EVENT DETECTION BASED ON NON-BROADCAST SPORTS VIDEO

Jinjun Wang[1, 2], Changsheng Xu[2], Engsiong Chng[1], Xinguo Yu[2] and Qi Tian[2]

[1]CeMNet, School of Computer Engineering, Nanyang Technological University, Singapore 639798
jjwang@pmail.ntu.edu.sg, aseschng@ntu.edu.sg
[2]Institute for Infocomm Research, 21 Heng Mui Keng Terrace, Singapore 119613
xucs@i2r.a-star.edu.sg

ABSTRACT

Most recent research work on sports video analysis focuses on broadcast video. The broadcast video is a post-produced video and has much additional editing information inserted. In this paper, we propose a novel event detection framework in sports game only based on the video taken by single (main) camera. Compared with event detection from broadcast video, the proposed framework is more challenging as interesting events needs to be automatically detected. The results of the research could be used for automatic replay generation in sports video. In this paper, a mid-level representation is first created based on low-level audio/visual feature and events related to the replay are then detected from this representation. We experiment our framework and some promising results are obtained.

1. INTRODUCTION

Identifying interesting events in sports video based on image, video and audio analysis techniques had attracted increasing research attentions in recent years. Among different sports types, soccer game gained more attention because of its commercial potential. A lot of work had been done for broadcast soccer video analysis. In [1] image features were analyzed and events like "Shooting", "Yellow/Red card" and "Penalty" were identified. In [2] a multi-level sports event detection system was proposed and both visual and audio information were used. Encouraging results were obtained from this model that had also been extended to other games. An overview of related work can be found in [3].

The above research efforts can ease the process of annotating the huge amount of live and achieved video material, facilitating the management of multimedia document, such as summarization and retrieval. However, they focus on broadcast video and rely on information from post-production, e.g. appearing of replay, multi-camera transition and caption. Thus these techniques are limited, e.g. they are not suitable for automatic sports video editing.

This has motivated us to build up an event detection framework using only the video recorded from single (main) camera. This research is challenging as analysis is now performed using features from main camera, post-production information is unavailable, e.g. we can not make use of the shot transition pattern to identify highlight as there is only one shot (i.e. the long shot) in the single camera video. In addition, a soccer game has loose structure so the detection of semantic event is difficult. Lastly, soccer video is always "noisy" – low level visual and audio features extracted are often affected by many factors such as audience noise, weather, luminance, etc. which further complicates analysis.

The paper is organized as follows. Section 2 gives the framework of the proposed approach. Section 3 and 4 describe the creation of mid-level representations and the detection of high-level events, respectively. Performance of every level of the model is tested in section 5. With the result, conclusions are made and future works are raised in Section 6.

2. FRAMEWORK

In analyzing broadcast sports video, multi-level and multi-model systems had been proved to be more robust in

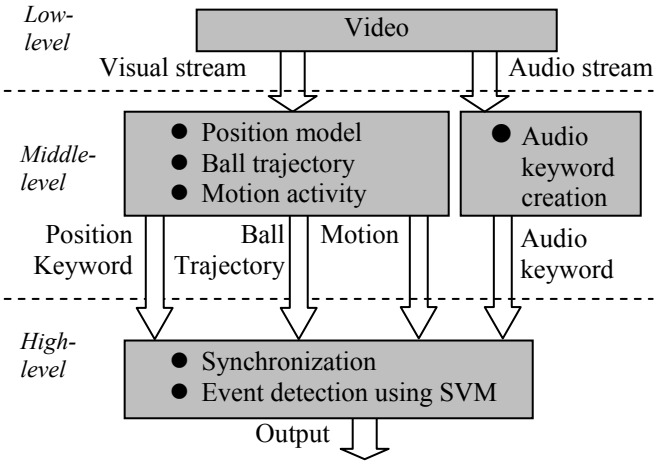


Figure 1: flowchart of our proposed model

bridging the large gap between low-level features and high-level semantics events [2]. We adopt this idea and make use of low-level features from both visual and audio domain. The proposed framework has 3 layers. In the low-level layer, video is divided into visual and audio streams and visual, motion and audio features are extracted. Mid-level representations are created from these features. Such representations, or called keywords, have had some coarse semantic meaning. In the high-level, Support Vector Machine (SVM) based classifier is applied to the keywords to detect important events. Figure 1 gives a diagram of our proposed framework.

3. MID-LEVEL REPRESENTATION CREATION

3.1 Position model

Since the main camera is always trying to capture the ball, to localize the view and/or register the current frame on the soccer field model can indicate the position where event happen. This is achieved by line detection, goalpost detection and center ellipse detection. In [1] three parallel field lines were detected to facilitate the penalty box area detection. In [4] the whole field is divided into 15 parts. Here we also divided the whole playground into 15 symmetrical areas as demonstrated in figure 2.

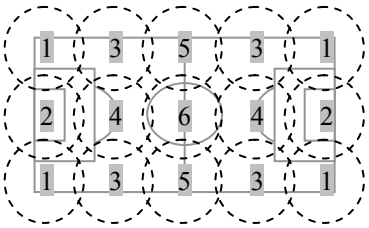
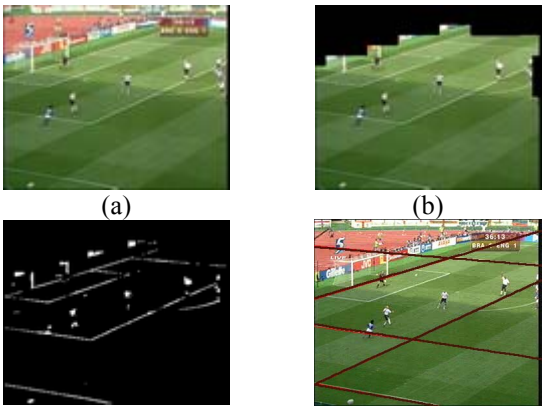


Figure 2: 15 areas on field model

Line detection: Each frame is first divided into blocks, 16x16 pixels in size. Dominant color analysis is then applied and blocks with less than half green pixels are black out, otherwise the block remains unchanged. A pixel with $(G - R) > T_1$ and $(G – B) > T_1$ is deemed as a green pixel, where R,G and B are the three color components of the pixel in RGB color space, and the threshold T_1 is empirically set to 20 in our system to be applicable for most soccer playground condition in our video database.

The color image is then converted to gray scale and edge detection is applied using Laplace-of-Gaussian (LOG) method. To reduce the effect of unbalanced luminance, the LOG edge detection threshold T_2 is updated adaptively for each block. An initial small threshold is used which is allowed to increase until no more than 50 edge pixels generated from each block (Typically a line such as field-line will generate 50 edge pixels within a 16x16 block). The edges are then thinned to 1 pixel width and finally the Hough Transform (HT) is used to detect lines. Fig. 3 illustrates the field-line detection process where the lines are highlighted in red (Figure 3d). The lines detected in each frame are represented in polar coordinates.



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

Figure 3: (a) frame with unbalanced luminance
(b) green region (c) edge detection (d) line detection

Goalpost detection: Although the goalpost can be detected by its two vertical goal posts using Hough transform, we have instead adopted a color based detection algorithm similar to [5]. The difference in our approach is that after the non-white pixels are removed, we applied heuristic constrains to locate goal post pixels. These constrains include:

i) White vertical lines should be long enough;
ii) White vertical lines should not be too long;
iii) There should be two vertical lines near each other;
iiii) These two lines should not be too near;

A detected goalmouth is shown in Figure 4 below (goal posts are highlighted in blue):

Figure 4: goalmouth detection

Center ellipse detection: Detection of center ellipse can help to determine the vertical position of the frame. We adopt the least-squares fitting method in [8] to detect the center ellipse.

Representation of position: As the low-level results from the video are always “noisy”, a Competition Network (CN) is used to filter out the noise and generate correct position keyword. The network has 15 nodes and each stands for a position in Figure 2. The coordinates of the detected lines, goalmouth position and center ellipse position form a vector as the input to the network. Each node generates a

response and proportionally prohibits all the other nodes. Once the final output of any node is above a threshold, a location related with the node that generates the highest output is claimed. Otherwise, the location remains the previous value. This is analogy to human precept: once we see enough things from the video, we get to know what position it is.

3.2 Ball trajectory

Detected and tracked position of the ball is a strong factor to recognize some events. For example, the relative position between the ball and goalpost indicates the events such as "goal" and "shooting" In this paper, the ball trajectory is an important mid-level representation for event detection described in section 4.

The ball trajectories are obtained by a novel trajectory-based ball-detection-and-tracking algorithm, which was presented in our previous work [6]. Unlike the object-based algorithms, this algorithm does not evaluate whether a sole object is a ball. Instead, it evaluates whether a candidate trajectory (a ball-trajectory candidate), which is generated from the candidate feature image by a candidate verification procedure based on Kalman filter, is a ball trajectory.

3.2 Motion activity

Motion pattern is closely related the semantic state of the game. E.g., the camera will stop moving when serious injuries occur, resulting in a period of no/low motion.

The motion information is readily available in compressed video. In our approach, we extract motion magnitude and direction features from P frames. For motion magnitude, average value of the motion vector from each MPEG block is used. The motion magnitude is then labeled as: no motion, low motion or high motion. For motion direction, an 8-bin motion histogram is created and the dominant direction is computed.

3.3 Audio keyword

There are some game-specific sounds such as applause, whistling that have strong relationships with the actions in the game. In our previous work [7], audio keywords like "Acclaim", "Whistle", "Commentator Speech" and "Silence" are created for sports game. The single camera video also contains audio information, thus in our model we created three audio keywords, namely "Acclaim", "whistle" and "Other" for high-level analysis.

The audio track is segmented using a sliding window of 20ms with no overlapping. Mel Frequency Cepstral Coefficients (MFCC) and Liner Prediction Coefficient (LPC) are extracted for each segment [7]. These features are sent to a Support Vector Machine (SVM) classifier to label the related segment with one of the above labels.

Since sports audio is "noisy", post-process of audio keywords sequence is applied. We filter out sudden changes in audio keyword that are considered as an error, using majority voting in 5 consecutive frames.

4. EVENT DETECTION

Our framework is tasked to detect the following events: "Goal", "Foul" and "Other" (player injure or sudden event). These events are selected as they are normally replayed in broadcast video hence is important to broadcaster.

To detect events from mid-level result, in [2] some heuristic rules are defined based on domain knowledge to map mid-level representation into high-level event. In [9] Hidden Markov Model (HMM) is used to detect events from mid-level keywords sequence. Besides, other classifiers like NN, SVM, Dynamic Bayesian Network (DBN) etc, are available in literature.

Although heuristic rules can sometimes achieve good results, it is not a generic approach hence it makes such system less robust. While HMM is good at managing temporal patterns, we notice that temporal pattern either does not exist in some low-level feature domains, or is heavily affected by noise in mid-level representation, thus resulting in HMM being not applicable in this problem.

Noting that certain events happen accompanying the appearance of certain patterns, e.g. "whistle" when "Foul", "long period of no/low motion" when "Injure", we choose SVM to detect events from mid-level keywords. SVM uses statistical learning and one of its major advantages is that it is robust to noisy data.

After the mid-level representations are created, synchronization operations are applied. Since audio keywords are created based on a smaller sliding window size than those based on video frame rate, we synchronize these different domain keywords according to time. The mid-level keywords are then put together to form a higher dimension feature vector for SVM. Since we use different set of keywords for different event, several SVM classifiers are used to detect the respective event.

5. EXPERIMENT

All the videos used are broadcast video of world cup 2002. We have edited them (to remove replays, close up, etc) to make them equivalent to the video taken by the main camera.

5.1 Accuracy of mid-level representation creation

Position model: totally 10 minutes videos (25 fps) are manually labeled. We then compare it with the result from

our position keyword creation model. The 15 symmetrical areas are labeled with 1 to 6 (as in Figure 2). The performance is listed in table 1 below:

Position	Precision	Position	Precision
1	86.3%	2	89.6%
3	84.2%	4	49.9%
5	77.7%	6	100.0%

Table 1: Accuracy of position model

Ball trajectory: The test is conducted on 15 long shot sequences (176 seconds in total). These sequences are representative in that they have different length, view type and ball appearance. Table 2 shows the overall tracking accuracy and more detailed results can be found in [6].

#det. & tracked	# false positive	accuracy
4283	25	98.8%

Table 2: Accuracy of ball trajectory

Motion activity: Since motion features are extracted directly from compressed video, it is objective and no more experiment is carried on it.

Audio keyword: We use 30 minutes audio to evaluate the audio keyword creation accuracy, 50%/50% as training/testing set. The classification result is shown in table 3 below:

	Acclaim	Whistle	Other
Accuracy	92.9%	89.8%	90.6%

Table 3: Accuracy of audio keywords creation

5.2 Event detection accuracy

We use 5.5 hours of video in this experiment, 3 types of event to be detected: "Goal", "Foul" and "Other". As mentioned in section 4, each event needs a SVM classifier for detection. In all the SVM classifier, radial basic function kernel is used:

$$k(x,y) = \exp(\frac{-|x-y|^2}{c^2}) \quad c = 8.0$$

Different set of features is heuristically chosen for the detection of respective event:

Goal: Position, motion and audio keywords are used.

Foul: Position and audio keywords are used.

Other: Position and motion keywords are used.

50%/50% of the whole video is used as training/testing set. The testing result is listed in table 4 below:

	Goal	Foul	Other
Event	4	105	12
Recall	100%	89.5%	80.0%
Precision	75.0%	73.4%	66.7%

Table 4: Accuracy of audio keywords creation

5.3 Discussion

The accuracy of mid-level representation creation is frame accuracy. In any event, few frames of error in mid-level representation do not cause the failure of detect this event to a certainty. This is because event detection normally needs several frames and single frame error can be rectified by time averaging. However, higher accuracy in mid-level representation can improve the performance of event detection.

6. CONCLUSION AND FUTURE WORK.

In this paper, a novel event detection framework in sports game is presented. The proposed framework allows sports highlight detection by visual and audio features from the single (main) camera video, thus extending the possible application areas to sports event detection system, such as coach video processing, broadcast video generation and sports broadcast management. The future work includes integration of more low-level features and mid-level representations to improve the performance, and implementation of an automatic replay generation system.

7. REFERENCES

[1] A. Ekin, A. Tekalp, and R. Mehrotra. "Automatic Soccer Video Analysis and Summarization", IEEE Trans. on Image Processing, vol. 12:7, pp: 796-807, 2003.

[2] L. Duan, M. Xu, T. Chua, Q. Tian, C. Xu, "A Mid- level Representation Frame-work for Semantic Sports Video Analysis", in Proc. of ACM Multimedia' 03, pp: 33-44. 2003.

[3] N. Adami, R. Leonardi, P. Migliorati, "An Overview of Multi-modal Techniques for the Characterization of Sport Programmes", Proc. SPIE – VCIP'03, pp. 1296-1306, July, 2003

[4] K. Wan, J. Lim, C. Xu, and X. Yu, "Real-Time Camera Field-View Tracking in Soccer Video", in Proc. of ICASSP '03, vol. 3, pp: 6-10, April 2003.

[5] D.Yow, B..Yeo, M.Yeung, and B.Liu, "Analysis and presentation of soccer highlight from digital video", In Proc. of Asian conf. on Comp. Vision (AVCV), 1995.

[6] X. Yu, C. Xu, H. Leong, Q. Tian, Q Tang, and K. Wan, "Trajectory-based Ball Detection and Tracking with Applications to Semantic Analysis of Broadcast Soccer Video", In Proc. of ACM Multimedia' 03, pp: 11-20, 2003

[7]M. Xu, N. C. Maddage, C. Xu; M. Kankanhalli, Q. Tian; "Creating Audio Keywords for Event Detection in Soccer Video", in Proc. of ICME '03. Proc. vol. 2, pp: 281 -284, July 2003.

[8] A W. Fitzgibbon, M Pilu, and R B. Fisher," Direct Least Square Fitting of Ellipses", IEEE Trans. on PAMI, vol. 21, pp: 476-480, 1999.

[9] J. Wang, C. Xu, E. Chng and Q. Tian, "Sports Highlight Detection from Keyword Sequences Using HMM", In Proc. of ICME'04, Taipei, June 2004.