HIGHLIGHT RANKING FOR RACQUET SPORTS VIDEO IN USER ATTENTION SUBSPACES BASED ON RELEVANCE FEEDBACK

Yijia Zheng[1]*, Guangyu Zhu*[2]*, Shuqiang Jiang*[3]*, Qingming Huang*[1, 3] *and Wen Gao*[4]

[1]Graduate School of the Chinese Academy of Sciences, Beijing, China
[2]School of Computer Science, Harbin Institute of Technology, Harbin, China
[3]Digital Media Lab, Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China
[4]School of Electronics Engineering and Computer Science, Peking University
{yjzheng,gyzhu,sqjiang,qmhuang,wgao}@jdl.ac.cn

ABSTRACT

In this paper, we propose a method to rank the highlights of broadcast racquet sports videos. Compared with previous work, we integrate relevance feedback into highlight ranking framework to effectively capture the user's interest in attention subspaces and generate personalized ranking result. First, we establish three user attention subspaces and extract audio, visual, temporal affective features to represent the human perception of highlight in each subspace. Then, the highlight ranking models are constructed using support vector regression (SVR) for the three subspaces respectively. Finally, the three submodels are linearly combined to generate the final ranking model. Relevance feedback technique is employed to adjust the weights of each submodel to obtain the result which is suitable to the user's preference. Experimental results demonstrate our approach is effective.

1. INTRODUCTION

Racquet sport is one of the most popular games which has huge numbers of audiences. It is significative to extract the exciting events from racquet sports video and rank them by impressive degree to save both the audience's browsing time and the download cost. Therefore, highlight ranking is an important research topic in sports video analysis field.

As an application of affective computing in sports video, existing research on highlight ranking is not much. Hanjalic [1] linearly combined three excitement components (motion activity, density of cuts and sound energy) to establish an excitement time curve for highlight modeling. Xiong et al. [2] formed an average relative entropy curve by fusing audio and motion. Highlights are defined as the local maximum of the excitement time curve. But it is not explicit to select the excitement components (affective features) and to tell to what extent the highlights reflect human perception. Xing et al. [3] presented a solution to analyze the racquet sports video highlights and proposed a subjective criterion to measure the performance of automatic highlight evaluation. However, the affective features they used were stressed on audio modality while human perception of highlights lies in multiple modalities in the video. Tong et al. [4] proposed a highlight ranking method in soccer games for video browsing. This method mainly based on the analysis of field games such as soccer. The extracted features and evaluation criterion can not be easily extended to racquet games.

Although a few approaches have been investigated on highlight ranking of racquet sports video, there is still no personalized scheme that has the ability of online learning and adjusting the ranking performance according to the requirements of different users. In order to solve this problem, we propose a novel highlight ranking scheme for racquet sports video based on relevance feedback in user attention subspaces. The highlights are ranked according to their impressive confidences and user's feedback by fully exploring their characteristics on audio, visual and temporal attention subspaces to make the ranking result more suitable for human personalization.

The rest of the paper is organized as follows. Section 2 introduces the overview of proposed approach. User attention subspaces partition and representation is presented in Section 3. Section 4 details highlight ranking model construction for each user attention subspace. The algorithm of highlight ranking based on relevance feedback is described in Section 5. Experimental results are reported in Section 6, conclusions and acknowledgement are drawn in Section 7 and 8.

2. OVERVIEW OF PROPOSED APPROACH

The framework of our proposed approach is shown in Fig. 1, which consists of three major models 1) user attention subspaces partition and representation, 2) highlight ranking model construction and 3) user relevance feedback.

Firstly, we predefine three user attention subspaces in terms of visual, audio and temporal modalities according to human perception of highlights. Proper visual, audio and

This work was supported by National High-Tech Research and Development Program (863 Program): 2006AA01Z117

temporal features are extracted to represent the affective characteristics of each subspace. Then, corresponding highlight ranking model is constructed for each subspace using SVR. Finally, linear combination of three models with proper weights is exploited to establish the final ranking model. Relevance feedback is employed to adjust each model's weight in order to effectively capture the user attention region in the predefined subspaces.

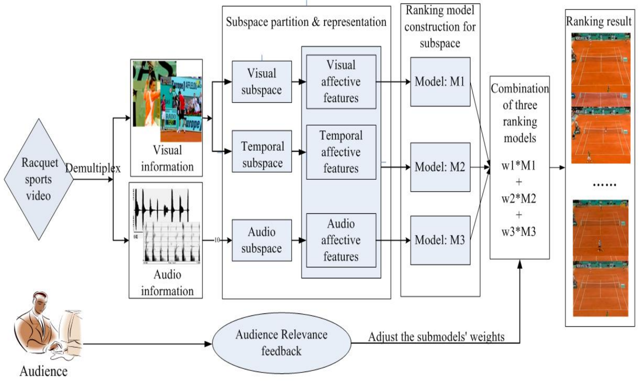
Subspace partition & representation
Ranking model construction for subspace
Ranking result
Racquet sports video
Demultiplex
Visual information
Audio information
Visual subspace
Temporal subspace
Audio subspace
Visual affective features
Temporal affective features
Audio affective features
Model: M1
Model: M2
Model: M3
Combination of three ranking models
w1*M1
+
w2*M2
+
w3*M3
......
Audience
Audience Relevance feedback
Adjust the submodels' weights

Fig 1. System Framework

3. USER ATTENTION SUBSPACES PARTITION AND REPRESENTATION

Human attention of events can be affected by various factors such as audio, visual and environment status [5]. Since the content of video is intrinsically multimodal, human's impression of highlights is affected by multimodal modalities. As the general experience for digital multimedia broadcasting, user's attention region may be different form person to person for one highlight segment. Some may consider the longest duration of highlight as the most impressive content. Others may choose the segment that has the loudest sound in the video. In our approach, we model such multimodal perception as user attention space which has different subspaces in terms of visual, audio and temporal modalities in broadcast sports video.

Treisman [5] demonstrated that user's perception of highlight impression is influenced by various factors coming from different information source. Visual, audio and temporal information are basic modalities in sports video that have significant influence on user's perception of highlights. In our approach, we thus predefine the user attention space based on three subspaces which are visual subspace (VSS), audio subspace (ASS) and temporal subspace (TSS) to express the user's comprehension on different aspects of highlight.

To effectively model the user attention subspaces, we extract affective features from sports video as the representation for each subspace. Motion as an effective visual feature can reflect the exciting degree of an event. Camera and player motion such as action and trajectory can be used in practice. We exploit MPEG-7 motion activity descriptor [3] and player direction switching rate [6] to represent the visual subspace in terms of camera motion and player trajectory respectively. The higher the direction switching rate and the larger the MPEG-7 motion activity descriptor, the more exciting the highlight. Actually, direction switching rate can also be replaced with swing switching rate [6] extracted from player action. Audio energy and pitch-related features are commonly used audio affective features in sports video [1-3]. Usually the higher the average energy of cheering and the speech's average pitch are, the more exciting the event indicates. The highlight duration is another affective feature that can reflect the exciting degree of highlights. Taking tennis for an example, a longer rally scene shows the players' contest is more exciting. This feature can be changed into the variation of shot length in field games such as soccer [4].

We extract seven affective features totally for attention subspaces representation including average MPEG-7 motion vector (AMMV), direction switching rate (DSR), average cheer energy (ACE), average pitch of excited speech (APES), highlight duration (HD), cheer duration (CD) and excited speech duration (ESD). These seven features are proved to be effective for affective video content analysis in [1-4], [6]. Using these affective features, the user attention subspaces are represented as shown in Table 1.

Table 1 User Attention Subspaces

Visual Subspace (VSS): {AMMV; DSR}
Audio Subspace (ASS): {ACE, APES}
Temporal Subspace (TSS): {HD, CD, ESD}

4. HIGHLIIGHT RANKING MODEL CONSTRUCTION FOR ATTENTION SUBSPACES

According to the predefined user attention subspaces, three ranking models are constructed respectively using SVR. They are visual submodel (VSM), audio submodel (ASM) and temporal submodel (TSM). The three models reflect users' different preference in attention subspaces. Our final highlight ranking model is the linear combination of three submodels which is called combined model hereinafter.

SVR is a nonlinear technique which has the advantage of requiring fewer training samples and having better generalization ability. It provides superior robustness and prediction accuracy for sparse and nonlinear data distribution [7]. The input of SVR model is the concatenation of extracted affective features and the output is the estimation of impressive value for highlight segments by computer.

5. RELEVANCE FEEDBACK FOR HIGHLIGHT RANKING IN USER ATTENTION SUBSPACES

To realize the personalized ranking performance according to user preference, we exploit the relevance feedback technique to capture the user's interest in three attention subspaces. According to [8] the relevance feedback mechanism can make the system understand the retrieval purpose of users, and find the most satisfied result according to user's individual requirement. Similarly it can help the ranking system to capture the user interest region in attention space effectively.

We define the final ranking model as the linear combination of models constructed on three attention subspaces:

$$M_{rally}(s) = w_{VSM} \cdot M_{VSM}(s) + w_{ASM} \cdot M_{ASM}(s) + w_{TSM} \cdot M_{TSM}(s) \qquad (1)$$

where $M_{VSM}(s)$, $M_{ASM}(s)$, $M_{TSM}(s)$ represent the ranking models for three attention subspaces and w_{VSM} , w_{ASM} , w_{TSM} are the corresponding weights respectively. $M_{rally}(s)$ is the exciting degree automatically estimated by computer with the reference of user's feedback. The detail of ranking process based on relevance feedback is described in Algorithm 1.

Using our approach, we can obtain a weight set $\{w_{VSM}, w_{ASM}, w_{TSM}\}$ for each user, which reflects his/her own perception of highlight impression in the attention subspaces. Then we can provide the user with the most exciting segments according to his/her individual preference. This method is different from the existing stereotyped video summarization technique, for it has the advantage of personalized adaptation and online learning ability.

6. EXPERIMENTS

As tennis is one of the most representative racquet sports, we conducted experiments on five video clips extracted from five different tennis matches of French Open 2005. The detail of experimental data is listed in Table 2.

Table 2. Videos for Experiment

Video	1	2	3	4	5
duration	08:52	12:59	26:52	26:32	01:38:35
highlights	21	31	53	59	209

To conveniently conduct the experiments we develop a prototype for highlight ranking and feedback interaction. All the highlight segments in tennis videos are detected using dominant color method [6]. To subjectively evaluate the highlights and obtain the ground truth, we designed a program for manual highlight confidence labeling. The scores of highlights are limited within the interval between 0.1 to 1.0. The more interesting a highlight is, the higher score it will be assigned to. We invited four subjects who have rich experience in tennis to independently score the impression according to their own understanding. The ground truth of the degree for a highlight segment is the average of all subjective scores. In our experiment, we randomly selected one clip to train the ranking model, other clips were used as test data. We conducted two contrastive experiments to evaluate the performance of our proposed approach.

To compare our approach with existing methods, we first conducted the ranking performance using a single model whose input contains all the affective features. We set M = 20 and *conf* = 0.7 in Algorithm 1 to obtain the top 20 most exciting highlight segments whose impressive confidence were above 0.7. The average *RA* of single ranking model is RA_1 = 85% for our four specific subjects. *RA* is the ranking accuracy defined in Algorithm 1, which represents the percentage of the highlight segments satisfied by user in the result set. We calculate the difference between ground truth and the estimation results obtained by our approach. The difference is defined as follows:

```
Step 1: Evaluate the impressive confidence for each highlight segment using Eq. (1) with the initial weights {1/3, 1/3, 1/3} and sort the highlights in descent order according to their exciting confidence.
Step 2: Select the top $M$ highlights whose impressive confidence are above $conf$ as the initial return set $HI = \{s_1, \ldots, s_M\}$ where $s_i$ is the selected return highlight segment. $1 \le i \le M$ , $M$ and $conf$ are inputted by user. Generate the user satisfied result $HS = \{s_1^{'}, \ldots, s_N^{'}\}$ where $s_i^{'} \in HI$ and $N$ is the number of user satisfied segments. Calculate the ranking accuracy $RA = N / M \times 100\%$ .
Step 3: Use VSM with the extracted visual affective features to evaluate $s_i \in HI$ respectively. Generate the highlight set $VHS = \{s_1^V, \ldots, s_P^V\}$ where $s_i^V \in HS$ and $M_{VSM}(s_i^V) \ge conf$ .
Step 4: Repeat Step 3 using ASM with audio affective features and TSM with temporal affective features to generate the sets $AHS = \{s_1^A, \ldots, s_Q^A\}$ and $THS = \{s_1^T, \ldots, s_R^T\}$ respectively.
Step 5: Adjust the weights of three models as $\{P/(P+Q+R), Q/(P+Q+R), R/(P+Q+R)\}$ and repeat Step 1 with the adjusted weights. Repeat Step 2 to calculate the new ranking accuracy $RA^{'}$ .
Step 6: If $|RA^{'} - RA| \le thres$ or user is satisfied, stop. Else go to Step 3.
```

Algorithm 1. Highlight ranking based on attention subspaces and relevance feedback.

$$dif = \sqrt{[\sum_{i=1}^{M} (s_{ai} - s_{gi})^2] / M} \qquad (2)$$

where s_{ai} is the ranking confidence of the *i*th highlight estimated by computer and s_{gi} is the corresponding ground truth. The average difference is dif_1 = 14.4% for single model calculated from our four subjects. The *dif* reflects the perception difference between computer automatic estimation and user subjective evaluation.

Then we employed the proposed approach based on user attention subspaces and relevance feedback with the initial weights $\{1/3, 1/3, 1/3\}$ for automatic highlight ranking for the first time. Similar to the previous test, the top 20 highlights whose ranking confidence was above 0.7 were output. Then the users selected the segments they satisfied. The obtained results without feedback for the combined model are RA_2= 75% and dif_2= 15.0%. The *dif curves* for single model and combined model contrasted with the corresponding ground truth are illustrated in Fig 2.

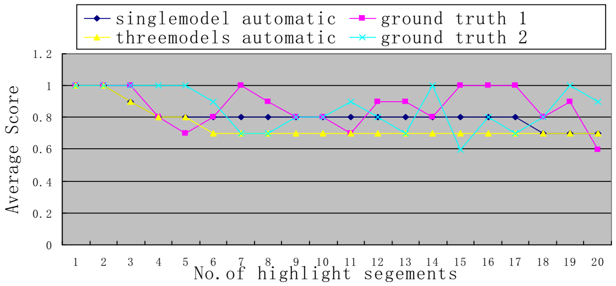

Fig 2. Comparative ranking results between single model and combined model without feedback interaction

In the second contrastive experiment we involved relevance feedback interaction. Each subject only needs to return the highlights he/she is satisfied to the system. The system adjusts the weights of submodels to provide a new round of top 20 highlights and their new confidence to subjects. After four times' feedback the system achieves the average RA_3= 95% and dif_3= 12.2%. The average *dif curve* of the fourth feedback interaction is shown in Fig 3 contrasted with the corresponding ground truth and the *feedback times vs. ranking accuracy curve* is shown in Fig 4.

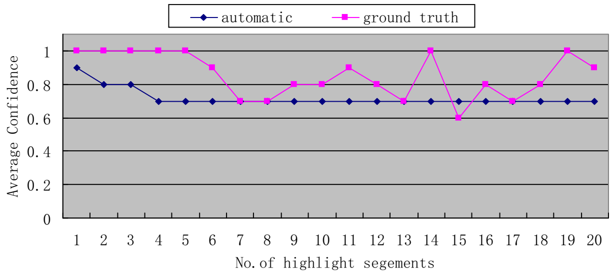

Fig 3. Combined model evaluation using relevance feedback

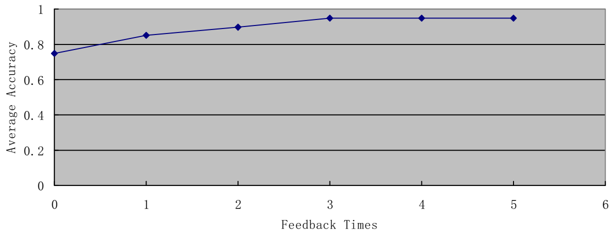

Fig 4. Feedback times---ranking accuracy curve

Based on the above experiments, we can see that the results of the combined ranking model without feedback interaction are not as well as the ones of single ranking model (RA_1>RA_2 and dif_1<dif_2). However, the results of combined model become more consistent with human perception than single model after some feedback interactions (RA_1<RA_3 and dif_1>dif_3). For our specific four subjects involved in the experiments, the average *RA* achieved 95% after four times feedback interaction. This demonstrates the relevance feedback effectively capture the user interest in the attention subspaces. The final weights set $\{w_{VAM}, w_{ASM}, w_{TSM}\}$ can be considered as the parameter of the personalized retrieval for the specific user.

7. CONCLUSION

This paper has presented a novel approach for highlight ranking which is different from previous work. We present the definition of user attention subspaces and employs relevance feedback method to capture the user interest region in attention space to make the ranking result more conforming to users' perception.

In future we will investigate more valuable affective features of the user attention subspaces and incorporate other sophisticated methods for highlight evaluation. The ranking work can be extended to other sports genres. After highlight ranking, users can perform hierarchical browsing according to their requirements, the browsing priority can be freely chosen by users according to their preference.

8. ACKNOWLEDGEMENT

This work was supported in part by Science100 Plan of Chinese Academy of Sciences: 99T3002T03, Beijing Natural Science Foundation: 4063041 and National 242 Project: 2006A09.

9. REFERENCES

[1] A. Hanjalic, "Generic approach to highlights extraction from a sports video," ICIP, vol. 1, pp. 1-4, 2003.

[2] Z. Xiong, R. Radhakrishnan, A. Divakaran, "Generation of Sports Highlights Using Motion Activity in Combination with A Common Audio Feature Extraction Framework," ICIP, vol. 1, pp. 5-8, 2003.

[3] L. Xing, H. Yu, Q. Huang, Q. Ye, A. Divakaran. "Subjective Evaluation Criterion for Selecting Affective Features and Modeling Highlights," Proceedings of SPIE, Vol. SPIE-6073, pp. 188-195, 2006

[4] X. Tong, Q. Lu, Y. Zhang, H. Lu, "Highlight Ranking for Sports Video Browsing," Proc. of ACM International Conference on Multimedia, pp. 519-522, 2005.

[5] AM. Treisman, G. Gelande, "A Feature-Integration Theory of Attention," Cognitive Psychology 12, pp. 97-136, 1980.

[6] G. Zhu, C. Xu, Q. Huang, W. Gao and L. Xing, "Player Action Recognition in Broadcast Tennis Video with Applications to Semantic Analysis of Sports Game," roc. of ACM International Conference on Multimedia, pp. 431-440, 2006.

[7] V. Cherkassky, Y. Ma, "Selecting of the loss function for robust linear regression," Neural Computation, 2002.

[8] Y. Rui, T.S. Huang, M. Ortega and S. Mehrotra, "Relevance feedback: A power tool for interactive content-based image retrieval," IEEE Trans. Circuits and Systems for Video Technology, vol. 8, no.5, pp. 644-655, 1998.