Story Unit Segmentation with Friendly Acoustic Perception*

Longchuan Yan[1,3], Jun Du[2], Qingming Huang[3], and Shuqiang Jiang[1]

[1] Institute of Computing Technology, Chinese Academy of Sciences, Beijing, 100080, China
[2] NEC Laboratories China, Beijing, 100084, China
[3] Graduate School of Chinese Academy of Sciences, Beijing, 100049, China
{lchyan, jdu, qmhuang, sqjiang}@jdl.ac.cn

Abstract. Automatic story unit segmentation is an essential technique for content based video retrieval and summarization. A good video story unit has complete content and natural boundary in visual and acoustic perception, respectively. In this paper, a method of acoustic perception friendly story unit segmentation for broadcast soccer video is proposed. The approach combines replay detection, view pattern and non-speech detection to segment story units. Firstly, a replay detection method is implemented to find the highlight events in soccer video. Secondly, based on positions of replay clips, an FSM (Fine State Machine) is used to obtain rough starting points of story units. Finally, audio boundary alignment is employed to locate natural audio boundaries for acoustic perception. The algorithm is tested on several broadcast soccer videos. The story units segmented by algorithms with and without audio alignment are compared in acoustic perception. The experimental results indicate the performance of the proposed algorithm is encouraging and effective.

Keyword: Video Processing, Story Unit Segmentation, Acoustic Perception, SVM.

1 Introduction

Story unit segmentation is to extract the events related video clips from a long video, which is a necessary component for video summarization, retrieval and indexing. The general story unit segmentation is still an open problem due to the limited generalization ability of pattern recognition algorithm and various video patterns. The pre-knowledge combined story unit segmentation algorithm can provide some good solutions in specific fields. As a major branch, the story unit segmentation in sports videos attracts much attention for its wide applications and tremendous commercial potentials. This paper utilizes the soccer video as representative example to illustrate the story unit segmentation issue.

* This work is supported by National High-Tech R&D Program (863 Program) under the grant No. 2006AA01Z117 and NEC Labs, China.

N. Sebe, Y. Liu, and Y. Zhuang (Eds.): MCAM 2007, LNCS 4577, pp. 223 – 231, 2007.

In the last decade, many approaches have been proposed for story unit segmentation of various sports video. B. Li et al. employed audio energy, scene cut and grass ratio to locate starting points of live video to extract the story unit for soccer video [1]. In [2-3], the view type and its transition relationship were used to detect the goal and "attack" events for soccer video. N. Noboru et al. used the CC (Closed Caption) text to aid the story segmentation for football video [4]. J. Wang et al. used SVM classifier to detect the event for soccer video taken by main camera and used play position and time duration to determine the event boundaries [5]. C. Liu et al. employed unsupervised scene cluster and their transition matrix to extract the story units in table tennis, diving and archery [6]. X. Tong et al. tried to segment highlight play-units for sports video browsing by some rules [7]. N. Babaguchi et al. used the dominant color and the ratio of the number of vertical lines to that of horizontal lines to link the replay and the live scene [8]. The above algorithms combine domain knowledge and some image or audio features to segment story units for specific video types. They employ the visual content as the story unit segmentation criterion, such as grass ratio, scene cut, view pattern, and etc. From the visual content, the story unit has complete content and natural boundaries. However, the algorithms pay little attention to the characteristics of acoustic perception of story unit. The story unit may have incomplete content or unnatural boundaries in audio stream, which is unfriendly in acoustic perception. The acoustic perception friendly story unit has complete content and natural boundary in audio stream, which give audience better acoustic perception.

In this paper, a method of acoustic perception friendly story unit segmentation is proposed. The algorithm combines production pattern, view pattern, and non-speech detection to locate the highlight events and segment the story units. To obtain good human perception, the algorithm tries to segment the story units with complete content and natural boundaries in image and audio streams, respectively. The story unit used in [1] is employed to represent highlight event, which is composed of the replay and its corresponding live video. The diagram of our algorithm is shown in Fig. 1. Firstly, the algorithm uses the replay detection to locate the highlight events. The replay is a well-known production pattern to enhance highlight events in soccer video, thus the replay detection can locate almost all the highlight events. Secondly, the algorithm determines the starting point of a live video by using view pattern. The

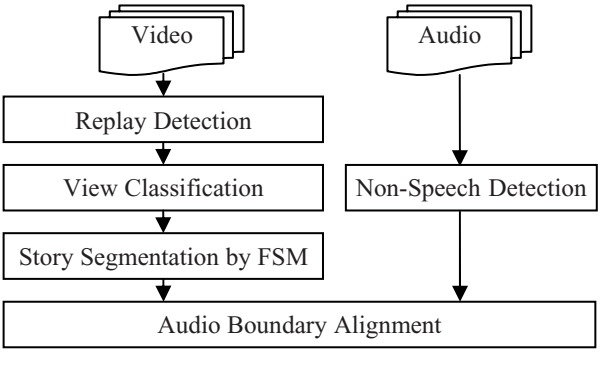

Fig. 1. Block diagram of story unit segmentation algorithm

grass ratio has been used to describe view type and story unit segmentation, but it is still a rough method. To obtain a fine view pattern description, the algorithm employs an SVM classifier to recognize the view type and uses an FSM to describe the view pattern. Lastly, the algorithm adjusts the audio boundaries of story units by selecting the non-speech points as final boundaries. The audio boundary alignments make the story unit more natural by keeping the integrity of speech. Compared with the grass ratio based method, the proposed algorithm is more robust by introducing the fine view classification and view pattern model. Furthermore, the algorithm can extract story units with the complete content and natural boundaries in image and audio streams, which can provide better visual-audio perception to audience.

The rest of the paper is organized as follows. Section 2 introduces story unit and its characteristics of acoustic perception. In Section 3, the view classification and non-speech detection are described, respectively. In Section 4, the proposed story segmentation algorithm is discussed in detail. Experimental results are presented in Section 5. Finally, Section 6 concludes the paper.

2 Story Unit and Its Characteristics of Acoustic Perception

The story unit used in this paper is defined as a replay and its corresponding live video. For visual content, the story unit makes audience view event clearly because replay often show event in different angles or views and the live video can offer audience to see the whole process of event.

For acoustic perception, announcer's speech in audio stream contains much useful information about a match. The story unit with the complete content and natural boundaries is acoustic perception friendly. The integral content means the audio clip contains enough information about the content of event in story unit. For the story unit, the replay part gives enough time for the announcer to illustrate the event though his/her speech always lags behind the content of image sequence. The natural boundary means the story boundary is on non-speech point. If the segmentation algorithm does not consider the acoustic perception, some story boundaries may be on phonations, such as points A and C in Fig. 2. They are unnatural story boundaries for acoustic perception. The proposed algorithm tries to select non-speech points as story boundaries for good acoustic perception.

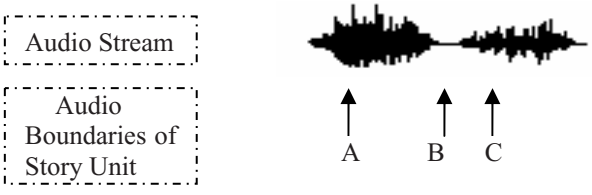

Fig. 2. Audio boundaries of story unit

3 Video and Audio Preprocessing

3.1 Replay Detection

In broadcast soccer video, the replay is often emphasized by adding two logos before and after the replay to attract the attention of audience. Detecting the replay through logo detection is an effective method to locate the highlight events. In this paper, our previous logo based replay detection method in [9] is implemented to detect the replay.

3.2 View Classification

Generally, there are four kinds of typical views in soccer video: namely global view, middle view, close-up view and outside view, as shown in Fig. 3. Here an SVM classifier is used to recognize the view type and local grass ratios are used as view classification features.

(a) (b) (c) (d)

Fig. 3. Typical views in soccer video (global view (a), middle view (b), close-up view(c) and outside view (d))

3.2.1 Adaptive Playfield Color Detection

The playfield color varies in different lighting or playfield conditions. To calculate the grass ratio correctly, a dominant color detection method is employed to adaptively determine the playfield color of each video. Firstly, the algorithm selects 100 frames randomly from the video and picks up the peak value of the Hue histogram of each frame. Then, the algorithm removes the peak values if they are not in the range to represent the Hue of playfield color. Finally, the average value and its range are calculated over the left peak Hue values to modal the playfield color.

3.2.2 Feature Selection

Grass ratio of image has been used as feature of view classification [2, 10], but the grass ratio does not contain spatial distribution and size information of objects. To

Fig. 4. Zones arrangement of local grass ratios

distinguish the views effectively, nine local grass ratios are introduced to describe the spatial distribution and size of objects. They are calculated by the number of grass pixels divided by that of all the pixels in a zone. The local zones arrangement is shown in Fig. 4.

3.2.3 View Classification

In this paper, the non-playfield images are regarded as the outside views if their global grass ratios are less than a predefined threshold. Other three types of views are recognized by an SVM classifier with RBF kernel function.

3.3 Non-speech Detection

Non-speech detection in an audio stream is a typical two-class recognition problem. We employ an SVM classifier to detect non-speech segments [11]. The feature extraction is a critical step of non-speech detection. The Shot-Time Energy, Zero-Crossing Rate, and 16-order Mel-frequency Cepstral Coefficients (MFCC) are extracted from each audio frame and combined as a feature vector. The length of the audio frame is 256 samples without overlapping. Before feature extraction, the audio signal is converted into a general format, which is 11,025 Hz, 16-bit and mono-channel.

To reduce abrupt misclassification of the SVM classifier, we use a 40-frame average filter to refine the classification results. In addition, we assume the minimal length of the non-speech segments is more than 40 frames (about one second). Otherwise, they are regarded as pauses between words in sentences.

4 Segmentation Method

In this section, we introduce the FSM based starting point detection of live video and audio boundary adjustment for friendly characteristics of acoustic perception.

4.1 Detection of Rough Starting Point of Live Video

In soccer videos, the view type transition has typical pattern in an event. We use an FSM to describe the view pattern to locate the rough starting points of live videos, as shown in Fig. 5. The view type V and view duration T are used to describe the story pattern in the FSM. There are three states in the FSM, namely E, L and B, which correspond to the ending candidate sub-clip, live story sub-clip and beginning candidate sub-clip, respectively.

In some cases, the live story sub-clip is too long relative to the length of replay. In other cases, the short zoom in or zoom out may affect the story unit segmentation. To avoid the two problems, two time constraints in the FSM are introduced. The length of the first shot in replay is used to calculate the time constraint T_c for constraining the search range of live video. It is more accurate than that of replay, because the replay may contain several shots to show an event. In addition, we use another time

constraint T_o to reduce the bad effect of short time zoom in or zoom out. In the proposed algorithm, the two time constraint thresholds in each replay are calculated by

$$T_c = \alpha \times T\_Shot \tag{1}$$

$$T_o = \beta \times T\_Shot \tag{2}$$

where *T_Shot* is the length of the first shot in a replay and α, β are the scale factors. In our experiments, α and β are set to 4 and 0.2, respectively.

For the detection of the starting points of live videos, view type and its duration are first identified, and then they are put into the FSM from the view type near the replay to the far. The FSM will stop when the starting point of live video is found.

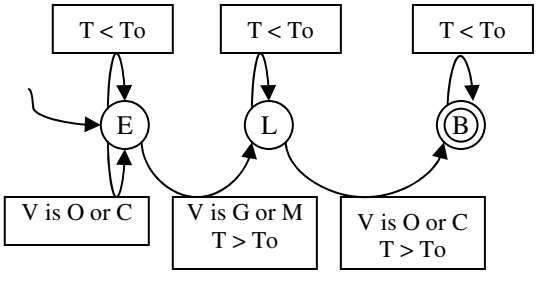

Fig. 5. FSM of view pattern in live video

4.2 Audio Boundary Alignment

The starting points of live videos produced by FSM and the ending points of replays are visual boundaries of story units. To obtain acoustic friendly boundaries of story units, the longest non-speech segment near the visual starting points or ending points is selected as the final story unit boundary. If there are no non-speech points near the boundaries, the visual boundaries are not adjusted any more.

5 Experimentation

For story unit segmentation, we use two broadcast soccer videos of FIFA 2006 to test the algorithm. The videos are compressed in MPEG-1 with 25 fps and frame resolution of 352×288. The middle experimental results on the view classification and the non-speech detection are first given below.

For view classification test, we collect 2,665 global view images, 846 middle view images and 534 close-up view images from different videos. About 1,400 images are used to train the classifier and others to test the classifier. The accuracy of view classification is from 88.6% to 99.0%. From experimental results, the classifier can

distinguish the three view types very well, though classification of the middle view is not as good as those of the other views.

For non-speech detection test, we manually label four audio sub-clips decompressed from soccer videos. The total length of audios is about 2,286 seconds. We select about 93 seconds audio as training set, and others as testing set. The average recall and precision are 79% and 87%, respectively. In the experiments, some transient pauses between words in sentences are the major obstacles for the high accurate non-speech detection, since it is difficult to label them exactly in training data set.

For the story unit segmentation, we use two metrics to evaluate the performance of the proposed algorithm, namely story integrity and natural audio boundary. The story integrity is that a story unit just contains one event. The natural audio boundary is that a story unit without speech abruption at its boundary. Since the two metrics are hard to compute, the subjective evaluation method is adopted. The evaluation result of story segmentation algorithm in story integrity is shown in Table 1. The #Story is the number of the whole stories in a video; the #P is the number of the segmented stories whose length is acceptable. The #S and #L are the numbers of the story units whose length are longer or shorter than their ideal length, respectively. Some key frames of segmented story units are shown in Fig. 6.

Table 1. Story Segmentation Results

Video name	# Story	# P	# S	# L
06_11_03.mpg	30	25	2	3
07_09_03.mpg	35	29	5	1

The numbers of natural audio boundaries story units produced by segmentation algorithms with and without audio boundary alignment are listed in Table 2. It can be seen that the audio boundary alignment can produce more story units with natural audio boundaries.

Table 2. Audio Boundary Evaluation

Video name	# Story	# natural story units with audio alignment	# natural story units without audio alignment
06_11_03.mpg	30	23	16
07_09_03.mpg	35	28	18

From the above two tables and Fig. 6, the algorithm can produce a promising performance. The proposed approach is based on some pattern recognition algorithms. When the view classification and non-speech detection are good, most story units can be extracted correctly. Otherwise, the story unit segmentation and boundary alignment are affected. Additionally, dynamic variation of view pattern in story unit may cause improper segmentations.

Fig. 6. Some Key Frames of Segmented Story Units

6 Conclusion

In this paper, an acoustic perception friendly story unit segmentation algorithm is proposed. The algorithm tries to segment the story units with complete content and natural boundaries in image and audio streams, respectively. For visual content, the view pattern and time constraints are used to segment the story units by an FSM. For audio stream, the audio boundary alignment is employed to avoid the speech abrupt cutting. From experiments, the algorithm can improve the acoustic perception of story units, especially in natural audio boundary. However, there is still much room to improve the story unit model and audio boundary alignment method. In future, we will seek some machine learning methods to model the story unit and take more attention to the audience perception of story unit.

References

1. Li, B., Errico, J.H., Pan, H., Sezan, I.: Bridging the semantic gap in sports video retrieval and summarization. Journal of Visual Communication & Image Representation, 393–424 (2004)
2. Ekin, A., Tekalo, A.M., Mehrotra, R.: Automatic Soccer Video Analysis and Summarization. IEEE transactions on Image Processing, vol. 12(7) (July 2003)
3. Ren, R., Jose, J.M.: Football Video Segmentation Based on Video Production Strategy. In: Proceedings of 27th European Conference on Information Retrieval, March 2005, Sandiago (2005)

4. Noboru, N., Babaguchi, N., Kitahashi, T.: Story Based Presentation for Broadcasted Sports Video and Automatic Story Segmentation. In: Proceedings of 2002 IEEE International Conference on Multimedia and Expo, August 2002, Switzerland (2002)
5. Wang, J., Xu, C., Chng, E., Wan, K., Tian, Q.: Automatic Replay Generation for Soccer Video Broadcasting. In: Proc. of 13th ACM International Conference on Multimedia (2004)
6. Liu, C., Huang, Q., Jiang, S., Zhang, W.: Extracting Story Units in Sports Video Based on Unsupervised Video Scene Clustering. In: Proc. of 2006 IEEE International Conference on Multimedia and Expo, Toronto, Canada, pp. 1613--1616 (2006)
7. Tong, X., u, Q., Zhang, Y., Lu, H.: Highlight Ranking for Sports Video Browsing. In: Proc. of 13th ACM International Conference on Multimedia, Singapore, November 2005 pp 519–522 (2005)
8. Babaguchi, N., Kawai, Y., Yasugi, Y., Kitahashi, T.: Linking Live and Replay Scenes in Broadcasted Sports Video. In: Proc. of ACM Multimedia 2000, Worshop on Multimedia Information Retrieval, Marina del Rey, pp. 205–208 (2000)
9. Zhao, Z., Jiang, S., Huang, Q., Zhu, G.: Highlight Summarization in Sports Video Based on Replay Detection. In: Proc. of 2006 IEEE International Conference on Multimedia and Expo, Toronto, Ontario, Canada, pp. 1613–1616 (2006)
10. Rjondronegoro, D., Chen, C., Pham, B.: Sports Video Summarization using Highlights and Play-Breaks. In: Proc. of 11th ACM International Conference on MIR (2003)
11. Lu, L., Zhang, H., Li, S.: Content-based audio classification and segmentation by using support vector machines. Multimedia Systems, pp. 482–492 (2003)