A SYSTEM FOR AUTOMATIC GENERATION OF MUSIC SPORTS-VIDEO

Weigang Zhang[1], Liyuan Xing[2], Qingming Huang[2], Wen Gao[1,2]

[1]School of Computer Science and Technology, Harbin Institute of Technology, Harbin 150001, China
[2]Graduate School of Chinese Academy of Sciences, Beijing 100080, China
{wgzhang, lyxing, qmhuang, wgao}@jdl.ac.cn

ABSTRACT

In this paper, we present a new representation of sports video abstract — *Music Sports-Video* (MSV), which provides exciting sports content accompanied with high quality background music for audiences and is available for high-quality audio-visual entertainment. We also propose a system generating MSV from user-provided sports video and music automatically. Firstly, the given sports video is segmented into a series of story units. Then all the story units are ordered by the predefined *Exciting Degree* (ED) and some high ED or user preferred story units are selected for MSV generation. Secondly, the ED of the given music is estimated by energy analysis on music beat. Thirdly, the selected story units are matched with music by their ED corresponding. Finally, the output MSV is rendered by connecting the selected exciting story units with appropriate transition effects, accompanied with the music. Experiments show encouraging results.

1. INTRODUCTION

Sports video is popular and appeals to large audiences. Its abstract, consisting of highlights or story units, is also welcome. Existing works on abstracting of sports video can only extract the highlights or story units from the original video footage, and the audio of the produced abstracts is usually discontinuous and sometimes even accompanied with background noises. Such poor audio drastically decreases their entertainment value. In fact, studies have shown that poor audio will significantly degrade the image quality of the perceived video while high quality audio can improve the perception of video content much [1]. Synchronizing the video and audio can enhance the perception of both. Thus, we present the new representation of sports video abstract—Music Sports-Video (MSV), in which the original poor audio is replaced by the user preferred high quality music. Furthermore, the video and music are synchronized by making the ED (exciting degree) of the video fit to the ED of the music, which makes the generated MSV more professional looking and entertaining. The work in this paper is an attempt to this new interesting area and we propose a system to generate MSV from user provided sports video and music automatically.

To achieve reasonable and pleasing MSVs, the following principles should be taken into consideration.

(1) The shots within each highlight clip or story unit of sports video are casual and changing their order may confuse the viewers. So the original shot order should not be altered. In other words, each highlight clip or story unit in the MSV should keep its integrality.

(2) Music should be suitably matched with the highlight clips or story units, that is, video and music should be synchronized and the most exciting part of music should be matched with the most exciting highlight clip or story unit. And the transitions between highlights or story units should occur at the onset of music beat.

In this paper we focus on the MSV generation from story units of sports video. Fig. 1 illustrates the flow chart of the proposed MSV generation system.

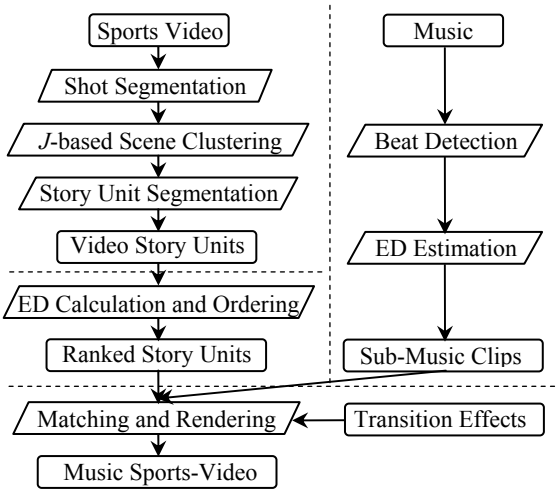

Figure 1: Flow chart of MSV generation system.

The rest of this paper is organized as follows. In section 2, we introduce story unit segmentation and ED ordering for sports video. Then music ED estimation is described in section 3. In section 4, we present the synthesis of video and music. Experiments are presented in section 5 and conclusions are drawn in section 6.

2. SPORTS VIDEO CONTENT ANALYSIS

Story units are extracted from the given sports video firstly. The following sections are a simple introduce of

0-7803-9332-5/05/$20.00 ©2005 IEEE

the story unit segmentation method presented in our previous work [2], followed by the ED calculation and ordering.

2.1. Shot segmentation

The given sports video is firstly segmented into N shots by the method of [3]. Five key-frames are then extracted from each shot at an equal temporal interval to represent the visual content of the shot.

2.2. *J*-based scene clustering

A sports game usually occurs in a specific field. Fixed visual scene content taken by several fixed cameras makes the sports video have well-defined temporal structures. In other words, there are several scenes appearing over and over in the sports video footage in turn. We observed that the alternation of shots from different scenes along the video timeline forms sports story units one by one. For example, in diving, a diving story unit is made up of four scenes —departure position scene, take off and dive scene, entry scene and score scene. Thus clustering the shots into different scenes will be helpful to story unit segmentation.

An unsupervised scene clustering method—*J-based scene clustering* is used in our system. The scene likeness of the shots is evaluated by their HSV histogram distance. Firstly, each shot is initialized as a scene. And then the two scenes whose distance is the smallest are merged into one scene cluster. This procedure is repeated until the merging stop criterion is satisfied. Then the merging process is stopped and the final scene clustering results are obtained. The stop criterion is defined based on a *J value* which is defined according to the Fisher Discriminant Function. The J value is the total scene cluster scatter, which describes the ratio of intra-cluster scatter to inter-cluster scatter of the scenes in the merging process. When the scene number is K_l in the merging process, the J value is defined as

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

Where J_t is the total inter-cluster scatter of the initial scene sequence, J_w^c is the intra-cluster scatter of scene cluster c. N_c is the shot number of scene cluster c. $\|\bullet\|$ represents the Euclidean distance. $\vec{s}_i^{\,c}$ and $\vec{s}_{mean}^{\,c}$ denote the feature value of shot i and the mean feature value of all shots in scene cluster c respectively. $\vec{s}_i$ denotes the feature value of the initial scene cluster i and $\vec{s}_{mean}$ denotes the mean feature value of all initial scenes in which each has only one shot.

At the beginning of the procedure, the intra-cluster scatter of all initial scenes is 0 and J_l is 0.0. With the increasing of intra-cluster scatter when two scenes are merged into one, J_l is rising. If all the scenes are merged into one scene cluster, J_l will reach the maximum 1.0. The smaller J_l is, the more similar the shots within each scene cluster are. Actually, it is expected that both J_l and the scene number are small. But in real situation, J_l is rising with the decreasing of scene number. As a tradeoff between J_l and the scene number, we choose the point where $J_l + k_l$ is the smallest as the best merging stop point, which is shown in Fig. 2. This is the defined stop criterion for the scene merging procedure. Here $k_l = K_l / N$.

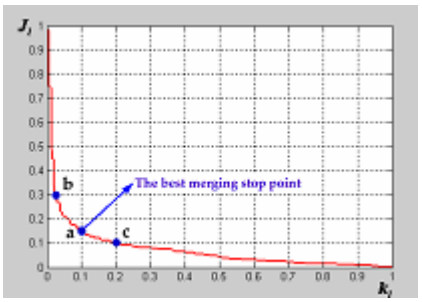

Figure 2: The relation curve of J_l and k_l on an archery video. (At point b, excessive merging takes place; at point c, inadequate merging takes place.)

2.3. Story unit segmentation

In this paper, a story unit is defined as a complete competition interval which starts with the preparation scene of a player and ends with the score scene. For instance, in diving video, a story unit begins with the player standing on the dive platform (or springboard), then follows with the actions of taking-off, diving and entering water, and ends with the score of the player announced.

When scene clustering is completed, the results are used to segment sports video into story units with additional domain knowledge. Firstly, scenes are weighed according to the number of shots they contain. Secondly, the n highest weighed scenes are selected as dominant scenes and labeled with *0,1…n-1*. When labeling a scene, its shots are also labeled as the same. Here scene *0* contains the most shots and n varies with the type of sports (For instance, for an archery video, n is 3). The unlabeled scenes and their shots are considered as noises and discarded. In this way, a labeled shot sequence along the original video timeline is obtained. For example, *{10101020202010101020202010…}* can be the sequence of an archery video. Thirdly, according to the periodicity of the shot label sequence, the video is segmented into story units.

2.4. ED calculation by audio analysis

After the story units are extracted, they are ranked in a descending order according to their ED for the MSV generation. Generally speaking, when players complete

their actions, the audiences will applaud. And the loudness and duration of these applauses show the exciting degree of the actions (or story units). The higher the loudness is and the longer the duration is, the more exiting the story unit is. Thus, the ED of story unit *i* can be calculated as

$$ED_i = \varepsilon_i \times \sqrt{\tau_i} \quad (2)$$

where ε_i stands for the mean spectrum energy which describes the loudness of the applause and τ_i is the duration of the applause.

Several ED ranking experiments are done on a few sports videos. The experimental results show that the ED is consistent with the score from the referees. A high ED corresponds to a high score. This indicates that the defined ED calculation method is reasonable.

3. MUSIC ANALYSIS

In our MSV generation system, to synchronize the video and music well, the structure and ED of the given music are analyzed after sports video content analysis.

3.1. Beat detection

Beat is the basic temporal structure unit of music. It is better to divide the music into sub-music clips at the onset of a beat. In our system, beat is the fundamental unit for music ED estimation.

So far, a lot of literature addresses the problem of beat detection. Our system adopts the method described in [4]. This method does not need any prior knowledge such as tempo, meter, musical style, etc. Its calculation of tempo and beat times is more robust than other methods. In the beat detection procedure, rhythmic events are firstly obtained by grouping the note onsets. Then, a set of tempo hypotheses is generated by examining the relationships between successive rhythmic events. Finally, beat times are calculated by employing a multiple hypotheses search, with an evaluation function selecting the hypothesis that fits the data best.

3.2. ED estimation for music

Generally, the exciting beat is with absolute loudness, sudden sound and fast tempo. The absolute loudness and sudden sound can be measured by average energy and energy peak separately [5]. And the fast tempo can be measured by tempo peak. In this paper, we utilize these three measures to estimate the ED of a beat. The ED of beat *k* is defined as

$$ED_k = \overline{EA_k} \times \overline{EP_k} \times \overline{TP_k} \quad (3)$$

where $\overline{EA_k}$, $\overline{EP_k}$ and $\overline{TP_k}$ are the normalized average energy, normalized energy peak and normalized tempo peak of beat *k*, respectively.

The music ED curve is drawn after beat ED estimation. It is helpful for the system to segment music into different exciting sub-music clips according to the variation of ED curve. Fig. 3 shows the ED curve of the piano music "Pour Elise", in which, high ED values correspond to high exciting beats.

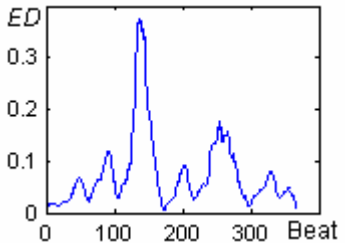

Figure 3: ED curve of piano music "Pour Elise" (after a 15-point moving average smoothing).

4. MSV GENERATION

4.1. Video-music matching

In order to synchronize the video and music, the ED of story units should be consistent with the music, that is, exciting story units should be matched with corresponding exciting sub-music clips. Fig. 4 shows how to match the video and music. Story units and sub-music clips are matched according to their ED. The highest ED story unit is matched with the sub-music clip with the highest ED peak. The second highest ED story unit is matched with the sub-music clip with the second highest ED peak. On the analogy of this, the remained matching is completed in the same way.

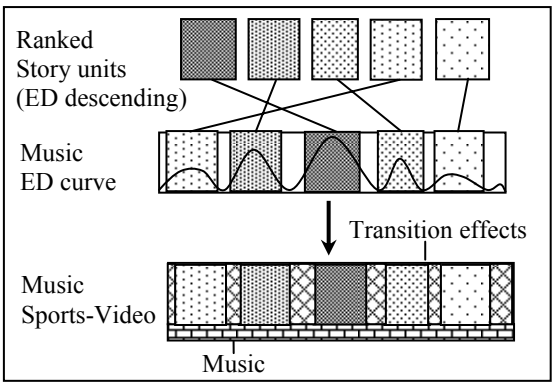

Figure 4: Synthesis of story units and music.

4.2. Final rendering

The output MSV is rendered after video-music matching. Transition effects are added to connect the story units. Their durations are adjusted according to the system requirement. Here what needs to pay more attention to is that the beginning and the end of each transition should occur at the beat onset to ensure the synchronization of

video and music. To make the MSV more professional looking, the following transition effect adding rules should be taken into consideration.

(1) "slow" fade or dissolve are the most commonly used transition effects.

(2) At the beginning and end of an MSV, slow "fade-in" and "fade-out" should be used respectively.

(3) In addition, captions such as sports type, production time and end information can be superimposed at the beginning or the end of MSV.

5. EXPERIMENTS

We have applied the MSV generation system to three kinds of sports videos including diving video, archery video and gymnastics video with different kinds of music including classical music, pop songs, etc. Fig. 5 illustrates the prototype system.

Figure 5: Prototype of automatic MSV generation system

As it is difficult to objectively evaluate the MSV generation system, subjective evaluation method [6] is used. The evaluation results are compared with that produced by Muvee AutoProducer [7]. Seven users are invited to give a satisfaction score on a scale of 1-10, with scale 1 corresponding to the worst score and scale 10 to the best. They do not know which video is generated by which system. Four raw sports videos are used, including two diving videos, an archery video and a gymnastics video. Thus there are 8 results to be evaluated, four of Muvee and four of our system. In Table 1, the average evaluation scores are presented, which indicate that our MSV results are generally better than that of Muvee.

The results of Muvee are focused more on the perception of music and the shots are disordered to match the music. In this way, the viewers will be confused and this degrades the perception quality of the produced musical video. While in our MSV generation system, each story unit keeps its integrality, and high quality music takes the place of the original poor audio, which enhances the perception of both video and music and makes the generated MSVs have higher entertainment value. Some MSVs generated by the proposed system are available at *http://www.jdl.ac.cn/en/project/spises/MSV.htm*.

Table 1: Subjective evaluation and comparison.

#	Video		Music		Avg Score	
	Content	Duration	Genre	Duration	Muvee	MSV
1	Diving-M	12'	Classical	4'16"	7.7	9.0
2	Diving-WS	12'13"	Classical	5'56"	7.6	7.7
3	Archery	22'14"	Pop Song	3'58"	7.8	7.8
4	Gymnastics	10'1"	Pop Song	4'34"	6.9	8.6
Avg	-	-	-	-	**7.5**	**8.3**

6. CONCLUSIONS

In this paper, an automatic Music Sports-Video generation system is proposed. The system abstracts the given sports video by using story units, uses high quality music to replace the original poor audio and provides better entertainment for audiences. In this system, a set of content-based sports video and music analysis algorithms are used to automatically generate professional looking MSVs. Furthermore, the system can also work in an interactive manner. User can not only select his favorite music, but also select his preferred story units, order them through user control, add his favorite transition effects (The music, story units and each transition effect can be previewed in advance), and obtain MSV in his own style. This system is a part of our project SPISES [8].

ACKNOWLEDGEMENTS

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

7. REFERENCES

[1]W.R.Neuman," Beyond HDTV: Exploring Subjective Responses to Very High Definition Television," MIT Media Laboratory Report, July. 1990.

[2]Weigang Zhang, Qixiang Ye, Liyuan Xing, Qingming Huang and Wen Gao, "Unsupervised Sports Video Scene Clustering and Its Applications to Story Units Detection," to appear in Proc. VCIP'2005, Beijing, China, July 12-15,2005.

[3]Zhang H J, Kankanhalli A, Smoliar S W, "Automatic Partitioning of Full-Motion Video," ACM/Springer Multimedia Systems, 1993, 1(1) :10～28.

[4]S. Dixon, "Automatic Extraction of Tempo and Beat from Expressive Performances," Journal of New Music Research, 30 (1), 2001, pp 39-58.

[5]Yu-fei Ma, Lie Lu, Hong-Jiang Zhang and Ming-jing Li. "An Attention Model for Video Summarization," ACM Multimedia 2002, pp. 533- 542, France, Dec 1-6, 2002.

[6]Xian-Sheng Hua, Lie Lu, Hong-Jiang Zhang, "Automatic Music Video Generation Based on Temporal Pattern Analysis," ACM Multimedia 2004, October 10-16, NY USA. 2004.

[7]Muvee AutoProducer, http://www.muvee.com.

[8]SPorts vIdeo Summarization and Enrichment System http://www.jdl.ac.cn/en/project/spises/SPISES.htm .