanding and affective features indicated by the user profile for personalized affective modeling. In order to achieve this, we first design a user interface to collect user profile by recording feedbacks. Then, we extract affective features from both audio and visual content of MTVs. Finally, we train and update two SVR models for the two affective components utilizing the user profile. Experiments prove that our proposed method is promising and more effective that the linear feature combination scheme. Our future work will be carried out to overcome the current scheme's limitations analyzed in the paper.

Acknowledgement

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

References

1. Moncrieff, S., Dorai, C., Venkatesh, S.: Affect Computing in Film Through Sound Energy Dynamics. In: ACM International Conference on Multimedia, pp. 525–527 (2001)
2. Xu, M., Chia, L.T., Jin, J.: Affective Content Analysis in Comedy and Horror Videos by Audio Emotional Event Detection. In: IEEE International Conference on Multimedia and Expo. (2005)
3. Lu, L., Liu, D., Zhang, H.J.: Automatic Mood Detection and Tracking of Music Audio Signals. IEEE Transactions on Audio and Language Processing 14(1) (2006)
4. Hanjalic, A., Xu, L.Q.: Affective Video Content Representation and Modeling. IEEE Transactions on Multimedia, 143–154 (February 2005)
5. Arifin, S., Cheung, P.Y.K.: A Computation Method for Video Segmentation Utilizing the Pleasure-Arousal-Dominance Emotional Information. In: ACM International Conference on Multimedia, pp. 68–77 (2007)
6. Wang, H.L., Cheong, L.F.: Affective Understanding in Film. IEEE Transactions on Circuits and Systems for Video Technology 16(6), 689–704 (2006)
7. Schlosberg, H.: Three Dimensions of Emotion. Psychol. Rev. 61(2) (March 1954)
8. Venkatesh, S., Adams, B., Phung, D., Dorai, C., Farrell, R.G., Agnihotri, L., Dimitrova, N.: You Tube and I Find"-Personalizing Multimedia Content Access. Proceedings of the IEEE 96, 697–711 (2008)
9. Sebe, N., Tian, Q.: Personalized Multimedia Retrieval: The New Trend? In: ACM International Workshop on Multimedia Information Retrieval, pp. 299–306 (September 2007)
10. Sullivan, D.O., Smyth, B., Wilson, D.C., McDonald, K., Smeaton, A.: Improving the Quality of the Personalized Electronic Program Guide. User Modeling and User-Adapted Interaction 14, 5–36 (2004)
11. Chen, L., Sycara, K.: WebMate: Personal Agent for Browsing and Searching. In: Proc. 2nd Int. Conf. on Autonomous Agents, pp. 132–139 (1998)
12. Magnini, B., Strapparava, C.: User Modeling for News Web Sites with Word Sense Based Techniques. User Modeling and User-Adapted Interaction 14 (2004)
13. Webb, G.I., Pazzani, M.J., Billsus, D.: Machine Learning for User Modeling. User Modeling and User-Adapted Interaction 11, 19–29 (2001)
14. Zhang, S.L., Tian, Q., Jiang, S.Q., Huang, Q.M., Gao, W.: Affective MTV Analysis Based on Arousal and Valence Features. In: IEEE International Conference on Multimedia and Expo., pp. 1369–1372 (2008)
15. Zhang, T., Kuo, C.C.J.: Audio content analysis for online audiovisual data segmentation and classification. IEEE Transactions on Speech and Audio Processing 9(4), 441–457 (2001)
16. Bordwell, D., Thompson, K.: Film Art: An Introduction, 7th edn. McGraw-Hill, New York (2004)
17. Burges, C.J.C.: A Tutorial on Support Vector Machine for Pattern Recognition. Data Mining and Knowledge Discovery 2, 121–167 (1998)
18. Smola, A.J., Scholkopf, B.: A Tutorial on Support Vector Regression. Statistics and Computing 14, 199–222 (2004)