HIGHLIGHT SUMMARIZATION IN SPORTS VIDEO BASED ON REPLAY DETECTION

Zhao Zhao[1], Shuqiang Jiang[1], Qingming Huang[1], Guangyu Zhu[2]

[1]Inst. of Computing Technology, Chinese Academy of Sciences, Beijing 100080, China
[2]School of Computer Science and Technology, Harbin Institute of Technology, Harbin 150001, China
{zzhao, sqjiang, qmhuang, gyzhu}@jdl.ac.cn

ABSTRACT

Highlight summarization technology has been studied widely in sports video analysis. In this paper, we propose a highlight summarization system based on replays. First the replay clips in the sports video are extracted as the highlight candidates. Then the features including audio energy and motion activity are employed to rank the arousal level of the replay clips. Finally we model the highlight with the arousal rank to generate summarization. The contribution of this paper concentrates on two aspects. Firstly, Event-Replay (ER) structure is proposed and some new features are employed to represent the arousal levels of ER for general sports video. Secondly a novel highlight model is proposed considering the inter-relation of ERs. The experiments evaluate the rationality of the system.

1. INTRODUCTION

With the increasing amount of multimedia content, people may not have enough time to browse all the details of the video content. It necessitates the development of summarization technology to offer users with a brief glance to the significant part of it. Among those technologies, the technology of sports video summarization is widely investigated since sports game holds a large amounts of audiences worldwide. As an important compiled clue, the replay shows the details of an important video segment with a slower speed. Thus it is widely employed in sports video analysis especially for event detection, highlight summarization etc. [1,2,3].

To detect replays from sports video, some early works focus on the characteristics of them such as motion vector [4] and replay structures [1]. However, these methods are not robust enough to be suitable for various kinds of sports video replay detection because replays in different sports video are various and compiled in different manners and can hardly be represented by such simple features. Therefore the recent approach is to detect the accompanying logo effect of the replays in sports videos to acquire the replay segmentations [5,6]. Since the replays have some semantic meanings, former works also focus on the event classification [2] and highlight ranking [3] based on replays. The authors of [3] studied the highlight ranking technology in soccer video using some specific static objects such as goalmouth, goal-net etc. However, little work concerns about a general highlight summarization model based on replays or considers the inter-relations among highlight segmentations. In past works, most highlight ranking methods directly select some features to rank the highlights discretely without considering the relationship among the events [8,9]. This strategy is often not semantically proper enough to represent the user's desire. In fact, users may want to require the "run" parts of the game which usually consist of a series of exciting events occurring densely in a comparatively short time. Thus we should consider the interactions between exciting events in order to find the "run" part to generate the highlight summarization. This highlight structure can be depicted as Fig. 1. The yellow sets are what we call "run" parts. They consist of several consecutive replays. The arrays in Fig. 1 represent the time axis. Our highlight summarization system is based on this structure.

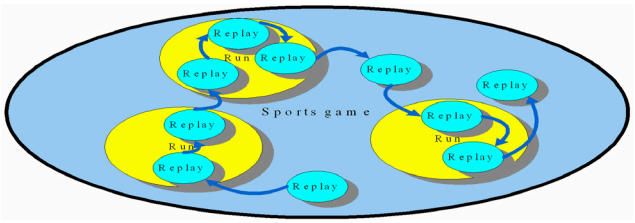


Figure 1: Highlight structure

In this paper we use the phrase "arousal" [10] to represent the exciting levels of the discrete events in order to discriminate with "highlight" which involves with the inter-relations among exciting events.

This paper contributes in two aspects. Firstly, Event-Replay (ER) structure is proposed and some new features are employed to represent the arousal levels of ER for general sports video thus make the system suitable for a general utilization. Secondly considering of the interactions among ERs we model the highlight and summarize it with the "run" segments.

The paper is organized as follows. Section 2 gives an approach based on logo detection to effectively detect replays. Then in Section 3 two affect features are employed to calculate the arousal level of the replays and then the highlight model is established. The experiment in Section 4 shows that the features being used can represent the subjective arousal level and the "run" segments can be

extracted effectively according to user's demand. At last Section 5 concludes this paper.

2. REPLAY DETECTION

Generally, a replay in a broadcast sports video is often accompanied with a pair of logo transitions which sweep off at the beginning and the end of it. In other words, if all logo transitions in a video are detected, the replay clips can be annotated effectively. Based on these observations, we first automatically generate the logo template from the video. Then all the logos are detected with the logo template. At last we pair the logos according to the usual length of the replays in order to detect them.

2.1. logo template generation

A logo transition is always a sweeping effect between two shots. Our algorithm is to detect the sweeping effect frames and classifies them into several clusters. Then the logo cluster is selected according to a judging criterion. The mean image of the frames in the logo cluster is set to be the logo template.

2.1.1. Detecting the sweeping effect

A. Compute the frame-to-frame difference.

B. If the difference exceeds threshold value h_1, preserve the frame.

C. Reject the frames which are non-consecutive in the preserved frame sequence or the length of the consecutive frames exceeds threshold value h_2.

Thus the sweeping effect frames set S_w is obtained.

2.1.2. Clustering

We utilize two features to calculate the distances between an arbitrary couple of frames in S_w. The features are calculated as follows:

A. Color histogram

The color histogram of a frame is computed in the HSV (Hue-Saturation-Value) color space with 256 bins quantization. The color histogram of frame *i* is denoted as H_i. Then we compute the color histogram distance of a couple of arbitrary frames in S_w. Here we use Euclidean distance. The distance between frame *i* and *j* is denoted as d^C_{ij}.

B. Edge

We first use edge detection algorithm to transit frame *i* to edge image. Then the edge image *i* is divided into small blocks(b^i_1, b^i_2,…). We count the number of edge pixels in a block. Thus we obtain the Block-Edge Histogram (BEH) E_i which can be depicted in Fig. 2.

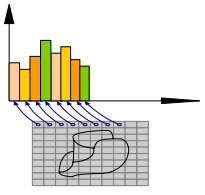


Figure 2: BEH

The edge distance between frame *i* and j is denoted as d^E_{ij}.

Then the total distance between frame *i* and *j* is:

$$d_{ij} = \omega_C d^C_{ij} + \omega_E d^E_{ij} \quad (1)$$

where ω_C, ω_E are the weights of color and edge distances respectively.

Then we obtain the distance matrix of the frames in S_w.

$$D_S = \{d_{ij}\} \quad (2)$$

After the distance matrix is acquired, we cluster the frames in S_w whose distances between them is smaller than a threshold value d_v. Then we obtain an initial cluster set including *N* number of classes of frames in S_w:

$$C = \{C_i \mid i = 1,2...N\} \quad (3)$$

Then we filter out the logo cluster C_i from C. It's apparent that the frames in the logo cluster C_i dispersedly distribute on the time axis. But the frames in other clusters usually consecutively distribute. So we define a scattered criterion of Cluster *i* as:

$$F_i = \log \sum_m (T^m_i - M_i)^2 / N_i \quad (4)$$

here T^m_i is the frame number, M_i is the mean value of the frame serial number in cluster *i*, N_i is the frame number of cluster i. We choose the cluster which has the smallest criterion F to be the logo template class.

Then the logo template could be computed as:

$$f_{logo} = \sum_m f^m_i / N_i \quad (5)$$

2.2. logo detection

In this process, we compute the distance between logo template and an arbitrary frame in the video. If the distance is smaller than a pre-defined threshold, it is recognized as a logo. The distance between logo template and a frame is depicted as Equation (1) with the same features.

2.3. replay recognition

After all logos are detected, we should pair them as the boundaries of replays and eliminate the wrong logos.

We test the temporal duration of video clips between logos one by one. If the temporal duration of a video clip is within the range of (T_l, T_h), the video clip is detected as a replay. T_l and T_h are empirically determined based on the fact that a replay usually does not last too long or too short.

The replay detection approach proposed above can detect replays in sports video effectively and rapidly which can be utilized in the next stage to summarize highlight.

3. HIGHLIGHT SUMMARIZATION

There is no doubt that the replay segmentation represents the most attracting events in a sports video. They can be a goal kick or a foul in a soccer game, a dunk shot in a basketball game or a four bagger in a baseball game. So we

obtain an important clue to the highlight as soon as all the replays are detected in a sports video. But users' demands are often multifarious and just browsing a replay sequence of a sports game normally could not satisfy their requirement. They may want to order for the "run" segments of the game. These "run" parts are often sets of exciting events which are temporally near and have short intervals between them as illustrated in Fig.1. So we not only rank the arousal level of the replays but also combine them into the semantic "run" parts. In this section, Event-Replay structure is first introduced. Then arousal levels of ERs are ranked and finally highlight summarization based on "run" part is extracted.

3.1. Replay arousal level ranking

3.1.1. The Event-Replay structure(ER)

Most of the exciting events in sports video are often compiled in the structure illustrated in Fig. 3. The first node is the broadcast of the event in a normal speed. It followed by some specific close-up shots of the key players, coaches, audiences called "link node". The last node is the slow motion replay of the event. The structure is called an Event-Replay structure(ER).

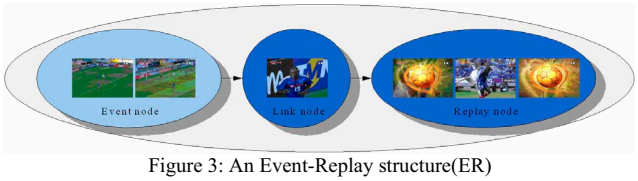


Figure 3: An Event-Replay structure(ER)

3.2.2. Arousal level calculating

Actually the highlight information is included in the event node of ER. It is observed that the link frames of ER are often close-up shots. So we first acquire a GMM model of the dominant color of the field in the game. Then, by judging the dominant color ratio of the key frames of the shots before a replay, event shots are determined. Finally we employ two fundamental features of the event node to calculate the arousal level of the ER in a simple linear model. The features selected are suitable to represent arousal level [10] of the sports videos.

1) The first feature is the average audio energy of a specific length T_a before the replay. It is reasonable because the audience often applause after an exciting event and the sportscaster usually comment in an agitated tone. The audio energy of replay *i* can be represented as:

$$A_i = \sum_m E(m) / T_a \tag{6}$$

Here $E(m)$ is the short time audio energy.

2) The second feature is the motion activity of the event which the replay represents. It is reasonable because a goal kick or a four beggar which has a higher arousal rank is usually accompanied with a more intense camera motion or a more rapid movement of the objects in the frames than events not so excited such as fouls.

Thus we calculate the global motion parameters of replay *i* to represent the motion activity M_i:

$$M_i = \sum_i \sum_j \alpha_j |G_i(j)| \tag{7}$$

Here $G_i(j)$ is the global motion parameter including pan, tilt, zoom, etc.

After these two features are calculated, the arousal level of replay i is denoted as:

$$R_i = \omega_A A_i + \omega_M M_i \tag{8}$$

Here both A_i and M_i has been normalized to the [0, 1] interval.

3.2. Highlight summarization

Considering the interactive effect between the exciting events, we can summarize the highlights of sports video.

The interactive structure can be depicted as quadruples:

$$I = \{ER, D, R, A\} \tag{9}$$

1) ER is the ERs in the video.

2) $D = ER \times ER$ represents the influences among ERs. The influence between ER_i and ER_j is depicted as:

$$d_{ij} = K / \sqrt{|t_i - t_j|} \tag{10}$$

Here K is a constant, t_i and t_j are the ending times of ER_i and ER_j . Apparently D is a symmetry matrix.

3) R is the arousal level of the ERs which is computed in equation (8).

4) A is the highlight mark of the ERs considering the interactions between ERs.

$$a_i = \sum_j d_{ij} R_i \tag{11}$$

Thus we can plot the highlight curve and finally summarize the video.

4. EXPERIMENTS

To evaluate the performance of the proposed system, we employed two sports videos for experiment. One is soccer game and the other is hockey game. Since highlight may be subjective to different people, we invite 4 observers to mark arousal level of the ERs in the videos independently. The mean values of the ranking results are considered as the ground truths of the arousal level which are normalized in the range of [0, 1] in order to be compared with the results computed by our proposed model. As aforementioned that the highlight ranking is various according to different observers, the similarity criteria are emphasized on the highlights that are generally acknowledged. Thus the ERs which have a larger value of arousal level in the ground truth are assigned to larger weights such as a goal kick in soccer or a penalty corner in a hockey game.

For that reason, the similarity criterion is defined as follows:

$$S = (1 - \frac{\sum_i R_i^g \left| R_i^g - R_i^c \right|}{N})^2 \qquad (12)$$

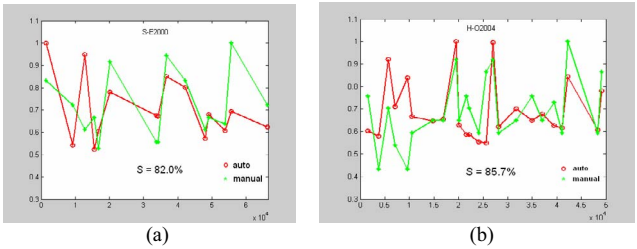
Here R_i^g and R_i^c are the manually marked arousal level ground truth of ER_i and the correspondingly computational value. *N* is the replay numbers.

Test data used in the system are soccer game Greece vs. France in EuroCup 2000 (S-E2000) and the hockey game New Zealand vs. Spain in Olympic Games 2004(H-O2004). Both are selected the first half of the game.

The replay detection results are shown in Table 1. The precision and recall are quite satisfactory and proper for further highlight summarization.

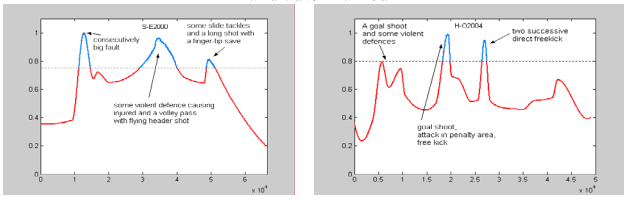
Table 1: Result of replay detection

Videos	Total	Correct	False	Prec. (%)	Recall (%)
S-E2000	19	19	0	100	100
H-O2004	25	24	3	88.9	96.0



(a) (b)

Figure 4: The comparison of arousal level between the manually marked and auto-marked



(a) (b)

Figure 5: Highlight curve considering the ER's interactive influence

Fig. 4 shows the automatically computed highlight rank vs. manually marked highlight rank. (In S-E2000, there are 4 replays that are not satisfied with ER structure and are discarded.) Fig. 5 shows the highlight curve of the sports video. As shown in Fig. 5(b), we select a threshold value and acquire two "run" parts. The consecutive goal shoot, an attack in penalty area and a free kick corporately make up the first "run" part. Two successive direct free kick compose the second "run" part. We can adjust the threshold value to acquire "run" parts and offer to users as the highlight summarization. Extensive experiments show that the results correspond to the human's subjective feeling and the extracted "run" segments are reasonable highlight summarizations.

5. CONCLUSIONS AND DISCUSSIONS

In this paper, we present a novel system of highlight summarization in sports video based on replay detection. The system is general to most genres of sports game because the replay is a reliable clue to the highlight and the features we utilize is not limited in a specific kind of sports game. In the system, we firstly propose an elaborate method to label the replays in the video for further highlight summarization. Then we utilize two basic features to represent the arousal level of the isolated event-replays (ERs) and propose a new method for highlight summarization considering the interactive effect among ERs. The experiments evaluate the rationality of the system. Future work will be emphasized on how to find more effective features to represent the arousal level and should also focus on modeling the highlight considering the interaction of the exciting events.

ACKNOWLEDGEMENTS

This work is partly supported by NEC Research China on "Context-based Multimedia Analysis and Retrieval Program", "Science 100 Plan" of Chinese Academy of Sciences and "Beijing Natural Science Foundation".

6. REFERENCES

[1] H.Pan,P.Beek,M.Sezan, "Detection of slow-motion replay segments in sports video for highlights generation," ICASSP2001, Salt Lake City, UT, May 2001

[2] A.Ekin, A.M.Tekalp, "Automatic soccer video analysis and summarization," Symp.Electronic Imaging:Science and Technology: Storage and Retrieval for Image and Video Databases IV, Jan.2003, CA.

[3] X.Tong, Q.Liu, "Highlight ranking for sports video browsing," Proc. of ACM Multimedia 05.

[4] V.Kobla, D.DeMenthon, D.Doermann, "Detection of slow-motion replay sequences for identifying sports videos," Proceedings of IEEE Third Workshop on Multimedia Signal Processing, pp.135-140, 1999.

[5] X.Tong,H.Lu, "Replay detection in broadcasting sports video," Proceedings of the Third International Conference on Image and Graphics. 2004.

[6] H. Pan, B. Li, and M. I. Sezan, "Automatic detection of replay segments in broadcast sports programs by detection of logos in scene transitions," Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing (ICASSP), 2002.

[7] X.Tong, Q.Liu, "Shot classification in sports video," Proceedings of ICSP 04, pp.1364-1367

[8] Alan Hanjalic,"Generic approach to highlights extraction from a sport video", Proc. IEEE ICIP'03, 14-17 Sept. 2003.

[9] Xiong.Z, Redhakrishnan.R,Divakaran.A, "Generation of sports highlights using motion activity in combination with a common audio extraction framework", Proc.IEEE ICIP Oct. 2003

[10] A.Hanj, L.Xu, "Affective video content representation and modeling," IEEE Tran. on Multimedia, VOL.11