Flexible Presentation of Videos Based on Affective Content Analysis

Sicheng Zhao, Hongxun Yao, Xiaoshuai Sun, Xiaolei Jiang, and Pengfei Xu

School of Computer Science and Technology, Harbin Institute of Technology,
No.92, West Dazhi Street, Harbin, P.R. China, 150001
{zsc,h.yao,xiaoshuaisun,xljiang,pfxu}@hit.edu.cn

Abstract. The explosion of multimedia contents has resulted in a great demand of video presentation. While most previous works focused on presenting certain type of videos or summarizing videos by event detection, we propose a novel method to present general videos of different genres based on affective content analysis. We first extract rich audio-visual affective features and select discriminative ones. Then we map effective features into corresponding affective states in an improved categorical emotion space using hidden conditional random fields (HCRFs). Finally we draw affective curves which tell the types and intensities of emotions. With the curves and related affective visualization techniques, we select the most affective shots and concatenate them to construct affective video presentation with a flexible and changeable type and length. Experiments on representative video database from the web demonstrate the effectiveness of the proposed method.

Keywords: Video presentation, affective analysis, emotion space, HCRFs.

1 Introduction

The explosion of multimedia contents has resulted in a great demand of video presentation. On one hand, viewers need to get a gist of video content, watch video highlights due to time limit and then make the decision to view the entire video (e.g. a movie) or not. On the other hand, video broadcast platforms, especially television stations, have to check substantial videos and select legal and valuable ones to play, which is a time-consuming and tedious task. Thus, effective video presentation techniques can make video reviewers' work more convenient and efficient.

Most previous works on content-based video presentation focused on certain type of videos, such as sports videos, home videos, or summarizing videos by event detection [1-5]. Liu *et al.* [1] proposed a novel flexible racquet sports video content summarization framework, by combining the structure event detection method with the highlight ranking algorithm. Zhao *et al.* [2] proposed a novel system of highlight summarization in sports videos based on replay detection. Based on videos' three properties: emotional tone, local main character and global main character, Xiang and Kankanhalli [3] employed affective analysis to automatically create adaptive presentations from home videos for three types of social groups: family, acquaintance

S. Li et al. (Eds.): MMM 2013, Part I, LNCS 7732, pp. 368–379, 2013.

and outsider. Wang *et al.* presented an approach for event driven web video summarization by tag localization and key-shot mining in [4], and turned a movie clip to comics automatically in [5].

Meanwhile, affective image and video content analysis has been paid much attention recently. Generally, all these works are based on two kinds of emotion models: categorical (discrete) emotion states, or dimensional (continuous) emotion space. In categorical (discrete) models [3, 6-8], emotions are usually considered to be one of a few basic categories, such as *fear*, *anger*, *happy*, *sad*, *etc*. Machajdik and Hanbury [6] exploited theoretical and empirical concepts from psychology and art theory to extract image features that are specific to the domain of artworks with emotional expression, and classified an image into eight affective categories: *amusement*, *awe*, *contentment*, *excitement*, *anger*, *disgust*, *fear* and *sad*. Kang [7] trained two Hidden Markov Models (HMMs) to detect affective states in movies. Xu *et al.* [8] utilized fuzzy clustering and HMMs to classify video shots in movies into five emotion types, including *fear*, *anger*, *sad*, *happy*, and *neutral*. Dimensional or continuous models mostly employ the 3-D Valence-Arousal-Control (VAC) emotion space [9] or 2-D Valence-Arousal (VA) emotion space for affective representation and modeling [10-13]. Hanjalic and Xu [10] modeled arousal and valence separately using linear feature combinations, and drew arousal curve and valence curve or combined affect curve. A novel computation method for affect-based video segmentation, which is designed based on the Pleasure-Arousal-Dominance (P-A-D) emotion model, was introduced in [11]. Zhang *et al.* [12] proposed affective information based movie browsing using affective visualization techniques. Nicolaou *et al.* [13] proposed a multi-layer hybrid framework for emotion classification that is able to model inter-dimensional correlations. Both emotion models have their advantages. The former one is easier for users to understand and label, while the latter one is more flexible and richer in descriptive power.

Some researchers have also considered viewers' affective reactions to video contents for video analysis [14-18]. Ma *et al.* [14] presented a generic framework of video summarization based on the modeling of viewer's attention. Joho *et al.* [15, 16] proposed a new approach for personal highlights detection and video summarization based on the analysis of facial activities of viewers, using the model of pronounced level and expression's change rate. Zhao *et al.* [17,18] presented a novel method to classify, index and recommend videos based on affective analysis, mainly on facial expression recognition of viewers.

However, few works have considered the affective analysis of video content for general video presentation of different genres, while affective information in videos is closely related with users' experiences and preferences [12]. In this paper, we propose a novel method to present general videos of different genres based on affective content analysis. The contributions lie in three aspects: 1. We propose an improved categorical emotion space, by combining the advantages of categorical emotion states and dimensional emotion space; 2. We propose a novel video presentation method with a flexible and changeable type and length; 3. Two HCRFs are trained to compute the labels of types and intensities of emotions, respectively.

The rest of this paper is organized as follows. Section 2 describes the proposed method in detail. We conduct relative experiments and analyze the results in Section 3. Finally, conclusion and future work are discussed in Section 4.

2 The Proposed Method

The framework of the proposed method is shown in Figure 1. First, we segment each video into video shots. Second, affective audio-visual features are extracted and selected. Then affective models are trained to map selected features into affective states in an improved categorical emotion space using HCRFs. Finally, we draw affective curves, visualize affective states, select the most affective shots, and cluster these shots to construct video presentation according to time order. Further, we can change the type and length of video presentation based on users' feedback.

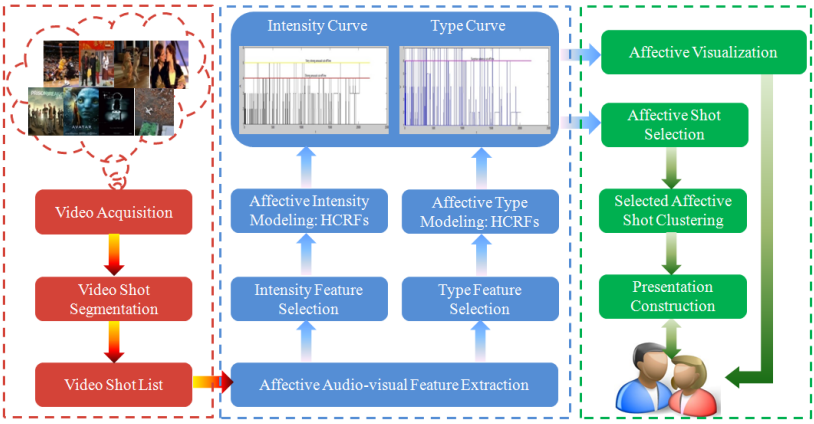


Fig. 1. The framework of the proposed method

2.1 An Improved Categorical Emotion Space and Related Affective Curves

An emotion can be evoked in a user when watching affective videos. Different scenes of different videos result in different emotion types and intensities of viewers. On one hand, categorical emotion states are easier for users to label and understand, while dimensional emotion space can express the intensity of emotions. On the other hand, the former one is computed in a classification framework, while the latter one is done regressively, which makes the computation more complex. By combining their advantages, we propose an improved categorical emotion space named IT emotion space, in which "I" represents the intensity or degree of emotion and "T" stands for the type of emotion. That is, we add discrete intensity dimension to the original categorical emotion states. Assume that the number of T is m, the number of I is n, and the total number of affective states is $N = m \times n$. If T contains the emotion type

of '*neutral*', N turns to $m \times n - m + 1$, as '*neutral*' does not have different intensities. For instance, in this paper, we assume that

T∈{'*neutral*', '*sadness*', '*anger*' , '*disgust*', '*fear*', '*surprise*', '*happiness*'}

and I∈{'*very low*', '*low*', '*mid*', '*strong*', '*very strong*'}.

Totally we can express 31 types of different affective states, as shown in Figure 2(a). For simplicity, we use T∈{1, 2, 3, 4, 5, 6, 7} and I∈{1, 2, 3, 4, 5} to represent different types and intensities of emotions respectively.

In this case, type and intensity of emotion are the functions of different audio and visual features. We model type and intensity time curve as

$$T(k) = F(f^k_{T\_audio}, f^k_{T\_visual}), I(k) = G(f^k_{I\_audio}, f^k_{I\_visual}) \quad (1)$$

where k, $f^k_{T\_audio}$, $f^k_{T\_visual}$, $f^k_{I\_audio}$, $f^k_{I\_visual}$ stand for the kth short, audio and visual features for type and intensity respectively.

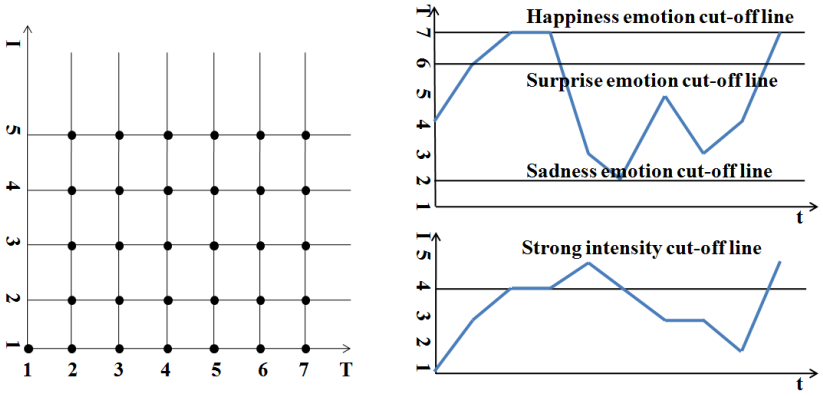


Fig. 2. (a) Illustration of the improved categorical emotion space (b)Related affective time curves with cut-off lines

2.2 Affective Feature Extraction

Based on related works of affective feature extraction [3, 6, 8, 10, 11], we extract 11visual features and 13 audio features, as shown in Table 1.

The *Audio Intensity Features* typically describe the intensity of the audio in video clips. They are closely related with the intensity of emotion and have been frequently used in audio affective analysis. *Mel Frequency Ceptral Coefficients* (*MFCC*) work well for excited and non-excited intensity detection. The *Audio Timbre Features* are frequently used to distinguish different types of sound production. *Rhythm* is an important tool used by artists to express their emotions, based on Onset detection and Drum detection. The *visual clues* in videos are also important in describing the affective states, and are commonly utilized to render the emotions and evoke certain

moods in audience. We extract color and texture visual features for emotion type. After the feature extraction, we employ Gaussian Normalization to normalize these features into [0, 1]. Besides, we also recognize facial expressions and estimate the expression intensities [17-19], as facial expression and its intensity tell the emotion type and intensity of actors to a great extent.

Table 1. Extracted affective features for emotions' type and intensity estimation

	Category		Name
Intensity	Audio	Intensity	Zero Crossing Rate (ZCR), Short Time Energy (STE) , MFCC
		Timbre	Sub-band Peak, Sub-band Valley, Sub-band Contrast
		Rhythm	Tempo, Rhythm Strength, Rhythm Contrast, Rhythm Regularity, Drum Amplitude
	Visual		Motion Intensity, Short Switch Rate, Frame Brightness
	Content		Facial expression intensity estimation
Type	Audio	Timbre	Pitch, Sub-band Peak, Sub-band Valley, Sub-band Contrast, Pitch STD
		Rhythm	Rhythm Regularity
	Visual	Color	Frame Brightness, Saturation, Color Energy, Color Emotion, Colorfulness, Hue
		Texture	Tamura, Wavelet textures, Gray-level Co-occurrence Matrix
	Content		Facial expression recognition

2.3 Affective Labeling

We utilize HCRFs to compute the type and intensity of emotion, based on the extracted audio and visual features. In nature, the main idea behind HCRFs is to enrich CRFs by adding hidden states to capture complex dependencies or implicit structures in the training samples [20, 21].

Suppose $F = \{f(y,h,x;\theta)\}$ be a set of feature functions, and $H = \{h_1, h_2, \cdots, h_m\}$ be a set of hidden variables. Equation (2) and (3) list the probability and potential function.

$$P(y \mid x,\theta) = \frac{\sum_h \exp(\varphi(y,h,x;\theta))}{\sum_{y'}\sum_h \exp(\varphi(y',h,x;\theta))} \tag{2}$$

$$\begin{aligned}\varphi(y,h,x;\theta) &= \sum_j f(x,j)\cdot\theta(h_j) + \sum_j \theta(y,h_j) + \sum_{(j,k)\in E} \theta(y,h_j,h_k), f(x,j) \\ &= [f_{j-1}^T, f_j^T - f_{j-1}^T,]^T\end{aligned} \tag{3}$$

Based on previous works on CRFs and HCRFs [20, 21], we use the following objective function for training the parameters:

$$\sum_i \log P(y_i \mid x_i, \theta) - \frac{1}{2\sigma^2} \| \theta \|^2 \tag{4}$$

Given a test shot with observations $x' = \{x_1, x_2, \cdots, x_n\}$ and trained parameters θ^* of the HCRFs, the label y' of this shot is recognized as follows:

$$y' = \arg\max_{y \in Y} P(y \mid x', \theta^*) \tag{5}$$

2.4 Affective Presentation Construction

Based on the computed affective type and intensity time curves, we are able to present videos flexibly. As affective presentation should contain those shots with high intensity levels, we use a cut-off line to segment the highlight shots. In this paper, we utilize 5 different intensity types, so in the intensity time curve, we draw a constant function i=*level*, *level*∈{1, 2, 3, 4, 5}. And the shots between the intersection points (the constant function and the intensity time curve, that is, $i(t)$=*level*) and the local maximums of the intensity time curve ($i(t)$=5) are the highlights. Further, we can change the value of *level* or cluster the selected shots using *k*-mean clustering to fit the demanded length of the video presentation according to users' time arrangement.

On the other hand, different people at different time have different moods, and want to see video presentation of different types. In this case, we provide a user feedback interface considering users' interests and moods. Users can choose the emotion types in which they are interested, and then we draw a constant function v=*value*, *value*∈{1, 2, 3, 4, 5, 6, 7}. And the shots at the intersection points of the constant function and the type time curve ($v(t)$=*value*) are the selected ones. Further, we can change *level* and *value* simultaneously to specify emotion type and intensity. One example of cut-off line in emotion transitions is given in Figure 2(b).

3 Experiments

In this section, we introduce the creation and labeling of the used dataset, conduct relative experiments, analyze our results and compare them with latest research.

3.1 Dataset and Labels

We select 20 representative videos (*Titanic*, *Avatar*, *Kung Fu Panda 2*, *The Ring*, *Got the Money Anyway* and so on) to create our database. They are of different video types, including *comedy*, *tragedy*, *horror*, *sport videos*, *etc*. The total length is about 40 hours and the average length is about 2 hours. As far as we know, our database is currently one of the most representative video databases [3, 7, 9-11].

We invite 20 volunteers (12 male, 8 female; 10 postgraduates, 10 undergraduates; 15 computer related major, 5 art related major) to label the affective ground truths. They are required to choose emotion type and intensity values (intensity in {1, 2, 3, 4, 5}, type in {1, 2, 3, 4, 5, 6, 7}) to describe each shot's affective states when they are watching the videos. As emotions evoked by videos are strongly influenced by different educations, cultures, backgrounds, *etc*, different participants may have different emotion types and intensities. The final evaluation value of emotion type and intensity for a shot is the latest integer of the average score of all viewers. In order to reduce the interference and confusion in subjective labeling, we provided a standard explanation of each emotion type and intensity. After watching one video, viewers can have one break to lower the impact of fatigue. Totally about 12000 shots' affective ground truths are labeled.

3.2 Results of Affective Labeling

We select six representative video genres to test our emotion type and intensity labeling accuracy: *comedy*, *tragedy*, *action*, *horror*, *exciting*, and *sports*. We separate the data into a training and testing set using K-fold Cross Validation (K=5). As different videos genres have different affective shot types, we test them respectively. In experiment, the number of hidden states of HCRF is selected 3. The experimental results are presented in Table 2, in which "—" represents that there is no such emotion type or intensity manual labels in related video genres. As T=1 stands for the emotion type of *neutral* (no emotion), we don't test its accuracy for simplicity.

From Table 2, it is clear to see that *horror* videos get the highest emotion type accuracy, *comedy* and *tragedy* lowest, and that the emotion intensity accuracy of *comedy*, *exciting*, and *sports* is higher. This is reasonable because the emotion types in *comedy* and *tragedy* are more likely expressed by conversations between characters, while the emotion types in *horror* videos are mainly expressed by our selected audio-visual features and the emotion intensities in *comedy*, *exciting*, and *sports* videos are commonly produced by motion and sound. The comparison of our method and [12] is shown in Figure 3, from which we can conclude our performance is better than [12], because we select more affective features, especially visual color and texture features, and facial expression and its intensity, and use HCRFs to capture the latent information of affective features and emotional labels.

Table 2. Emotion type and intensity estimation accuracy on the testing set (%)

	type of emotion						intensity of emotion				
	2	3	4	5	6	7	1	2	3	4	5
comedy	—	—	70.0	—	—	76.5	—	66.6	78.6	78.0	60.0
tragedy	75.0	60.0	60.0	50.0	65.0	73.3	60.0	60.0	77.5	73.3	60.0
action	—	75.0	75.0	70.0	77.5	75.0	70.0	75.0	82.5	80.0	70.0
horror	85.7	—	—	86.2	81.2	—	—	70.0	72.0	70.0	73.3
exciting	—	63.6	61.5	77.7	76.2	72.7	60.0	63.6	71.9	76.2	70.0
sports	—	71.4	66.6	—	81.6	70.6	—	72.7	72.0	82.2	75.0

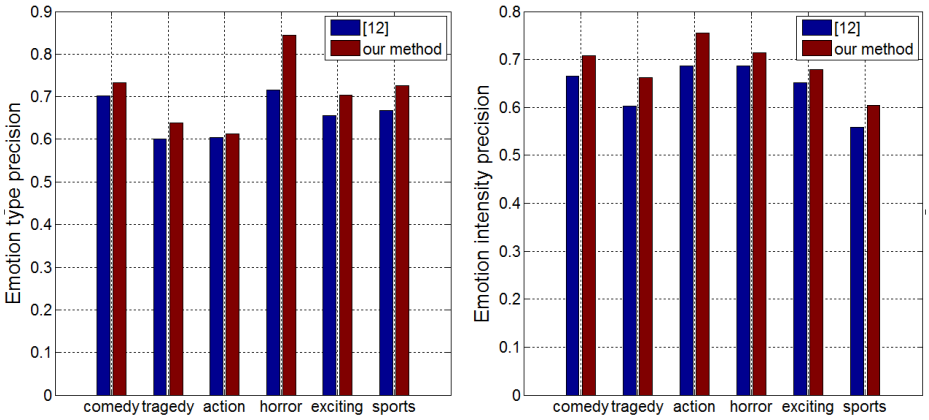
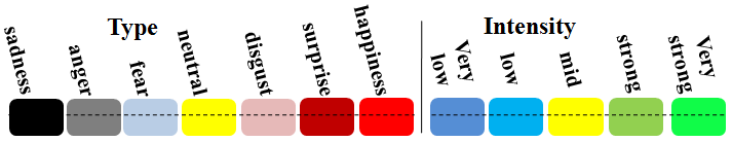


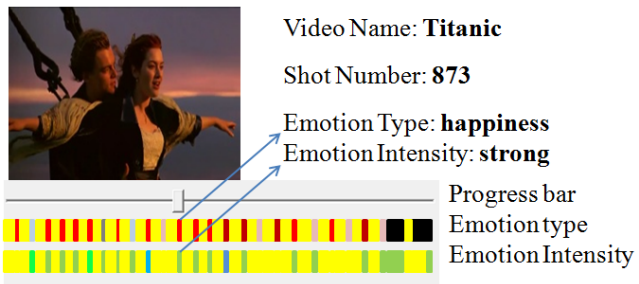
Fig. 3. Comparison of the proposed method and [12] in emotion type and intensity precision

3.3 Affective Visualization

Similar to [12], we use an affective visualization technique to make users understand videos' affective states better. We employ commonly used color to represent affective states on time axis. As we use the proposed IT emotion space, containing 7 emotion types and 5 levels of emotion intensity, we simply use 7 and 5 different colors to represent relevant type and intensity, respectively. A progress bar and two color bars



(a)Different colors for different emotion types and intensities



(b) An instance of affective visualization on time axis

Fig. 4. Illustration of easy-understanding affective visualization results

are shown in Figure 4 to explain the affective states of different contents. Thus, users can browse the affective video content and navigate to their interested shots conveniently. Compared to [12], our visualization is much easier and friendlier for users to understand the affective content, because they can directly see the type and intensity of emotion they are dealing with, instead of the hard-understanding values of VA space or PAD space for the public.

3.4 Results of Affective Presentation

Based on the computed emotion type and intensity labels, we draw affective type and intensity curves and use cut-off lines to segment those shots with high intensities and selected emotion types. Take *Titanic* for instance, the affective type and intensity curves are shown in Figure 5. We use cut off-line i=4 to cut the intensity curve, and get the related key frames and corresponding emotion types, partly shown in Figure 6. Similar result of comedy *Got the Money Anyway* is shown in Figure 7. We can also use emotion type cut off-line to get the shots with the interested moods, such as t=6 in Figure 5.

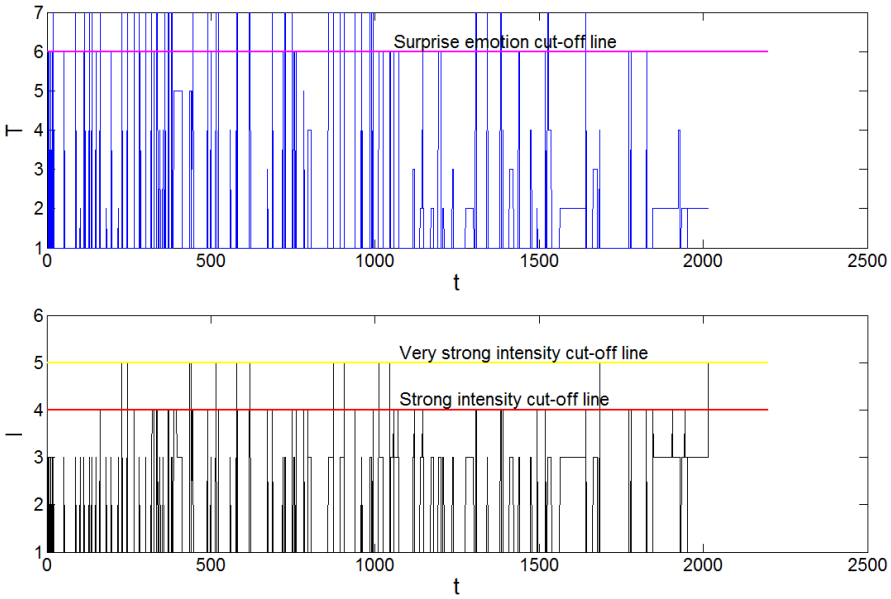


Fig. 5. Illustration of emotion type and intensity time curves and segmentation with cut-off lines. The magenta line in the above figure is the surprise emotion cut-off line. The interaction points of the magenta line and the blue type curve correspond to the shots which evoke surprise emotion in viewers. The yellow and red lines in the bottom figure are the very strong intensity cut-off line and strong intensity cut-off line. The interaction points of the yellow line and the black intensity curve correspond to the shots which evoke very strong emotion intensities in viewers. The interaction points of the red line and the black intensity curve correspond to the shots which evoke strong emotion intensities in viewers.

Fig. 6. Examples of some affective key frames of tragedy <Titanic>, and related emotion types

Fig. 7. Examples of some affective key frames of comedy <Got the Money Anyway>, and related emotion types

In order to evaluate the effectiveness of the proposed method, we use the subject agreement (SAR). We asked the 20 volunteers and another 50 users to watch the constructed video presentations and the provided affective video visualization, and score them with a ten-level Likert scale (1-10, 1: not satisfied at all; 10: very satisfied). Then we compute the average score of the two groups respectively. The satisfaction score are 8.86 and 8.25. Through the results of user study, we can see that the proposed presentation method is able to satisfy users' demand.

4 Conclusion and Future Work

In this paper, we propose a novel general video presentation method based on affective content analysis. According to computed affective labels of emotion type and intensity in the proposed easy-understanding emotion space, we draw affective curves which tell the types and intensities of emotion changes and transitions. With

the curves and related affective visualization techniques, we select the most affective shots and concatenate them to construct a flexible affective video presentation according to time order. The presentation results are satisfying.

As viewers of different ages and different races differ in affective semantics understanding, how to present videos affectively according to viewers' experiences is worth studying. How to tackle the case that different viewers may have different emotion types and intensities is a challenging topic. Maybe assigning one shot with different emotion types and intensities in an emotion vector is an appropriate answer. Combining video's affective contents and viewers' affective reactions to them effectively can present better. Further, how to map low-level affective audio-visual features into affective states more precisely (such as affective audio-visual words and latent topic driving model [22]) is also our future research task.

Acknowledgments. The work was supported by the National Natural Science Foundation of China (No. 61071180) and Key Program (No.61133003).

References

1. Liu, C., Huang, Q., Jiang, S., et al.: A framework for flexible summarization of racquet sports video using multiple modalities. Computer Vision and Image Understanding 113(3), 415–424 (2009)
2. Zhao, Z., Jiang, S., Huang, Q., Zhu, G.: Highlight summarization in sports video based on replay detection. In: IEEE International Conference on Multimedia & Expo, pp. 1613–1616 (2006)
3. Xiang, X., Kankanhalli, M.: Affect-based adaptive presentation of home videos. In: ACM Multimedia, pp. 553–562 (2011)
4. Wang, M., Hong, R., Li, G., Zha, Z., Yan, S., Chua, T.: Event Driven Web Video Summarization by Tag Localization and Key-Shot Identification. IEEE Transactions on Multimedia 14(4), 975–985 (2012)
5. Wang, M., Hong, R., Yuan, X., Yan, S., Chua, T.: Movie2Comics: Towards a Lively Video Content Presentation. IEEE Transactions on Multimedia 14(3), 858–870 (2012)
6. Machajdik, J., Hanbury, A.: Affective image classification using features inspired by psychology and art theory. In: ACM Multimedia, pp. 83–92 (2010)
7. Kang, H.: Affective Content Detection Using HMMs. In: ACM Multimedia, pp. 259–262 (2003)
8. Xu, M., Jin, J., Luo, S.: Hierarchical Movie Affective Content Analysis Based on Arousal and Valence Features. In: ACM Multimedia, pp. 677–680 (2008)
9. Schlosberg, H.: Three Dimensions of Emotion. Psychological Review 61(2), 81–88 (1954)
10. Hanjalic, A., Xu, L.: Affective Video Content Representation and Modeling. IEEE Transactions on Multimedia 7(1), 143–154 (2005)
11. Arifin, S., Cheung, P.Y.K.: A Computation Method for Video Segmentation Utilizing the Pleasure-Arousal-Dominance Emotional Information. In: ACM Multimedia, pp. 68–77 (2007)
12. Zhang, S., Tian, Q., Huang, Q., Gao, W., Li, S.: Utilizing affective analysis for efficient movie browsing. In: IEEE International Conference on Image Processing, pp. 1853–1856 (2009)

13. Nicolaou, M.A., Gunes, H., Pantic, M.: A Multi-layer Hybrid Framework for Dimensional Emotion Classification. In: ACM Multimedia, pp. 933–936 (2011)
14. Ma, Y.-F., Lu, L., Zhang, H.-J., Li, M.: A User Attention Model for Video Summarization. In: ACM Multimedia (2002)
15. Joho, H., Staiano, J., Sebe, N., Jose, J.M.: Looking at the viewer: analysing facial activity to detect personal highlights of multimedia contents. Multimedia Tools Application 51(2), 505–523 (2011)
16. Joho, H., Jose, J.M., Valenti, R., Sebe, N.: Exploiting facial expressions for affective video summarization. In: ACM International Conference on Image and Video Retrieval (2009)
17. Zhao, S., Yao, H., Sun, X., Xu, P., Liu, X., Ji, R.: Video Indexing and Recommendation Based on Affective Analysis of Viewers. In: ACM Multimedia, pp. 1473–1476 (2011)
18. Zhao, S., Yao, H., Sun, X.: Video Classification and Recommendation Based on Affective Analysis of Viewers. Neurocomputing (to appear, 2012)
19. Yang, P., Liu, Q., Metaxas, D.N.: RankBoost with l_1 regularization for Facial Expression Recognition and Intensity Estimation. In: IEEE International Conference on Computer Vision, pp. 1018–1025 (2009)
20. Lafferty, J., McCallum, A., Pereira, F.: Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In: IEEE International Conference on Machine Learning (2001)
21. Quattoni, A., Collins, M., Darrell, T.: Conditional random fields for object recognition. In: Neural Information Processing Systems (2004)
22. Irie, G., Satou, T., Kojima, A., Yamasaki, T., Aizawa, K.: Affective Audio-Visual Words and Latent Topic Driving Model for Realizing Movie Affective Scene Classification. IEEE Transactions on Multimedia 16(2), 523–535 (2010)