EXTRACTING STORY UNITS IN SPORTS VIDEO BASED ON UNSUPERVISED VIDEO SCENE CLUSTERING

Chunxi Liu[1], *Qingming Huang*[1,2], *Shuqiang Jiang*[2], *Weigang Zhang*[3]

[1]Graduate University of Chinese Academy of Sciences, Beijing 100049, China
[2]Institute of Computing Technology, Chinese Academy of Sciences, Beijing 10080, China
[3]School of Computer, Harbin Institute of Technology at Weihai, Weihai 264209, China
{cxliu, qmhuang, sqjiang, wgzhang}@jdl.ac.cn

ABSTRACT

Many sports videos such as archery, diving and tennis have repetitive structure patterns. They are reliable clues to generate highlights, summarization and automatic annotation. In this paper, we present a novel approach to analyze these structure patterns in sports video to extract story units. First, an unsupervised scene clustering method for sports video is adopted to automatically categorize the video shots into several disparate scenes. Then, the clustering results are modeled by a transition matrix. Finally, the key scene shots are detected to analyze the structure patterns and extract the story units. Experimental results on several types of broadcast sports video demonstrate that our approach is effective.

1. INTRODUCTION

In recent years, with the explosive growing on video data being produced, distributed and made available all over the world, there is an emerging need of effective tools for analyzing, searching, and retrieving video of interest according to video content. As a popular video genre, sports video attracts much attention for its large audience and tremendous commercial potential.

Compared with other videos such as news, movie, and documentary, sports video has its own characteristics. A sports game usually happens in a specified field, with well-defined temporal structures. We can categorize the sports video shots into several dominant scenes according to the space-constrained field of the sports game. For example, there always exist three dominant scenes in table tennis video as shown in Fig. 1: close-up of player *A*, court-view and close-up of player *B*. On the other hand, repetitive patterns can be found in most of the sports video according to their production order or intrinsic rules of games, especially for sports types such as archery, diving, *et al.* These repetitive patterns usually reveal a chain of actions, which compose the story units of the sports video. In this paper, we define story unit as a few consecutive shots to describe a logical meaningful event. It is the same as the scene definition in [11], in which scene is defined as one of the subdivisions of a play as a division of an act presenting continuous action in one place. For example, a story unit in diving video is defined as

the player standing on the dive platform → *the actions of taking off* → *diving* → *entering water*

This reveals the whole action of diving from the beginning to the end. Actually, for the same kind of sports video, story units do not vary significantly from game to game. Based on our observation, the last shot of each story unit plays the key role and always indicates the occurrence of the unit. This shot always comes from one dominant scene cluster and is called as key scene shot here. For example, the key scene shots in archery are the target scenes, and in diving are the enter water scenes. Therefore, if we can extract these key scene shots, the story units could be obtained easily.

Fig. 1. Three main dominant scenes in table tennis

Lots of researches have been carried out for video structure analysis. Statistical models such as the Hidden Markov Models have been used for analysis the structure of the video [1][2][3]. M. Yeung *et al*. [4] use the time constrained clustering and STG to extract story units in long programs. Naoko *et al*. [5] focus on grasping the story from the speech transcript by using a probabilistic framework based on Bayesian networks. A. Divakaran *et al*. [6] try to combine unsupervised and supervised learning to extract sports highlights and other events. Recently, Wang *et al*. [7] use templates based method to detect repetitive patterns in sport video.

In this paper, we propose an efficient story unit extraction method based on the result of unsupervised video scene clustering. In our method, cluster transition matrix is constructed to detect the key scene shots and extract the story unit. Our approach is general for sports video with repetitive patterns and does not need any supervision. Compared with

the work in [7], our results for story unit extraction not only have a repetitive pattern but also are more accordance with human's expectation and demonstrate more meaningful information.

The rest of the paper is organized as follows: section 2 provides an overview of the system. Section 3 discusses the unsupervised sports video scene clustering algorithm. In section 4, the story unit extraction method is described in detail. Experimental results are presented and discussed in section 5. Finally, we conclude the paper in section 6.

2. SYSTEM OVERVIEW

The flowchart of the proposed story unit extraction method in sports video is shown in Fig. 2. First, the video sequence is automatically segmented into shots and five key frames are extracted to represent the content for each shot. Then, the unsupervised video scene clustering algorithm is performed to obtain the video scenes. Finally, story units are extracted based on the proposed model-based approach from the video sequence.

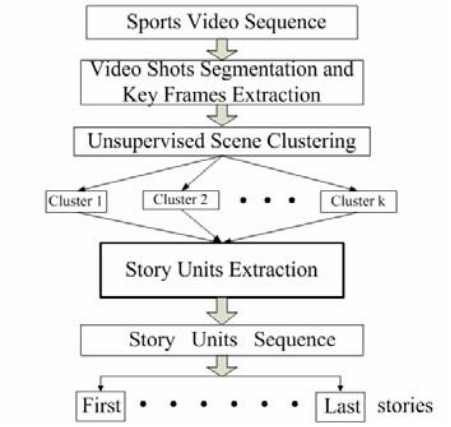

Fig. 2. Flow chart of the story unit extraction system

3. UNSUPERVISED SPORTS VIDEO SCENE CLUSTERING

Such process is achieved by using the unsupervised scene clustering method in [9] with revision. Each shot is represented by 5 key-frames. Color histogram (256 bins) in HSV color space is employed as the low level features of these frames. We use the Euclidian distance between two shots (or scenes) to represent their scene similarity. To determine the stopping point, we define a J value based on Fisher Discriminant Function in (1).

$$J_l = \frac{\sum_{c=0}^{K_l} J_w^c}{J_t} = \frac{\sum_{c=0}^{K_l}\sum_{i=0}^{N_c}\left\|\vec{s}_i^{\,c} - \vec{s}_{mean}^{\,c}\right\|}{\sum_{i=0}^{N}\left\|\vec{s}_i - \vec{s}_{mean}\right\|} \quad (1)$$

where J_t is the inter-cluster scatter of the initial scene sequence, J_w^c is the inter-cluster scatter of the scene cluster C. N is the total scene number in the initial scene sequence, N_c is the shot number of scene cluster C. $\|.\|$ represents the Euclidean distance. $\vec{S}_i^c$ ($\vec{S}_i$) denotes shot i in scene cluster C (the initial scene sequence), and $\vec{S}_{mean}^c$ (S_{mean}) denote the mean feature of shots in scene C (the initial scene sequence).

The value of J_l represents the total scatter of scene cluster, which describes the ratio of intra-cluster scatter to inter-cluster scatter of the scenes in the merging processing. Actually, it is expected that both J_l and the scene number K_l are small. While in real situation, the smaller the scene number K_l is, the larger the J_l value will be. As a tradeoff between J_l and K_l, we choose the point where $J_l + K_l$ is the smallest as the best merging stop point.

In the merging process, instead of clustering the total shots in one iteration, we cluster the first 100 shots by the algorithm described above and assign the rest shots into the initial clusters according to the nearest-neighboring criterion. When both of the following conditions are satisfied, a new cluster can be created. 1) The distance between the shot and the center of its nearest cluster is ten times further than the average point to center distance in this cluster. 2) The shot number of this cluster is bigger than 20. This is suitable because there are repetitive structures in sports video. The clustering result reveals that the revised algorithm is more robust with the increase of the shot number and gives a better result.

4. STORY UNITS EXTRACTION

4.1 Story unit model

Sports video consists of story units with non-story unit in the interval which is illustrated in Fig. 3.

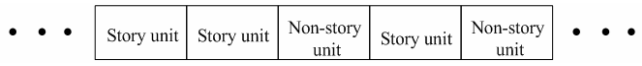

Fig. 3. Sport video in terms of "story unit"

After the clustering procedure, we sort the resulting clusters in a descending order according to their shot number. Each cluster is labeled with the serial number. Then we label each shot with the same number as the cluster which it belongs to. The first few clusters in the sequence whose total shot number is bigger than 90 percent of the total video shots number are called as the dominant scene clusters. The other clusters are called noise clusters and the shots in it are called as noise shots. Combined with the shot sequence in the timeline, the sport video with transition model is created with a typical example illustrated in Fig. 4.

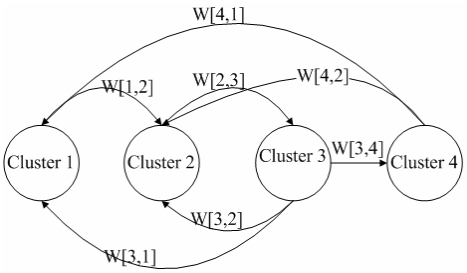

Fig. 4. An illustration of scene cluster transition model

In Fig. 4, cluster1, cluster 2 and cluster 3 are the three dominant scene clusters while cluster 4 is a noise cluster. Each edge $W[i, j]$ is associated with a weight, which denotes the transition times from cluster *i* to cluster *j*. Each transition from cluster *i* to cluster *j* denotes that one shot S_h in cluster *i* and the next shot S_{h+1} in the timeline belongs to cluster *j*. $W[i, j]$ satisfies Equation (2), where *N* is the total shot number in the sports video and *C* is the total cluster number.

$$\sum_{i=0}^{c}\sum_{j=0}^{c} W[i, j] = N - 1 \qquad (2)$$

A cluster transition matrix shown in Table 1 is generated from the example in Fig. 4. Most of the sports videos can be represented by a transition matrix after scene clustering.

Table 1. An example Transition matrix of the transition model in Fig.4 **(**the total shot number is

	Cluster1	Cluster2	Cluster3	Cluster4
Cluster1	W[1,1]=0	W[1,2]=9	W[1,3]=0	W[1,4]=0
Cluster2	W[2,1]=0	W[2,2]=0	W[2,3]=20	W[2,4]=0
Cluster3	W[3,1]=7	W[3,2]=9	W[3,3]=0	W[3,4]=4
Cluster4	W[4,1]=2	W[4,2]=2	W[4,3]=0	W[4,4]=0

4.2 Story unit extraction

An important step of our algorithm is to detect the key scene shots in the timeline. As described above, these shots come from one dominant scene cluster. First, E_i^{in} and E_i^{out} are defined as

$$E_i^{in} = \sum_{j=0}^{c} W[j, i] \quad and \quad E_i^{out} = \sum_{j=0}^{c} W[i, j] \qquad (3)$$

where E_i^{in} represents the total number of edge weight that enters one scene cluster and E_i^{out} represents the total number of edge weight that out of the cluster. If all the story units appear in the same pattern structure without noise shots in video, then for each cluster, $E_i^{in} = E_i^{out}$.But this rarely happens in actual broadcast sports video, which usually have a few different story units and some noise shots.

Before further processing, transition matrix is refined and only the biggest number in each row is kept except the ones in the diagonal. To remove the effects of noise shots, values in the matrix smaller than a predefined threshold are deleted. Then we calculate the ratio of E_i^{in} / E_i^{out} for each cluster. It has three cases:

1. $E_i^{in} = E_i^{out}$. These scene clusters, which are composed of by the shots between the first story unit shot and the last shot, should have an equal number of enter and out edge weight.
2. $E_i^{in} < E_i^{out}$. These scene clusters, which are composed of by the first story unit shots, should have less enter edge weight than the out edge weight.
3. $E_i^{in} > E_i^{out}$. The out edge weight of the scene cluster, which is composed of by the key scene shots, must be smaller than the enter edge weight. And the key scene shots cluster must have the biggest E_i^{in} / E_i^{out} ratio among all the clusters.

The detail of our algorithm is described in Fig. 5.

1. Segment the sports video into shots, $S = \{s_1, s_2, \ldots, s_n\}$. Extract five key-frames in each shot to represent the shot and calculate the normalized 256-bin HSV histograms for each frame, H_m^i(H-16, S-4, V-4), m=1,,5} and i=*1*,,*n*.
2. Perform scene clustering to categorize the segmented shots into *C* scene clusters, C_i, i={1,2,..,c}.Re-sort the resulting scenes, then label each cluster and each shot.
3. Calculate the transition matrix according to the scene clustering result and refine it.
4. Calculate the E_i^{in} / E_i^{out} for each scene cluster. The scene shots cluster with the biggest ratio is the key scene shots cluster that we are looking for.
5. Search the key scene shots S_i^k along the timeline which denotes shot *i* belongs to key scene shot cluster *k*. Then look up the shot S_{i-1} before this shot. Search the transition matrix, if W[r, k]>0 and S_{i-1} is not the key scene shot, mark this shot. Continue this search procedure until we can not find any new shot, then these marked shots consist of one story unit along the timeline.
6. Go to find the next key scene shot and repeat the process until the video ends.

Fig. 5. Story unit extraction algorithm

In this way, story units of sports video are extracted, which could be used in highlight summarization, video indexing and so on.

5. EXPERIMENT RESULT

We apply the proposed method to extract story unit in three kinds of broadcast sports videos including archery (22min), table tennis (33min) and diving (31min). All the test videos

are compressed in MPEG-1 with 25 fps and frame resolution of 352×288. All the videos are recorded from the living broadcast programs of the Olympic Games 2004. Some story units extracted by our approach are presented in Fig. 6.

Fig. 6.Some story units extracted from archery and diving (the story number and the shot number are marked under the picture)

To measure the story unit extraction performance, the ground truth is manually labeled. We use the following criterions to evaluate our approach. Experimental results are shown in Table 2.

$$\text{Precision}=\frac{\#\text{of correctly extracted story units}}{\#\text{of extracted story units}} \quad (4)$$

$$\text{Recall}=\frac{\#\text{of correctly extracted story units}}{\#\text{of actual story units}} \quad (5)$$

Table 2. Story unit extraction performance evaluation

video	Archery	Table tennis	diving	average
P	100%	100%	83.9%	97.1%
R	96.5%	100%	100%	98.8%

From Table 2, it can be found that the results are encouraging. Our approach relies on the scene clustering results. When the clustering results are good, most story units can be extracted correctly. On the other hand, when they are not so satisfactory, the story unit extraction result will be influenced. Moreover, our approach is based on the repetitive patterns of sports video when the pattern is not so clear our result will also be affected. Evaluating the performance of the story unit extraction method is such a subjective work that there are not unified objective criterions to evaluate the story unit extraction results. The defined criterions in this paper just provide a primary result to show the effectiveness of our method.

In the diving video, every actual dive story unit is followed by lots of replay marked by logos. In our actual experiment, in order to removal the effects of the replay, we preprocess the diving video by using the method in [10].

6. CONCLUSION

In this paper, we have proposed an effective model-based approach to extract story unit from sports video streams. In this approach, an unsupervised scene clustering method is first used to cluster the segmented shots into several scene shots clusters. Then each cluster and its shots are labeled. The scene transition matrix is generated from the clustering result. We refine the matrix and find out the key scene cluster. Finally, the key scene shots are detected along the timeline and the story units are extracted around them. In future work, audio feature may also be applied to help extracting story units in sports video.

7. ACKNOWLEDGMENT

This work is partly supported by NEC Research China on "Context-based semantic analysis and retrieval program", Science 100 Plan of Chinese Academy of Science and Beijing Natural Science Foundation: 4063041.

8. REFERENCES

[1] E. Kijak, G. Gravier, P. Gros, L. Oisel and F. Bimbot, "HMM based structure of tennis videos using visual and audio cues," in *Proc. int. conf. on Multimedia and Expo,* 2003.

[2] M. Naphade, T. Huang, "Discovering recurrent events in video using unsupervised methods, in *Proc. int. conf. ICIP,* 2002.

[3] E. Kijak, Oisel , P. Gros, "Hierarchical structure analysis of sport video using HMMS,"in *Proc. int. conf. ICIP*, 2003.

[4] M.Yeung, B.Yeo and B. Liu, "Extracting story units from long programs for video browsing and navigation,"in *Proc. int. conf. Multimedia Computing and Systems*, 1996.

[5] N. Nitta, N. Babaguchi, and T. Kitahashi, "Story based representation for broadcasted sports video and automatic story segmentation,"in *Proc. int. conf. ICME*.2002.

[6] A. Divakaran, K. Perker, S. Chang, R. Radhakrishnan, and L. Xie, "video mining:pattern discovery versus pattern recognition," in *Proc. int. conf. ICIP*, 2004.

[7] P. Wang, Z. Liu, S. Yang, "A probabilistic template-based approach to discovering repetitive patterns in broadcast videos,"*ACM MM*, 2005.

[8] H. Zhang, A. Kankanhalli, S. W. Smoliar, "Automatic partitioning of full-motion video," *ACM/Springer Multemedia Systems*, 1993.

[9] W. Zhang, Q. Ye, L. Xing. Q. Huang and W. Gao, "Usupervised sports video scene clustering and its application to story units detection," in Proc. SPIE – VCIP, 2005.

[10] X. Tong, H. Lu, Q. Liu, H. Jin, "Replay detection in broadcasting sports video," in *Proc. int. conf Image and Grphics,* 2004.

[11] Webster. Webster Dictionary. Available: http://www.m-w.com