GOAL DETECTION IN SOCCER VIDEO USING AUDIO/VISUAL KEYWORDS

Yu-Lin Kang[#], Joo-Hwee Lim[#], Mohan S. Kankanhalli[*], Chang-Sheng Xu[#], Qi Tian[#]*

[#]*Institute for Infocomm Research*
21 Heng Mui Keng Terrace
Singapore 119613
{yulin,joohwee,xucs,tian}@i2r.a-star.edu.sg

[*]*School of Computing*
National University of Singapore
Kent Ridge, Singapore 119260
mohan@comp.nus.edu.sg

Abstract

In this paper, we propose a two-level framework to detect interesting events automatically based on Audio and Visual Keywords (AVKs). The first level extracts low-level features such as motion, color, texture, pitch etc to detect video segments boundaries and label segments as audio and visual keywords. Next, we extract the exciting break portions from the AVK sequence. Then, we use two Hidden Markov Models (HMM) to model the exciting break portions with and without goal event respectively. We have applied the proposed approach to the detection of goal event in six half matches of soccer videos (270 minutes, 14 goals) from FIFA 2002 and UEFA 2002 and achieve 90% precision and 100% recall respectively.

1. Introduction

The amount of accessible video information has been increasing rapidly. People quickly get lost in myriad of video data as it is voluminous, and hence it is very time-consuming to locate a relevant video segment linearly. Thus automatic detection of semantic events in video will be very useful. In particular, an increasing number of event detection algorithms are being developed for sports video. In the case of the soccer game that attracts a global viewer-ship, research effort has been focused on extracting high-level structures [1,2] and detecting key highlights to facilitate annotation and browsing [3-6].

Although many approaches have been presented for event detection in sports video, there is still room for improvement from both system modeling and experimental result point of views. From the system modeling point of view, most of the event detection systems known to us share two common features. First the modeling of high-level events such as play-break, corner kicks, goals etc are anchored directly on low-level features such as motion and colors, leaving a large semantic gap between computable features and content meaning as understood by humans. Second some of these systems tend to engineer the analysis process with very specific domain knowledge to achieve more accurate object and/or event recognition. This kind of highly domain-dependent approach makes the development process and resulting system very much ad-hoc and not reusable. From the experimental result point of view, Table 1 shows the precision and recall of goal detection in soccer videos reported by some of the relevant publications presented recently. As we can see, the first two systems do not achieve 100% recall which means that some goal events might be missed. The third system achieves 100% recall, but its precision is not high. The fourth system achieves 100% recall and 100% precision, but it needs tracked temporal position information of the players and ball during a soccer game segment to be input manually.

Table 1 Precision and recall reported by other publications

Reference	Precision	Recall
[3]	77.8%	93.3%
[4]	80.0%	95.0%
[5]	50%	100%
[6]	100%	100%

In this paper, we propose a two-level event detection framework and demonstrate it on soccer videos. Our goal is to make our system adaptable to different events in different domains. To achieve our goal, we introduce a mid-level representation called audio and visual keywords (AVK) that can be learned and detected from video segments. Based on AVK, a computational system that realizes the framework comprises two levels of processing:

1. The first level focuses on video segmentation and AVK classification. The video stream is partitioned into video segments and each segment is labeled with one or more AVKs with certainty values at this level. In the simpler case considered in this paper, for each video segment, we assign two visual keywords and one audio keyword to it. In other words, the first level parses the video stream and outputs a sequence of AVKs.

2. The second level deals with event detection. In general, the probabilistic mapping between the AVK sequence and the events can be modeled either statistically (e.g. HMM) or syntactically (e.g. grammar). In this paper, we first extract the exciting break portions from the AVK sequence. Then, we use two Hidden Markov Models (HMMs) to model the exciting break portions with and without goal event respectively.

The paper is organized as follow. First, we define the set of audio and visual keywords and discuss their classification in next section. In Section 3, we explain how we extract the exciting break portions from AVK sequence and model the goal portions with HMM. Finally, we present promising experimental results in Section 4 followed by a conclusion.

0-7803-8554-3/04/$20.00 ©2004 IEEE.

2. Audio and Visual Keywords

The notion of visual keywords was initially introduced for content-based image retrieval [7]. In the case of images, visual keywords are salient image regions that exhibit semantic meanings and that can be learned from sample images to span a new indexing space of semantic axes such as face, crowd, building, sky, foliage, water etc. In the context of video, Audio and Visual Keywords (AVKs) are extended to represent recurrent and meaningful spatio-temporal patterns of video segments. They are characterized using low-level features such as motion, color, pitch etc and detected using classifiers trained a prior.

2.1 Visual Keywords for Soccer Video

We define a set of simple and atomic semantic labels called visual keywords for soccer videos. From the focus of the camera and the moving status of the camera point of views, we classify the visual keywords into two categories: static visual keywords (Table 2) and dynamic visual keywords (Table 3).

Table 2 Static visual keywords defined for soccer videos

Keywords	Abbreviation
Far view of whole field	FW
Far view of half field	FH
Mid range view (whole body visible)	MW
Close-up view (inside field)	IF
Close-up view (edge field)	EF
Close-up view (outside field)	OF

Generally, "far view" indicates that the game is playing and no special event happens so the camera captures the field from far to show the whole status of the game. "Mid range view" always indicates the potential defense and attack so that the camera captures players and ball to follow the actions closely. "Close-up view" indicates that the game might be paused due to the foul or the events like goal, corner-kick etc so that camera captures the players closely to follow their emotions and actions.

Table 3 Dynamic visual keywords defined for soccer videos

Keywords	Abbreviation
Still camera	ST
Moving camera	MV
Fast moving camera	FM

In essence, dynamic visual keywords based on motion features intend to describe the camera's motion. Generally, if the game is in play, the camera always follows the ball. If the game is in break, the camera tends to capture the people in the game. Hence, if the camera moves very fast, it indicates that either the ball is moving very fast or the players are running. For example: given a "far view" video segment, if the camera is moving, it indicates that the game is playing and the camera is following the ball; if the camera is not moving, it indicates that the ball is static or moving slowly which might indicate the preparation stage before the free-kick or corner-kick in which the camera tries to capture the distribution of the players from far.

2.2 Audio Keywords for Soccer Video

We define three audio keywords for our system: "Plain", "Exciting" and "Very Exciting" for soccer videos. Initially, we also include another audio keyword " Whistle" in our vocabulary. According to soccer games rules, most of the highlights happen along with different kinds of whistling. Ideally, detection of whistling should facilitate the event detection in soccer videos greatly. Unfortunately, the sound of the whistling is always overwhelmed by the noise of the audience and environment. Hence, we remove the "whistle" from our audio keywords vocabulary.

2.3 Classification of AVK

For the first step of our processing, we perform conventional shot classification using color histogram approach to the video stream to segment video stream into video shots. Then, we compulsorily insert shot boundaries within shots whose length is longer than 100 frames to further segment the shot into shorter segments evenly. For instance, a 150-frame shot will be further segmented into 2 video segments, 75-frame each. In the end, we label each video segment with one static visual keyword, one dynamic visual keyword and one audio keyword.

For static visual keyword classification, we first convert all the P-Frames in the video segment into edge-based binary maps by setting all the edge points into white points and other points into black points. We also convert all the P-Frames into color-based binary maps by mapping all the dominant color points into black points and non-dominant color points into white points. Then, we detect the playing field area and segment the Regions of Interest (ROIs) within the playing field area. Finally two support vector machine classifiers and some decision rules are applied to the position of the playing field and the properties of the ROIs such as size, position, texture ratio, etc to label each P-Frame with one static visual keyword. (Fig.1)

We label each P-Frame of the video segment with one static visual keyword. Then, the static visual keyword that is labeled to majority of P-frames is taken as the static visual keyword labeled to the whole video segment. The details of classification of static visual keyword are presented in a separate paper [8], and we shall not repeat its details here. We have applied the method proposed in [8] to 3495 soccer video shots and achieved 94.3% and 97.2% average precision and recall respectively.

By calculating the mean and standard deviation of the number of motion vectors within different direction regions and the average magnitude of all the motion vectors, we label each video segment with one dynamic visual keyword.

For the audio keywords, we first segment audio stream into audio segments of same intervals. Next, we calculate the pitch and the excitement intensity [9] of the audio signal within each audio segment. Then, since the length of the audio segment is much shorter than the average length of the video segments, we use video segment as our basic segment and calculate the average

excitement intensity of the audio segments within each video segment. In the end, we label each video segment with one audio keyword according to the average excitement intensity of the video segment.

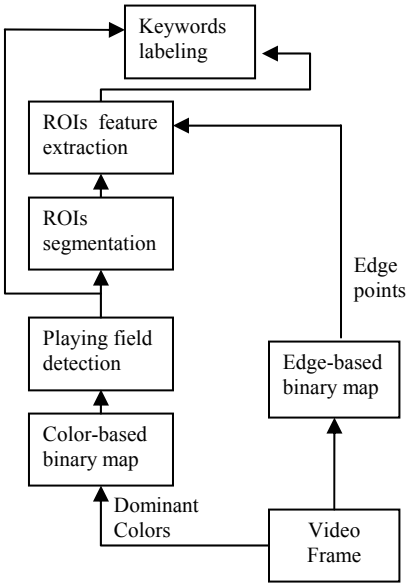

Fig.1 Flow diagram for static visual keywords labeling

3. Event Detection

In this paper, we focus on the use of statistical model for event detection. More precisely, Hidden Markov Models (HMM) are applied to AVK sequences in order to detect the goal event automatically. The AVK sequences that follow the goal events share similar AVK pattern. Generally, after the goal, the game will pause for a while (around 30-60 seconds). During that break period, the camera first zooms into the players to capture their emotions and people cheer for the goal. Next, two to three slow motion replays are presented to show the actions of the goalkeeper and shooter to the audience again. Then, the focus of the camera might go back to the field to show the exciting emotion of the players again for several seconds. In the end, the game resumes.

In this section, we will first describe how we extract exciting break portions from the AVK sequences. Next, we introduce how we map AVK sequence into 13-dimension feature vectors. Then, we describe how we use two HMMs to model the goal pattern. In the end, we report our experimental results.

3.1 Exciting Break Portion Extraction

Generally, long "far view" segment always indicates that the game is in play and short "far view" segment is sometimes used during a break. Hence, we extract play portions by detecting four or more consecutive "far view" video segments. For break portions, we scan the static visual keyword sequence from the beginning to the end sequentially. When we spot a "far view" segment, a portion that starts from the first non-"far view" segment thereafter ends at the start of the next play portion is extracted and regarded as a break portion. (Fig. 2)

After break portions extraction, we use audio keyword to further extract exciting break portions. For each break portion, we compute the number of "EX" and "VE" keywords that are labeled to it, denoted as EX_{num} and VE_{num}. The excitement intensity and excitement intensity ratio of this break portion is computed as:

$$Excitement = 2 \times VE_{num} + EX_{num} \quad (1)$$

$$Ratio = \frac{Excitement}{Length} \quad (2)$$

where Length is the number of the video segments within the break portion.

By setting thresholds for excitement intensity ratio (T_{Ratio}) and excitement intensity ($T_{Excitement}$) respectively, we extract the exciting break portions.

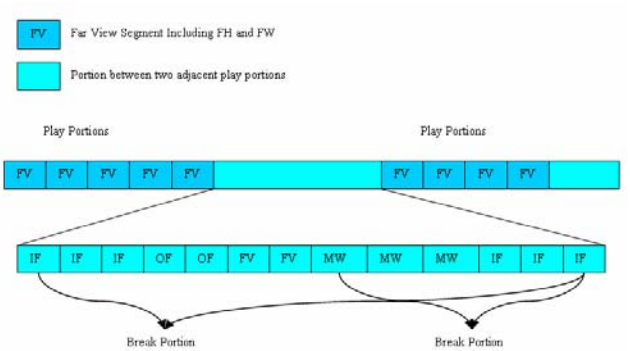

Fig. 2 Break portions extraction

3.2 Feature Vector

For each video segment, we label one static visual keyword, one dynamic visual keyword and one audio keyword. Including the length of the video segment, we use a 13-dimensions feature vector to represent one video segment.

We have defined 12 AVKs in total and the first 12-dimensions correspond to the 12 AVKs. Given a video segment, only the dimensions that correspond to the AVKs labeled to the video segment are set to one and, other dimensions are all set to zero. The last dimension is used to describe the length of the video segment by a number between zero and one, which is the normalized version of the number of the frames of the video segment.

3.3 Goal and Non-Goal HMM

Hidden Markov Model is a powerful tool for analyzing the sequential data. It has been applied to many sports video research work with significant success. In our case, we use two five-state left-right HMMs [10] to model the exciting break portions with

goal event (goal model) and without goal event (non-goal model) respectively. We denote goal model likelihood with G and non-goal model likelihood with N hereafter. Observations sent to HMMs are modeled as single Gaussians.

In practice, HTK [10] is used for HMM modeling. The initial values of the parameters of the HMMs are estimated by repeatedly using Viterbi alignment to segment the training observations and then recomputing the parameters by pooling the vectors in each segment. Then, Baum-Welch algorithm is used to re-estimate the parameters of the HMMs. For each exciting break portion, we evaluate its feature vector likelihood under both two HMMs and we say the goal event is spotted within this exciting break portion if its G is bigger than its N.

4. Experiments and Evaluation

Six half matches of the soccer video (270 minutes, 15 goals) from FIFA 2002 and UEFA 2002 are used in our experiment. The soccer videos are all in MPEG-1 format, 352×288 pixels, 25 frames/second.

AVK sequences of four half matches are labeled automatically. Since these four half matches have 9 goals only, we manually label two more AVK sequences of two half matches with 6 goals. For the purpose of cross validation, for each one of the four automatically labeled AVK sequences, we use the other five AVK sequences as training data to detect goal event from current AVK sequence.

Exciting break portions are extracted from all the six AVK sequences automatically by different sets of threshold settings. In practice, best performance is achieved when the thresholds of T_{Ratio} and $T_{Excitement}$ are set to 0.4 and 9 respectively (Table 4).

Table 4 Result for goal detection (T_{Ratio}=0.4, $T_{Excitement}$=9)

Video	Goal	Correct	Miss	False Alarm	Precision	Recall
GER vs ENG	3	3	0	0	100%	100%
LEV vs LIV	4	4	0	0	100%	100%
LIV vs LEV	1	1	0	0	100%	100%
USA vs GER	1	1	0	1	50%	100%
Total	9	9	0	1	90%	100%

When we set T_{Ratio} and $T_{Excitement}$ to appropriate values, most of the exciting break portions extracted from the AVK sequence are the corner-kick, free-kick, etc. These portions share similar structure patterns. If T_{Ratio} and $T_{Excitement}$ are set to too low, some portions with different structure patterns might also be extracted which will bring noises to the HMM and lower the system performance.

From our experiments, we can see that our approach achieves 90% precision and 100% recall when T_{Ratio} and $T_{Excitement}$ is set to 0.4 and 9 respectively. When T_{Ratio} and $T_{Excitement}$ is relaxed to lower values, the precision and recall degenerate. Since, generally, the excitement intensity ratio and excitement intensity of goal events is higher than 0.6 and 12, we think 0.4 and 9 is a reasonable setting.

5. Conclusion

We have proposed a two-level framework that uses audio and visual keywords to detect goal from soccer video automatically. The notion of audio and visual keywords indeed facilitates the use of HMM-based statistical approach to detect events in soccer video. Our future work includes integration with text caption features in keyword classification; dealing with uncertainties from video segmentation and classification, and applying the framework to event detection in other domains.

6. References

[1] L.X. Xie et al., "Structure Analysis of Soccer Video with Hidden Markov Models", In *Proc. of ICASSP'2002*, 2002.

[2] P. Xu et al., "Algorithms and Systems for Segmentation and Structure Analysis in Soccer Video", In *Proc. of ICME'2001*, 2001.

[3] J.Assfalg, M.Bertini, C. Colombo, A. Del Bimbo, W. Nunziati. "Automatic Extraction and Annotation of Soccer Video Highlights", In *Proc of ICIP* 2003

[4] A. Ekin and A. M. Tekalp, "Generic Event Detection in Sports Video Using Cinematic Features", *2nd IEEE Workshop on Event Mining: Detection and Recognition of Events in Video,* pp.34 June 2003.

[5] L.Y Duan, Min Xu, T.S Chua, Q Tian, C.S Xu, "A Mid-level Representation Framework for Semantic Sports Video Analysis", ACM Multimedia, 2003.

[6] V. Tovinkere, R. J. Qian, "Detecting Semantic Events in Soccer Games: Toward a Complete Solution", In *Proc of ICME* 2001

[7] J.H. Lim, "Building Visual Vocabulary for Image Indexation and Query Formulation. *Pattern Analysis and Applications (Special Issue on Image Indexation)*, ***4****(2/3)*: 125-139, 2001.

[8] Y.L Kang, J.H. Lim, Q. Tian, M.S Kankanhalli, C.S Xu, "Visual Keywords Labeling in Soccer Video" submitted to ICPR, 2004.

[9] K. Wan, N. Maddage, C.S. Xu, "Characterization of dominant speech for automatic sports highlight generation" submitted to ICASSP, 2004.

[10] S. Young, G. Evermann, D. Kershaw, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev and P. Woodland, "The HTK book" version 3.2, CUED, Speech Group, 2002.