Current and Emerging Topics in Sports Video Processing

Xinguo Yu
Institute for Infocomm Research,
21 Heng Mui Keng Terrace, Singapore 119613
xinguo@i2r.a-star.edu.sg

Dirk Farin
Technical University Eindhoven
P.O. Box 513, 5600 MB Eindhoven, Netherlands,
d.s.farin@tue.nl

Abstract

Sports video processing is an interesting topic for research, since the clearly defined *game rules* in sports provide the rich domain knowledge for analysis. Moreover, it is interesting because many specialized applications for sports video processing are emerging. This paper gives an overview of sports video research, where we describe both basic algorithmic techniques and applications.

1. Introduction

Sports broadcasts constitute a major percentage in the total of public and commercial television broadcasts. The growing demands of the viewers require advances in video capturing, storage, delivery, and video processing ability. With the advent of large storage capabilities and more TV channels with full coverage of large sport events, the organization and search in these data-sets becomes more appealing. This situation poses a challenging research problem: how to quickly find the interesting video segments for various consumers with differing preferences. In other words, the consumers want a system that allows them to retrieve specific segments quickly from the huge volume of available sports video, thus saving time. Other growing demands of viewers are new enhancement and presentation techniques that provide a better viewing experience, or that even generate the feeling to actually take part in the sports event, instead of only watching a transmission of a video captured by a single camera. These presentations require research in the fields of reconstruction, virtual reality generation, enrichment, etc.

Sports videos have a game-specific structure because sports games are under rules and regulations. Furthermore, most of the games take place in restricted playfields with a defined layout. With the specific domain knowledge, sports video analysis can provide results of higher semantic levels compared to the analysis of other kinds of videos, for example, arbitrary home video. The playfield also provides a great aid in reconstructing the sports video using computer vision technology. Since a playfield is restricted, several fixed cameras can cover the whole playfield to capture enough information for a 3D (*three-dimensional*) reconstruction of the sports event. We can coarsely classify sports video research into the following two main goals:

- **Indexing and Retrieval Systems**: to build indexing and retrieval systems that can find requested sports scenes based on high-level semantic queries.
- **Augmented Reality Presentation**: to present the sports videos with additional information to provide new viewing experience to the users. A goal is to give the impression that the viewer is virtually part of the event.

In the following sections, we first describe some basic algorithmic building blocks that reappear continuously in different contexts. Subsequently, we describe the main applications of sports video processing, and we close with some concluding remarks. Because of the vast amount of research conducted in this area, the presented topics and references can only be a restricted selection.

2. Algorithmic Building Blocks

2.1. Event and highlight detection

Event detection examines the sports video for events like goal-shootings, fouls, scoring a point, etc. Usually, various features of the video are used to infer the occurrences of events. The analysis of events can be carried out by combining various attributes of the video, including its structure, events, and other content properties. The content analysis problem can be addressed in two steps. First, we have to identify relations between the features and the semantic concepts that we want to detect. Second, we must find the procedure that captures the relation between the features and the semantic concept.

Highlights in sports videos are the special events that the viewers are especially interested in. The definition of what is interesting differs for each viewer. While sport fans are more interested in events like goals or spectacular points in tennis, the coaches might be more interested in the errors of the players to help them improve their capabilities. Highlight detection is the process to find the occurrences of highlights and their starting and ending times in the video.

In event detection, features can be extracted from three channels: video, audio, and text. Each channel provides some cues that correlate with the occurrence of events. In other words, we can detect the events through inferring from these cues. We briefly describe the different approaches and methods of designing the event detection algorithms by pointing out opposing concepts below.

Low-Level Feature vs. **Object-Related Feature**: Low-level features are the *cinematic* features [38], acquired directly from the input video by using simple feature extractors. Examples are the dominant color, dominant texture, motion vectors, audio volume, keywords of text, etc. Many algorithms have been developed [1, 28, 32, 37] based on these low-level features. The low-level feature-based algorithms are popular because they are simple and efficient. However, for many events we cannot achieve satisfactory results by using such algorithms, because the events have a complex semantic nature and there are no clear relations between low-level features and these semantic concepts.

0-7803-9332-5/05/$20.00 ©2005 IEEE

Object-related features are attributes of objects such as ball locations and player shapes, acquired by more complex algorithms. Event detectors based on object-related features often use both object-related and low-level features [13, 40]. An important object-related feature is the position of players and of the ball (see below). For example, a goal in a soccer game is defined as the event that the ball passes one of the goalmouths on the soccer field. Also a growing number of algorithms were designed to detect events with the aid of ball location or ball trajectory. Object-related features are usually difficult to be obtained [6, 11-12, 39-40, 42], but more events can be detected by using object-related features.

Single Channel vs. **Multiple Channels**: Some events can be detected using the features from a single channel [37]. However, for most events, we have to combine the features from multiple channels to detect the specific events [3, 8-9, 20, 22, 35-36]. The usually available channels are video and audio; the text channel can be available in the case of analysis of the commentary text in the internet, the superimposed text, and commentary speech [35, 45].

Game-Specific vs. Generic: Most of the developed event-detectors are game-specific because different games have different events. These detectors need to use the domain knowledge of the particular game to infer the events [22, 39-41]. Obviously, it is better if an algorithm can detect the events of multiple games. Unfortunately, such algorithms are hard to attain. A compromise to design such systems is to build frameworks that can be configured to different games [8, 11-12].

Feature Model vs. Shot Pattern: Some algorithms directly model the relations between the features and the events. Such algorithms are intuitive and efficient. However, some events have no intuitive relations with the features but require context information for the detection. One kind of algorithms uses the shot patterns to capture the context information [7-8]. They take four steps to detect the events. First, they segment the video into shots. Second, they extract the features within each shot. Third, they classify the shots into the predefined categories. Fourth, they infer events using a procedure that models the shot patterns.

Learning-Based vs. Normal: Most current algorithms use hard-wired procedures to capture the relations between the features and the events. They are efficient but not generic. Learning-based algorithms capture the relations between the features and the events by statistical analysis and optimization [34, 37]. In general, the machine learning approach is chosen when it is difficult to intuitively define the relations between the features and the events.

Broadcast Video vs. Non-broadcast Video: In the literature, most of the algorithms were designed for broadcast sports video from which they can use the editing information as an additional source of information. For example, such algorithms can use the information from the text, speech, and the graphics (with the known meanings) superimposed in the video [45]. The algorithms for non-broadcast video have only video and audio before editing [33]. Hence, they face more research challenges because they have less information to be used. However, the on-site systems for sports have to use such algorithms because they have to produce the results prior to broadcasting.

2.2. Structure Analysis

Video structure denotes the temporal grouping of the video. An example for a sport with a clear temporal structure is tennis. Each match comprises several sets, which are again divided in games, which are built from points. Other sports like soccer have less structure.
In sports video research, video structure analysis is not equivalent to simple cut detection, but it combines the knowledge about the rules of the game with general video structure analysis methods. An important structure is the distinction between play and break periods, because the consumers can save a lot of time by skipping the break periods [34].

2.3. Object Detection and Segmentation

In sports videos, the important acting objects are well defined. In some sports, only the position of the players is of interest (soccer, tennis), for other sports like gymnastics or diving, the shape and movements of the athletes is the main interest. Automatic analysis of these sports requires automatic segmentation algorithms that provide exact object shapes and algorithms that can infer behavior from the object shape [11-12]. These accurate object boundaries are also required for synthesizing new views in a 3D reconstruction [19].

2.4. Ball and Player Tracking

Because of the importance of the ball and player locations, specialized algorithms have been developed to track their position. Especially the tracking of a small ball is a difficult problem [7] and algorithms have been developed that detect the ball trajectory through several frames in one step [39-40] to increase the algorithm robustness.

2.5. Camera Calibration

Tracking the players only determines their position in the image coordinate system, but does not provide information about their position on the playfield, or more generally in the real world. To determine the geometric mapping between the image coordinates and the real-world positions, the camera parameters are required. Since the playfield itself is usually marked with certain lines whose real-world position is known, these lines can be used to carry out the camera calibration [11-12, 26].

2.6. 3D Reconstruction

One step further, multiple cameras can be used to obtain views from different position and compute a complete 3D reconstruction of the playfield [37]. Usually, fixed multiple cameras are used, which simplifies their calibration, reduces the computing time, and improves the accuracy of camera calibration. More importantly, they can cover the whole playfield while current broadcast sports videos only display a portion of large playfields such as soccer fields.

3. Applications of Sports Video Processing

3.1. Abstracting

Video abstracting is the process of creating a shortened version of a video that still comprises the essential information from the complete video. The video abstract is a very important tool for sports video, as several periods during some games may be boring to the consumers and watching the abstract can save a lot of time. Generating a video abstract seems to be strongly related to finding the highlight events, but techniques have been proposed that estimate the excitement of a scene directly. Several researchers have used audio features to detect the highlights, because audio contains information about audience applause and of the excitement in the voice of the commentator. For example, Rui *et al.* [28] used audio features such as excited speech and baseball hits to detect the highlights.

Instead of synthesizing a fixed abstract, an alternative approach is to use the structural analysis to detect semantic units. These can be used to build video indices and the table of contents according to different categorization schemes. The advantage of this approach is that the user has more control about what he wants to view and the probability to miss an interesting scene is smaller.

3.2. Tactics and performance analysis

Tactics analysis is to understand the tactics that teams or individual players have used. Performance analysis is to evaluate the performance of a team or a player through analyzing their motion and activity in games. Coaches and players are interested in such results for improving their performance in later games. The consumers are interested in such results for enjoying sports video with additional statistical information. The main tasks in tactics and performance analysis are to find the traces and the actions of players. Past work includes team behavior analysis [18, 29], team and player possession analysis [41, 43], player and ball detection, and tracking [19, 39-40]. The essential algorithmic techniques for tactics analysis are ball and player tracking as well as camera calibration [11-12] to find the corresponding real-world position of the players based on their position in the video.

3.3. Augmented Reality Presentation of Sports

The presentation of sport videos can be improved with two basic techniques. One is the 3D reconstruction of the game to provide arbitrary views and thus to increase the viewing experience to consumers [4, 19, 21]. The second technique is to insert some illustrations into the original video or to provide the illustration in extra windows to help consumers to understand the video easier [25, 44]. An example that is already used in practice is to superimpose virtual lines onto the field, for example, the distance between a player and the goal in football [46]. Another kind of enrichment is to insert advertisements into the original video to improve the commercial value of the video [31].

Sports video reconstruction provides 3D video to consumers. The advances in computer vision, computer graphics, and image processing have provided the techniques to build 3D reconstruction systems for sports. 3D reconstruction systems have been designed to serve different purposes. For example, they can be designed to generate virtual 3D views such that a viewer can examine a scene from the viewpoint of the referee or a player on the field [6, 15, 19, 21]. They can also be designed to generate 3D reconstructed video such that it gives a better viewing experience than the original video [44].

3.4. Sports video for small devices

The transmission of sport events on small devices like PDA, 3G/UMTS phones will become more popular in the future. Because the screen size of these devices is limited, different content is required, compared with TV broadcasts [32]. Since the production of special content is expensive, automatic transcoding applications that, for example, focus the view to the most important image area, will be employed. Because small devices have only a narrow transmission bandwidth, it is likely that a class of special compression algorithms for small devices will be developed in the near future.

3.5. Referee assistance

Referee assistant systems target to help the referee in cases of difficult decisions and partially replace the work of the referee. Such systems can use either specialized electrical sensors, or they can use more flexible real-time video analysis systems. Currently, some tennis championships like Wimbledon, US Open and Australian Open use video systems to display the landing positions that are very close to the boundary lines of the court, where the technology is provided by Hawkeye [47]. Hence, such systems are useful to help the referee to decide if a doubtful decision was correct. For other sports, like an offside position in soccer, some decisions cannot be decided at fixed lines, but should be detected at varying positions on the field. These cases require a more complex and powerful video solution. We expect that referee assistance systems will become more important in the future since they ensure an unprejudiced decision and they support the large commercial value.

4. Concluding Remarks

This paper has discussed the main techniques and approaches in sports video research. The specialty of sports video processing lies in the available domain knowledge of sports. A lot of work has already been carried out on content analysis of sports videos, and the work on enhancement and enrichment of sports video is growing quickly due to the great demands of customers. The development of video analysis systems still concentrates on the extraction of low-level features. We believe that the future direction of research will concentrate on a more specialized, but also more in-depth analysis. For the enrichment and 3D reconstruction of sports events, current systems exist for specialized applications. However, more general systems that are adaptable to cover a larger variety of applications will be an important design goal for future systems. Probably, more emerging topics will appear in enhance the quality and value of the sports video in the near future.

5. References

[1] J. Assfalg, M. Bertini, C. Colombo, A. Del Bimbo, Semantic annotation of sports videos, IEEE Multimedia, vol.9, p.52-60, April 2002

[2] J. Assfalg, M. Bertini, C. Colombo, A. Del Bimbo, and W. Nunziati. Semantic annotation of soccer videos: automatic highlights identification, Computer Vision and Image Understanding, vol 92, pp295-305, 2003.

[3] N. Babaguchi, Y. Kawai, and T. Kitahashi, Event based indexing of broadcasted sports video by intermodal collaboration, IEEE Trans. on Multimedia, 4(1): 68-75, 2002.

[4] T. Bebie and H. Bieri, A video-based 3D-reconstruction of soccer games, Eurographics, vol. 19 (3), 2000.

[5] S. F. Chang. The Holy Grail of content-based media analysis, IEEE Multimedia, vol. 9:2, 6-10, 2002.

[6] T. D'Orazio, C. Guarangnella, M. Leo, and A. Distante. A new algorithm for ball recognition using circle Hough transform and neural classifier, Pattern Recognition, vol. 37, pp393-408, 2004.

[7] L. Duan, M. Xu, and Q. Tian, semantic shot classification in sports video, SPIE: (SRMD) 03, pp. 300-313, 2003.

[8] L. Duan, M. Xu, T. S. Chua, Q. Tian, and C. Xu, A mid-level representation framework for semantic sports video analysis. ACM Multimedia 2003: 33-44.

[9] A. Ekin, A.M. Tekalp, and R. Mehrotra. Automatic soccer video analysis and summarization, IEEE Trans. on Image Processing, Vol. 12:7(2003), 796-807.

[10] D. Farin, T. Haenselmann, S. Kopf, G. Kühne, and W. Effelsberg, Segmentation and classification of moving video objects, in B. Furth and O. Marques (eds.): "Video databases, design and applications", CRC Press, 2003.

[11] D. Farin, J. Han, and P.H.N. de With, Fast camera calibration for the analysis of sport sequence, ICME 2005.

[12] D. Farin, S. Krabbe, W. Effelsberg, and P.H.N. de With, Robust camera calibration for sport videos using court models, SPIE, vol. 5307, pp. 80-91, 2004.

[13] Y. Gong, T. S. Lim, H. C. Chua, H. J. Zhang, and M. Sakauchi, Automatic parsing of TV soccer programs, 2nd Int. C. on Multimedia Comp. and Sys., pp.167-174, 1995.

[14] M. Han, W. Hua, W. Xu, Y. Gong, An integrated baseball digest system using maximum entropy method, ACM MM 02, Dec 01-06, 2002, Juan-les-Pins, France

[15] N. Inamoto and H. Saito. Free viewpoint video synthesis and presentation from multiple sporting videos, ICME 2005.

[16] N. Inamoto, H. Saito. Immersive observation of virtualized soccer match at real stadium model, 2rd IEEE & ACM Int. Symp. on Mixed & Aug. Reality, pp.188-197, 7-10 Oct. 2003, Japan

[17] S. Iwase and H. Saito, Tracking soccer players based on homography among multiple views, SPIE, vol.5150, pp283-292, 2003.

[18] T. Kawashima, K. Yoshino and Y. Aoki. Qualitative image analysis of group behavior, CVPR 94. pp. 690-693, 1994

[19] T. Koyama, I. Kitahara, and Y. Ohta. Live 3D video in soccer stadium, 2rd IEEE & ACM Int. Symp. on Mixed and Aug. Reality, pp.178-186, 7-10 Oct. 2003, Japan.

[20] W.-N. Lie and S.-H. Shia. Combining caption and visual features for semantic event classification of baseball video, ICME 2005

[21] K. Matsui, M. Iwase, M. Agata, T. Tanaka, and N. Ohnishi. Soccer image sequence computed by a virtual camera, CVPR 1998.

[22] S. Nepal, U. Srinivasan, G. Reynolds, Automatic detection of 'goal' segments in basketball videos, ACM MM01, pp261-269, 2001.

[23] Y. Ohno, J. Miura and Y. Shirai. Tracking players and estimation of 3D position of a ball in soccer games, ICPR 00, vol. 1, pp. 145-148, 3-7 Sept. 2000.

[24] G. S. Pingali, Y. Jean, and I. Carlbom. Real time tracking for enhanced tennis broadcasts, CVPR, pp.260-265, 1998.

[25] G. S. Pingali, A. Opalach, Y. Jean, and I. Carlbom. Visualization of sports using motion trajectories: providing insights into performance, style, and strategy. Proc. 12th IEEE Visualization Conf., pp. 75-82, 544, San Diego, Oct. 2001.

[26] I. Reid, and A. Zisserman, Goal-directed video metrology, ECCV 96 (vol II), LNCS 1065, pp. 647-658, 1996.

[27] J. Ren, J. Orwell, G. A. Jones, and M. Xu. A general framework for 3D soccer ball estimation and tracking, ICIP04, pp.1935-1938, 2004, Singapore.

[28] Y. Rui, A. Gupta, A. Acero, Automatically extracting highlights for TV Baseball programs, ACM MM00, pp105-115, Oct. 2000.

[29] T. Taki, J. Hasegawa and T. Fukumura. Development of motion analysis system for quantitative evaluation of teamwork in soccer games, ICIP 96, pp.815-818, 1996.

[30] V. Tovinkere and R. J. Qian. Detecting semantic events in soccer games: Towards a complete solution, ICME01, pp.1040-1043, 2001.

[31] K. W. Wan, X. Yan, X. Yu, and C. Xu. Robust goalmouth detection for virtual content insertion, ACM MM03, pp 468-469, 2003.

[32] K. W. Wan, X. Yan, and C. Xu. Automatic mobile sports highlights, ICME 2005

[33]. J. Wang, C. Xu, E. Chng, K. W. Wan and Q. Tian. Automatic replay generation for soccer video broadcasting, ACM MM 04, pp. 32-39, 2004.

[34] L. Xie, P. Xu, S. F. Chang, A. Divakaran, and H. Sun. Structure analysis of soccer video with domain knowledge and hidden Markov models, P.R. Letter, vol 25, pp767-775, 2004.

[35] Z. Xiong, R. Radhakrishnan, A. Divakaran, and T. S. Huang. Highlights extraction from sports video based on an audio-visual marker detection framework, ICME 2005.

[36] H. Xu, T.-H. Fong, and T.-S. Chua. Fusion of multiple asynchronous information sources for event detection in team sports video, ICME 2005.

[37] M. Xu, N. C. Maddage, C. Xu, M. Kankanhalli, and Q. Tian. Creating audio keywords for event detection in soccer video, ICME 2003, Vol II, 281-284.

[38] Y.-Q. Yang, Y.-D. Lu, and W. Chen. A framework for automatic detection of soccer goal event based on cinematic template, Int'l Conf. on Machine Learning and Cybernetics, Shanghai, 26-29, Aug. 2004.

[38] X. Yu, C. H. Sim, J. R. Wang, and F. C. Loong. A trajectory-based ball detection and tracking algorithm in broadcast tennis video. ICIP04, pp1049-1052, 2004.

[39] X. Yu, C. Xu, H. W. Leong, Q. Tian, Q. Tang and K. W. Wan. Trajectory-based ball detection and tracking with applications to semantic analysis of broadcast soccer video, ACM MM03, pp. 11-20, 2003.

[40] X. Yu, H. W. Leong, J. H. Lim, Q. Tian, and Z. Jiang. Team possession analysis for broadcast soccer video based on ball trajectory, PCM 03, pp. 1811-1815, 2003.

[41] X. Yu, H. W. Leong, C. Xu, and Q. Tian. A robust and accumulator-free ellipse Hough transform, ACM MM04, pp. 256-259, 2004.

[42] X. Yu, T. S. Hay, Xin Yan, and E. Chng. A player-possession acquisition system for broadcast soccer video, ICME 2005.

[43] X. Yu, Xin Yan, T. S. Hay, and H. W. Leong. 3D reconstruction and enrichment of broadcast soccer video, ACM MM04, pp260-263, 2004.

[44] D. Zhang, S. F. Chang, Event detection in baseball video using superimposed caption recognition, ACM MM02, Dec. 01-06, 2002, Juan-les-Pins, France

[45] Sportvision Inc., "System for enhancing a video presentation of a live event", U.S. Patent 6,597,406, January 26, 2001,

[46] http://www.hawkeyeinnovations.co.uk.