SportsVBR: A Content-Based TV Sports Video Browsing and Retrieval System

Liu Huayong[1], Zhang Hui[2]

[1] Department of Computer Science, Central China Normal University, Wuhan 430079, Hubei, PR China
hyliuwuhee@hotmail.com
[2] Finance Department of Business School, Wuhan University, Wuhan 430072, Hubei, PR China
zhanghui994@sohu.com

Abstract. An advanced content-based sports video browsing and retrieval system, SportsVBR, is proposed in this work. Its main features include event-based sports video browsing and keyword-based sports video retrieval. The paper first defines the basic structure of our SportsVBR system, and then introduces a novel approach that integrates multimodal analysis, such as visual streams analysis, speech recognition, speech signal processing and text extraction to realize event-based video clips selection. The experimental results for sports video of world cup football games indicate that multimodal analysis is effective for video browsing and retrieval by quickly browsing event-based video clips and inputting keywords according to a predefined sports vocabulary database. The system is proved to be helpful and effective for the overall understanding of the sports video content.

1 Introduction

With the deep development and abroad application of multimedia technology, video digital has become a very important information expressing form in the information system. With the remarkable increase of video data, it is becoming important to index and store them considering their retrieval and recycling. In order to enable detail retrieval, understanding the semantic contents of the video is inevitable. And automatic indexing and retrieval of video information based on content is a very challenging research area with many research efforts addressing various relevant issues [1, 2, 3, 4, 5]. The most difficult problem is: what does content mean? Or, more specifically, how should one characterize visual or auditory content present in a video, and how to extract them for building useful, high-level annotations to facilitate content-based indexing and retrieval of video segments from huge digital video libraries? It is generally accepted that content is too subjective to be characterized completely because it is often concerned about objects, background, domain, context, etc. This is one of the main reasons why the problem of content-based access is still largely unsolved. Ref. [6] has presented a video browsing method, which is conducted through a VCR-like

interface. It is tedious and time-consuming, and is not concerned about high-level semantic content understanding. Ref. [7] has presented a video's structured browsing and querying system called videowser, but it also has not realized effective content-based retrieval and just considers the image analysis, so it is not a really content-based video browsing and retrieval system.

Sports video of TV programs is an important resource for the sports fans or the special sports analysis experts. But what the consumer wants to see are the interesting segments of the sports video, and so how to extract the interesting segments, that is to say, how to index and retrieve the content of events is the problem that we want to solve in this paper. There are lots of related works that are concerned with the retrieval of sports video such as Y. Gong et al. [8] that presents a method to automatic parse the soccer programs using domain knowledge, Yoshinori Ohno et al. [9] that describes a system to track soccer players and a ball by using color information from video images, etc. But most of the recent reported work related to sports video focus on one of video features and do not truly constitute a content-based multimedia research method.

Though much research work has been made towards developing automatic video searching system in recent years, however, because of the numerous video program variations, it is still a very difficult work to design a general-purpose system for all types of video programs. In this paper, we focus on TV sports video as a particularly important category of video programs and design a content-based sports video browsing and retrieval system, SportsVBR, which is convenient for users to fast browsing and retrieving sports video. Combining audio-visual features and caption text information, the system can automatically selects the interesting events. Then using automatically extracted text caption and results of speech recognition as index files, SportsVBR supports keyword-based sports event retrieval. The system also supports event-based sports video clips browsing and generates key-frame-based video abstract for each clip.

The rest of the paper is organized as follows. In Sec. 2, we present an overview of our system. We first present an algorithm in Sec. 3 to select video clips that may contain events. Then we describe how to get keywords such as "goal" or "penalty kick" by speech recognition and detect interesting segments by computing the short time average energy and other parameters of audio. At last, a method of extracting textual transcript within video images is introduced to detect events and use these textual words to generate the indexing keywords. Based on sections above, we present the approach of content-based browsing and retrieval of sports video in Sec. 4, and in Sec. 5, the interface of SportsVBR and its functions are given, and conclude the paper in Sec. 6.

2 Overview of the System

Fig. 1 shows the block diagram of our system. The modules shown within the dotted lines form the core part of our system and also are the main subject of this paper. Our system analyzes the sports video by dividing it into video and audio streams respectively. In video streams, it processes the visual features

and extracts the textual transcripts to detect the shots that probably contain the events. Visual features are not sufficient for detecting events, so the textual transcripts detection can improve the accuracy and it also use to generate the textual indices for users to query the video events clips. In audio streams, we realize the speech recognition of special words such as "goal" or "penalty kick" and use these words to generate the textual indices, too. In audio signal processing module, we compute several parameters of audio signal to find the interesting parts of sports video more accurately. After we find the video events clips, we organize them in our system for content-based browsing and retrieval.

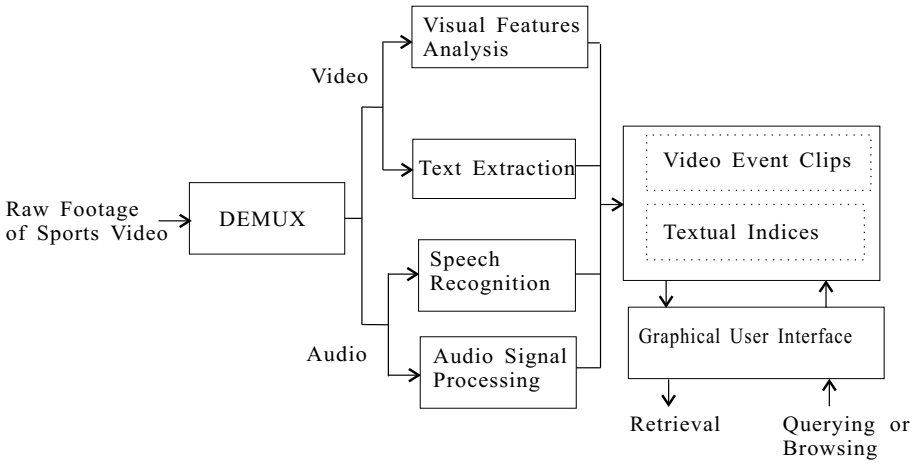

Fig. 1. The block diagram of our system

3 TV Sports Video Multimodal Analysis

Sports video indexing based on events is a kind of indexing by semantical contents, and we think that an event is defined over a time interval, not just a time point. Further, a current event is closely related to preceding events or subsequent ones. In football games, there are lots of events can be considered from different points of view. In our research work, we are particularly concerned with the events that maybe change the score and are interesting for fans and coaches or kinematics researchers: i) penalty kicks (PK), ii) free kicks next to goal box (FK), and iii) corner kicks (CK). These are typical events shown in TV news and magazine programs summarizing a match: they thrill the speakers and the audience, as they embody attack actions in proximity of the goal box area that might eventually lead to the scoring of a goal. Hence, penalty kicks and corners are also often used to calculate statistics supporting the evaluation of the degree of aggressiveness of the two teams.

3.1 Visual Streams Analysis

For video analysis, visual streams processing are aimed at shot-by-shot indexing, so we try to discover a shot that is similar to the target event by matching of feature vectors. First, the sports video stream is segmented into shots. The fundamental element in video processing is usually a shot, which is defined as image sequence what the pickup camera records during one continuous movement. The shots are given by detecting shot change operations such as cuts or dissolves. Second, a shot to be matched is selected from all the segmented shots. For the shot, we extract $N1$ image frames from its head and $N2$ frames from its tail. Because for the events described above, the ball is set to a steady state for kick in the starts of these events, and the ends are the goal box scene or a field scene taken behind a goal post. So the scenes are transitions from a camera-steady frame containing a steady ball to a scene containing goal box or scene taken behind a goal post. Each frame is divided into 4×4 rectangular blocks. Color distribution in each block is given as feature parameters. A feature vector f_n of an image frame n is formed as

$$f_n = (R^n_{1\times1}, G^n_{1\times1}, B^n_{1\times1}, R^n_{1\times2}, G^n_{1\times2}, B^n_{1\times2}, ..., R^n_{4\times4}, G^n_{4\times4}, B^n_{4\times4}) \, . \quad (1)$$

where $R^n_{k\times l}, G^n_{k\times l}, B^n_{k\times l}$ are average RGB values in the kth$\times l$th block. We build a feature vector F of a shot by concatenating each vector of $N(= N1 + N2)$ image frames. In symbols

$$F = (f_1, f_2, ..., f_N) \, . \quad (2)$$

Finally, we measure the distance between the vector of the shot and that of the example image sequence, which may be viewed as a temporal image model of the target event. Let G denote the latter vector. Of course, the dimension of both vectors is identical. In the matching, the distance is given by

$$d(F, G) = \|F - G\|^2 \, . \quad (3)$$

If d is smaller than some threshold, the shot is indexed by the target event: PK, FK, or CK. The example sequence is provided for each target event. It is a crucial problem how we should obtain it. The ideal way may be learning from example stream. At the current stage, we employ a simple way, selecting a sequence randomly from the concerned stream.

3.2 Audio Streams Analysis

The fundamental purpose of this method is to detect segments from TV sports programs where interesting events described above occur. Because only use the visual streams analysis, it is not so sufficient to detect these interesting events exactly. Adding audio streams analysis can improve the accuracy rate of event detect greatly. This typically means increased crowd activity manifested as increased energy level in audio. However, energy level alone is not sufficient in

detecting meaningful segments; further processing for recognition of keywords such as "goal" or "penalty kick" is necessary in order to detect the interesting events.

Our algorithm can run with sequences sampled at rates much lower than the typical rates. For that, we sub-sample the 44.1kHz data and work with 441Hz sequences. This further improves the computational efficiency and makes it possible to implement this method on platforms with less computation power. We process the audio streams in units of segments that are one second long. We use a form of audio energy that we call $InterestingLevel$, $X(A)$ which is computed as average absolute amplitude for each segment: if A is a segment of the audio level, the interesting level, $X(A)$=Ave(Abs(A)) where Abs is the absolute value function. Then, a sliding window of five units (equivalent to five seconds) is used to compute the average levels, as in most situations true events last for at least five seconds. Then segments with averages above a certain threshold are combined to form a sequence. The threshold we used is half of the max value over the entire program (approximately three hours). The measure of importance used in ranking for each segment is the summation of all average energy values through the range. This ensures that energy as well as the duration is taken into account in picking that segment. The parameters used in the algorithm such as the five second interval or the level of threshold is experimental and further optimization may be possible.

Recognition of keywords such as "goal" or "penalty kick" is helpful to detect these interesting events. In our framework we integrate the speech recognition engine that is compiled with the API functions, which are provided by the speech recognition development kit Microsoft Speech SDK 5.0 of Microsoft corps. This engine recognizes the commentator's voices in each shot to get the frame containing the interesting events.

3.3 Text Extraction

For retrieval of sports video, we should extract the meaningful texts appeared in the frames within interesting events and other situations. Because these specific texts including explain an athlete or reflect the score change, etc, and they appear in a fixed subregion of an image frame, contain textual information that expresses the on-going scenes. An example is displayed in Fig. 2. The text segmentation and extraction algorithm is described in detail in another paper of mine [10].

Fig. 2. An example of the meaningful texts

Using test data set, world cup football games selected from our video database randomly, which lasts three hours or so in total, we obtain an accuracy rate of 91.3% and a recall rate of 97.5% for sports video event clip selection.

4 Content-Based Browsing and Retrieval of Sports Video

According to analysis of sports video above, we present content-based sports video browsing and retrieval. For this paper, we combine the method of content-based browsing event-based video clips and the method of querying by inputting keywords.

Browsing video is an important retrieval method to obtain video content. Through the event detection of sports video, we can get the three interesting events: PK, FK and CK. Then the clips containing the three events are stored into the system database for users' browsing. Key frame can let users know general meanings of a video clip quickly, and now there are lots of algorithms that introduce how to extract key frame. Considering the specification of TV sports programs, we select three kinds of frames as key frames in our framework. The first is the frame that is closest to the middle point in temporal space in $N1$ frames of every event shot. The second is the corresponding frame in $N2$ frames. And the third is the frame that is closest to the middle point in temporal space in every event shot. So the key frames and the video clips of interesting events forms a kind of video abstract for users' retrieval.

Querying by users' inputting keywords is another efficient method for video retrieval. Keywords obtained from the speech recognition module and text extraction form full-text search indices, and some keywords are predefined in a small sports video vocabulary database that are built according to our observation results for a month of CCTV sports programs and the database can be expanded in future work. When users input keywords whatever they thought about such as "goal" or "corner kick" to query sports video, and then the system can provide the exact video abstracts.

5 The Interface of SportsVBR and Its Functions

Fig. 3 shows the interface of our system. It consists of four sections, the display window, the video control window, the sports event clips display window and the key word inputting query window. User is able to input sports video files, and set the start and end of the video files. Clicking the play button in the video control window, the sports video event clip selection of the video files is automatically finished. The results of extracted key frames of the selected event clips are shown in the sports event clips display window. The display window is used to play a video file in realizing event clip selection. And also is used to play a event clip selected from the selected results or to show a key frame selected from the extracted key frames. User is able to input keywords to search event clips, for example, the word "goal", and then the retrieval result is shown in the

display window. Click the key frame and the corresponding event clip can be played in the display window.

Fig. 3. The interface of our system

6 Conclusions

Content-based video browsing and retrieval for video flows is a hot spot in the recent researches of video database. This paper develops a system, called SportsVBR, to realize fast and efficient sports video browsing and querying based on event-based video clips selection. The system is designed for parsing TV sports video, but its integration strategy of audio-visual cues, the analysis of text event detection, as well as the methods of content-based video browsing and retrieval can also be applied to the scene segmentation and video retrieval of other video types in future work.

References

1. S.W.Smoliar and H.J.Zhang: Content-Based Video Indexing and Retrieval. Multimedia (1994) **1**(2): 356–365
2. Michael G. Christel: Visual Digests for News Video Libraries. In Proc. of the ACM Multimedia'99 Conference (1999) 303–311
3. Alexander G.Hauptmann, Michael J. Witbrock: Story Segmentation and Detection of Commercials In Broadcast News Video. In ADL-98 Advances in Digital Libraries Conference (1998) 168–179
4. Cuneyet Taskiran, Jau-Yuen Chen, Charles A.Bouman et al: Compressed Video Database Structured for Active Browsing and Search. In Proc. of IEEE Int'l Conf. on Image Processing (ICIP'98) (1998) 133–137
5. H.J.Zhang, Y.H.Gong, S.W.Smoliar et al: Automatic Parsing of News Video. In Proc. of IEEE Int'l Conf. on Multimedia Computing and Systems (1994) 45–54

6. Y.H.Chang, D.Coggins, D.Pitt et al: An Open-Systems Approach to Video on Demand. IEEE Communications Magazine (1994) 68–80
7. Wu Lingqi, Li Guohui: Video's Structured Browsing and Querying System: Videowser. Mini-Micro System (2001) 112–115
8. Y.Gong, L.T.Sin, C.H.Chuan et al: Automatic Parsing of TV Soccer Programs. ICMCS'95 (1995) 167–174
9. Yoshinori Ohno, Jun Miura and Yoshiaki Shirai: Tracking Players and a Ball in Soccer Games. In Proc. of the 1999 IEEE Int'l Conf. on Multisensor Fusion and Integration for Intelligent Systems (1999) 147–152
10. Liu Huayong, Zhou Dongru. Content-based News Video Story Segmentation and Video Retrieval. In Proc. of SPIE Second International Conference on Image and Graphics (2002) 1038–1044