A General Framework for Sports Video Summarization with its Application to Soccer

Baoxin Li, Hao Pan, Ibrahim Sezan

Sharp Laboratories of America
Camas, WA 98607, USA

ABSTRACT

We propose a general framework for indexing and summarizing sports broadcast programs and its specific application to soccer. The framework is based on a high-level model of sports broadcast video using the concept of an event, defined according to domain-specific knowledge for different types of sports. Thus it covers most sports including those that have the action-and-stop pattern (e.g., baseball) and those containing continuous actions (e.g., soccer). In particular, within this general framework, using soccer as an example, we propose a novel approach to automatic event detection, which is based on automatic analysis of the visual and aural signals in the media. An MPEG-7 compliant prototype browsing system has been implemented to demonstrate the results.

1. INTRODUCTION

With the increasing amount of audio-visual data that are broadcast or available in a prerecorded format, there is an emerging need for efficient media management including browsing, filtering, indexing, and retrieval. Among the challenges facing current media management systems is the automatic extraction of the media content descriptions. This paper propose a general framework for automatic event detection and video summarization in the sports video domain, and its specific application to soccer.

We start with a general model of a sports broadcast video at a high level, with event/non-event breakdown of the video. According to this model, a sports broadcast video is composed of sparse event segments that are interleaved with other not-so-interesting segments, which are referred to as non-event segments. The definition of an event can be specialized to a specific sport based on specific domain knowledge (e.g., an event can be a pitch in baseball or a goal in soccer). This is a unifying model due to the fact that it applies to different types of sports, including those that have the action-stop pattern, e.g., baseball and American football, and those that are continuous without obvious breaks, e.g., continuous-action sports such as soccer and ice hockey. Automatic event detection algorithms can be developed on the basis of this general model. In particular, we propose a novel algorithm for soccer event detection using the proposed continuous-action model.

Once event segments are detected, indexing is performed on the basis of starting and end points of these segments. Video summarization is implemented by concatenating event segments thereby forming a condensed version of the video. In comparison with existing methods and especially those that aim at extracting only the "exciting" segments of sports programs automatically (e.g., [1,2]), the proposed framework extracts events that are either objectively defined using the domain knowledge of the underlying sports, or defined based-on the selection that a human operator makes during the production of the broadcast program, e.g., replays provided during the broadcast. Hence the results obtained by the proposed approach are more robust in practice.

In Section 2, we provide an event-based modeling of sports broadcast video supporting both stop-action and continuous-action sports. In Section 3, we develop a new algorithm for soccer in order to demonstrate the application of the model to continuous-action sports. Experiments are presented in Section 4. We conclude with discussion and conclusions in Section 5.

2. MODELING SPORTS VIDEO USING EVENTS

In many types of sports, although a typical game broadcast lasts say, a few hours, only parts of this time contain real actions. These important parts occur semi-periodically but sparsely during the game. The remaining time is typically less important (e.g., change of players, time-outs, etc). These types of sports are so-called "action-and-stop" sports. For an action-and-stop sport, if all actions have been extracted, then a user can follow, understand, and even enjoy the game by viewing only the action clips. Examples include baseball and American football, where all pitches and all plays, respectively, contain every detail that a user wants to know about the underlying game. On the other hand, there are also many other sports that are almost continuous without any break, and thus it is impossible to cut out any portion of the video without compromising a user's ability for grasping every detail of the game. However, for this type of sports, since the video is full of consecutive actions, what a user would desire is some reasonable classification of the actions, which allows access to those parts of the video that are more interesting or exciting than other parts of the video. Examples of continuous-action sports include soccer, ice hockey, etc.

In consideration of the above distinction between action-and-stop sports and continuous-action sports, we extend our earlier work [3] (which is for the former type of sports) by introducing a general modeling, using "event" in an abstract sense. We model the video as a sequence of "events" interleaved with "non-events", with "event" being defined as the basic segment of time during which an important or exciting action occurs, as illustrated in Fig. 1. When instantiated for a specific sport, the event can be, for example, a pitch in a baseball game, and a goal or a goal attempt in soccer. Obviously, an event is a complete action and can contain multiple video shots; thus the modeling is at a higher level than the breakdown of video into shots.

For an action-and-stop type sport such as baseball, by defining event to be every pitch, the modeling in Fig. 1 effectively breaks down an input video into pitches and segments of idle time. In this case, the modeling is reduced to that proposed in [3].

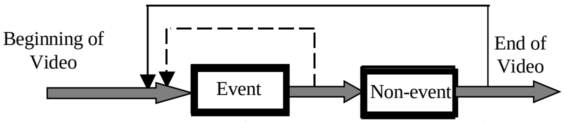

Figure 1. **A general model of a sports video in terms of "event". The inner loop (in dashed lines) indicates the possibility that two or more important actions can occur consecutively.**

For a continuous-action game, there is no break per se. Therefore, the problem becomes how to define event so that it represents those actions that are of greater importance and interest. While there are different methods for detecting "exciting" segments in video by using, for example, audio information, we claim that "excitement" is a subjective concept and thus is not easy to model, let alone to detect automatically by a computer. We propose a new approach to defining event in this case. We define important events as those actions in a game that are replayed by the broadcaster. By this strategy, we effectively shift the task of judging the importance/excitement of an action to the broadcaster, who is generally in a much better position to make this judgment than any automatic algorithm. This method assumes that the actions that are replayed by a broadcaster are typically more exciting or important than other actions. We believe that this is a reasonable assumption. In this case, the modeling in Fig. 1 is effectively reduced to that of Fig. 2.

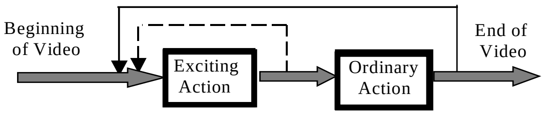

Figure 2. The effective modeling for continuous-action sports.

Obviously, compared with Fig. 1, here we know that event in Fig. 1 refers to "exciting action", and "non-event" refers to "ordinary (less exciting) action". It is for this reason, we have emphasized that event in the general modeling is used in an abstract sense, since there may be actual actions going on during a non-event period. For action-and-stop sports, even if we can use the play-centric modeling (as in [3]), it is also possible to use the modeling in Fig. 2. In the latter case, the modeling would focus on distinguishing, for example, exciting pitches from others in a baseball game, rather than detecting every pitch.

The proposed modeling is readily applicable to media database management applications, where common operations such as indexing, retrieval, logging, and annotation, etc. can all benefit from the breakdown of a video into the event and non-event segments. For example, events can form meaningful indexing points for semantic analysis, and video summarization can be achieved by concatenating the event segments.

2.1 Comparison with Earlier Work

Compared with earlier work in this area, the proposed modeling is (a) more general as it unifies two major types of sports, namely action-and-stop and continuous-action sports, and (b) it has obvious advantages as we describe below. For action-and-stop type of sports, the proposed approach is an extension of earlier work, where individual papers either handle a specific sport [4], or can handle only play/break type of patterns [3,5]. The proposed modeling is also different from prior work in the case of continuous-action sports ([1,2,6]) in that our modeling defines exciting actions based on the selection made by a production expert, namely on replays. Thus in terms of accuracy or meaningfulness in determining whether an action is exciting, our modeling has a unique advantage over the above mentioned methods that practically rely on the automatic understanding of the content by a computer. We rely on automatic detection of replay segments using low-level features and production patterns, as described in the next section, which is a much more suitable task for a computer than understanding the content.

3. EVENT DETECTION IN SOCCER VIDEO

With the modeling in Fig. 1 and Fig. 2, the major task of video analysis becomes the detection of all the events in a given video. While it is straightforward to show that the general framework can be specialized to action-stop type of sports such as baseball (see [3] for more discussion), in this section, we propose a novel event detection algorithm for broadcast video of soccer, which is used as an example for continuous-action sports. The algorithm is largely based on video/audio analysis of the input video, and they make use of two types of prior knowledge in extracting semantics from broadcast sports video: (1) domain knowledge, and (2) production knowledge. The detection algorithms represent these two types of knowledge in terms of rule bases where rules are expressed in terms of low-level visual and aural features that are automatically computed from the media.

As discussed in Section 2, for continuous-action sports, the event in the modeling of Fig. 1 is actually defined as exciting/important actions that have been picked out by the broadcaster for re-playing. In soccer, these exciting actions most likely contain goal attempts and surely contain goals. We utilize the knowledge of the established production patterns in soccer broadcast programs. The sequential relationship between an exciting action and its replay is shown in Fig. 3. An event includes an exciting action and the setup action leading to that action. An exciting action is usually followed by one or more close-up shots of the key player(s), and/or the audience, and/or the coach, and/or the referee, and then followed by the replay of the action. In other words, an exciting action has two versions, live and replay, with some close-up segments separating the two parts. We make use of this triplet production pattern in detecting the exciting events. In the following, we describe the detection of a replay segment, its associated close-up segment, and the associated exciting action segment, respectively. The overall diagram of the algorithm is shown in Fig. 4.

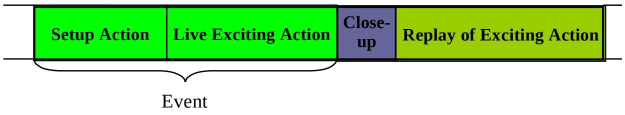

Figure 3. The relationship between an exciting action and its replay. Events are defined as the action that is replayed plus the setup action leading to that action.

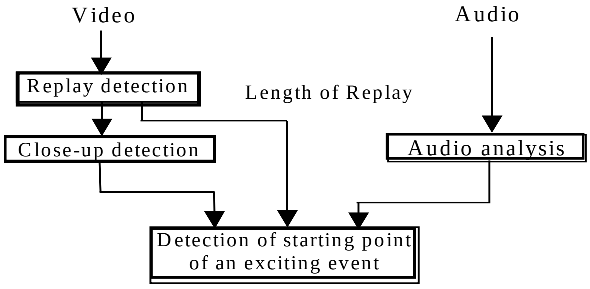

Figure 4. The general framework of the soccer algorithm.

3.1 Detection of replay segments

Replay detection is the first step in the algorithm. We have developed methods for detecting replays with sufficient reliability and accuracy ([7,8]). Due to space limitation, we can only briefly mention that the method of [7] is mainly based on the detection of slow-motion generated by field/frame repetition, and that the method of [8] is mainly based the detection of transitional logos that sandwich replay segments. More details can be found in these references.

3.2 Detection of close-up segments

We use two visual features to detect close-up segments: (1) the ratio of the number of pixels belonging to the soccer field to the total number of pixels in a frame; and (2) scene cuts, discussed respectively as follows.

(1) Ratio of the field pixels to the total number of pixels:
In close-up segments, players (or referees) are shot from a close distance. Therefore, the frame is not dominated by the colors of the soccer field. Otherwise, the field is shot from a long distance and thus a large portion of the frames is the soccer field, as shown in Fig. 5. The dominant color of a soccer field varies from time to time and game to game and can be calibrated ([8]). In computing the above ratio for distinguishing close-ups, we assume that the dominant color is green.

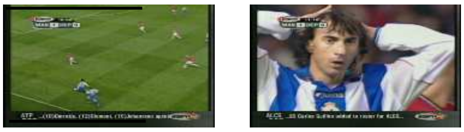

Figure 5. A frame from a live segment (left); and a frame from the associated close-up segment (right).

The color of green can be defined in any appropriate color space. In this paper, green is calculated in the normalized RGB color space, and only the normalized R and G color components, as defined below, are considered. This is intended to alleviate the effects of varying color in the case of different fields, weather conditions, camera settings, etc. A pixel is classified as being green if its normalized (*R'*,*G'*) components satisfy the following conditions: $R'<0.5$, $G'>0.5$, and $R'+G'\leq 1.0$, where

$$R' = \frac{R}{R+G+B}, \quad G' = \frac{G}{R+G+B} \qquad (1)$$

Fig. 6 shows a typical example of the computed ratio over a period of time. One can see that in normal shots, the ratio is close to 1. In close-up shots, the ratio is close to 0; in replay segments, the ratio varies.

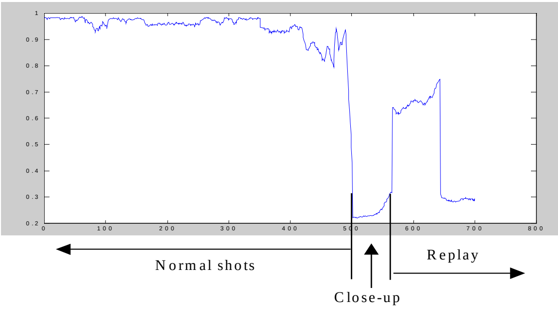

Figure 6. Ratio of the number of green pixels to the number of pixels in a frame.

(2) Scene cuts:
Usually there is a sudden scene cut between the live action and the close-up segment. This scene cut is at the end of the live action and at the starting point of the close-up segment. Sometimes, a close-up segment between the live action and its replay consists of several different shots of players, coaches, and audiences, and therefore consists of several scene cuts, as illustrated in Fig. 7. Close-up segments are characterized by the low ratio of the number of green pixels to the number of total pixels in a frame. The earliest scene cut in this time period with low {number of green pixels/total number of pixels} ratio is chosen as the starting point of the close-up segment and the end point of the live action. In the example shown in Fig. 7, it is the Scene cut 0.

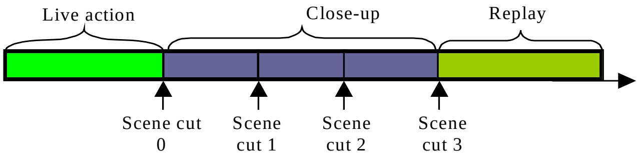

Figure 7. Demonstration of determination of the starting point of the close-up segment and the end-point of the live exciting action.

Any appropriate scene cut detection algorithm can be used. In our implementation, we utilize the difference between color histograms of every two adjacent frames. The color histogram is calculated in the normalized RGB color space and each of the normalized R and G color components is evenly divided into 16 bins. As a result, there are a total number of 256 bins. The difference between two color histograms is measured using the mean-square error (MSE) criterion. A large change (larger than a threshold computed dynamically from differences in recent past frames) in the color histogram difference identifies a scene cut .

3.3 Detection of starting points of exciting events

The last step of the proposed soccer algorithm is to detect the starting points of exciting event segments. Although the semantic content of the live version and replay version of an exciting action is the same, the camera angles used to capture them in the broadcast video are usually different. In other words, frames in replays look structurally very different from those in the live version. This is demonstrated in Fig. 8 using frames taken from a typical soccer broadcast. Even with the state-of-the-art content analysis technology, it is extremely difficult, if not impossible, to reach the conclusion that the contents of these frames belong to the same physical action, using purely only the video as input. Therefore, although the replay segment has been detected, one can hardly expect to find the corresponding live action based on content analysis. In the following, we propose a method for determining the starting point of the event, i.e., the starting point of the set-up action.

Due to the nature of soccer games, except for a few cases, such as a free kick and a corner shot, there is no objective way of defining the starting instances of exciting events. Besides, before exciting actions take place, the video shots in broadcast soccer games are typically unbroken. A shot usually lasts for several minutes without scene transitions. Therefore, one cannot use scene cuts or transitions in detecting the starting points of the event segments. In our algorithm, we choose the starting points of the event associated with the detected replay segments as either 15 or 30 seconds prior to the starting points of the close-up segments. We make use of the following two criteria in choosing between 15 and 30 seconds. If either of the two criteria favors 30 seconds, then we place the starting point at 30 seconds before the end point of the action. Otherwise we place the starting point at 15 seconds before the end point of the action.

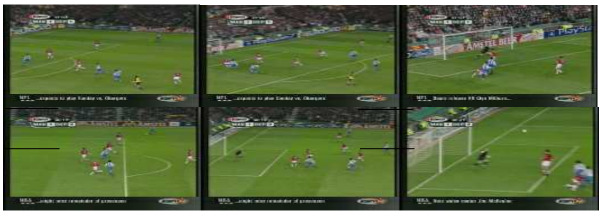

Figure 8. Images of a live broadcast of a goal (top row) and images of the corresponding replay (bottom row).

Criterion 1. Energy spectrum of the audio signal. An exciting action and its setup are typically accompanied by the audience's and/or the commentators' increased level of excitement, which is usually reflected in the audio tracks. In our experiments, the energy in the 1k~6.4kHz frequency band is calculated by taking FFT of the Hamming-window filtered input audio signal. If the average audio energy over the 30-second duration is larger than the average energy over the 15-second duration, then a 30-second long event segment is favored; otherwise, a 15-seconds long event segment is favored. (We also investigated the possibility of using the audio pitch contour to reflect the degree of the excitement of commentators. However, in our experiments, two commentators with significantly different pitch baffled the method)

Criterion 2. Duration of detected replay segments. We have observed that the lengths of replays are proportional to the excitement of the actions in the replays. When an action is more exciting, it is reasonable to give it more time to see how the event is built up. Thus if the duration of a detected replay exceeds certain threshold, a 30-second event segment is favored.

3.4 Generating the game summary

We have used the following four methods for generating highlight summaries:
(1) Concatenation of {exciting event segment + close-up segment + replay segment}, for all replays ;
(2) Concatenation of {exciting event segment + close-up segment}, for all replays;
(3) Concatenation of {exciting event segment}, for all replays;
(4) Concatenation of all replay segments.

4. EXPERIMENTAL RESULTS

We have performed extensive experiments by applying the proposed algorithms to soccer video from different broadcasters, and video captured from different recording sources. Excellent results have been obtained. Particularly, for two sample sequences of two different games, with a total duration 120 minutes, the algorithm detected all the five goals in the games. All other 21 detected actions were exciting goal attempts. The resulting summaries were evaluated by a group of soccer fans in our laboratory. They all agreed that the summaries provided them with all the relevant and exciting events and that their viewing experience was smooth in that the summary flowed naturally. (While hard-core fans will mostly prefer to enjoy the *live* broadcast at its entirety; they nay enjoy the summaries as a quick review of the game. In fact, in our tests, the fans enjoyed watching the summaries and the exciting parts of the game over and over again to recreate and share their excitement with others.) On the average, the summaries that include the replay segments were less than 18% of the original video in length, which means a user can review all the exciting events and their replays of a full game by spending only 16 minutes on watching the summary.

All the data used in our experiments are MPEG-encoded streams of 320x240 frame resolution, captured by an inexpensive TV-tuner PC card. All the algorithmic modules further reduce the input resolution to 160x120 before computation. This suggests that the proposed algorithms do not require very high quality input. In our experiments, the event detection algorithm achieved faster than 30-frames/second computational performances on a Pentium III-800MHz PC.

An MPEG-7 compliant prototype browsing system has been implemented to visualize the results, which will be demonstrated during the conference.

5. CONCLUSIONS

We have proposed a general framework for sports video analysis, which includes a unifying model for modeling both action-stop and continuous action sports. We have demonstrated the generic nature of the model by successfully applying it to different sports, and particularly to soccer in this paper. We have evaluated the performance of the proposed framework and algorithm through extensive experiments. Another important advantage of the proposed event based model is the fact that it facilitates synchronization and merging with independently generated rich metadata, which typically have an event-based granularity. Upon synchronization and merging, the resultant composite metadata contain not only video indexing points of event segments, but also independent rich metadata that are typically generated by human experts [9].

6. REFERENCES

[1] D. Yow, B-L. Yeo, M. Yeung, and B. Liu, "Analysis and Presentation of Soccer Highlights From Digital Video", *Proc. 2nd Asian Conference on Computer Vision*, 1995.

[2] Y. Rui, A. Gupta, and A. Acero, "Automatically extracting highlights for TV baseball programs," *Proc. ACM Multimedia 2000*, Los Angeles, CA, 2000.

[3] B. Li and M. I. Sezan, "Event detection and summarization in sports video," *Proc. IEEE Workshop on Content-Based Access to Video and Image Libraries*, 2001.

[4] T. Kawashima, K. Tateyama, T. Iijima, and Y. Aoki, "Indexing of Baseball Telecast for Content-based Video Retrieval", *Proc. IEEE-ICIP* 1998.

[5] D. Zhong and S-F. Chang, "Structure Analysis of Sports Video Using Domain Models", *Proc. IEEE International Conf. on Multimedia and Expo*, 2001, Tokyo, Japan.

[6] L. Xie, S.-F. Chang, "Structure analysis of soccer video with hidden Markov models," *Proc. IEEE-ICASSP 2002.*

[7] H. Pan, P. van Beek, and M.I. Sezan, "Detection of Slow-motion Replay Segments in Sports Video for Highlights Generation", *Proc. IEEE-ICASSP*, 2001.

[8] 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 ICASSP 2002*.

[9] B. Li, J. Erricco, H. Pan, and M. Ibrahim Sezan, "Bridging The Semantic Gap in Sports Video", *Proc. of SPIE/IS&T Electronic Imaging*, January 2003.