DYNAMIC BAYESIAN NETWORK BASED EVENT DETECTION FOR SOCCER HIGHLIGHT EXTRACTION

Fei Wang[1,3], Yu-Fei Ma[2], Hong-Jiang Zhang[2], Jin-Tao Li[1]

[1]Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100080, China
[2]Microsoft Research Asia, 5F Sigma Center, 49 Zhichun Road, Beijing 100080, China
[3]Graduate School of the Chinese Academy of Sciences, Beijing 100039, China

ABSTRACT

In this paper, we propose a novel approach to event detection in soccer videos using Dynamic Bayesian Networks (DBNs). Based on such high level semantics, say, events, more meaningful soccer highlights are extracted. As a powerful statistical tool for time series signal processing, DBNs provide us a feasible method to model sports events by combining contextual information and prior knowledge. In particular, we first develop a DBN model to interpret high-level events composed of low-level primitives in a soccer video. Then, we select a set of robust statistical features as observation input. Finally, the DBN model is leaned to figure out the most likely series of events. The effectiveness of the proposed method has been demonstrated by our experiments.

1. INTRODUCTION

A sports video usually has a long period, but only few important segments may attract viewers, especially for soccer. Therefore, extracting highlights from a sports video is highly desirable. In the literature, automatic highlight extraction techniques can be classified into four categories, i.e. replay based, audio based, model based, and event based. Replay based approaches assume that actions replayed by a broadcaster are typically highlights [1]. Exploring the correlation between highlights and announcers' excited speech, Rui proposed an audio based approach [2]. Ma *et al.* addressed this issue by using user attention model. The highlights are detected at places where strong responses are evoked by the contents that may attract human attentions [3]. However, lacking of exact semantics is the main drawback in these approaches. In order to extract highlights with high level semantics, event based approaches have been developed [4, 5]: Sports events have explicit definition, for example, shoot in soccer, home run in baseball, etc, so to detect these specific events in sports videos can provide viewers more meaningful highlights.

Our previous work includes [6] which used Bayesian Networks (BNs) to classify frames into some typical scenes in a soccer video and [7] which proposed a HMM based framework for sports game event detection. Although BNs take advantage of Bayesian inference to obtain good classification, the contextual information in time line is not sufficiently utilized. HMMs are good at temporal signal analysis, such as speech recognition. However, the expression capability of HMMs becomes limited when they are employed in video content analysis, because video is a kind of signal containing both spatial and temporal information. Compared to the prior work, in this paper we demonstrate the usage of DBNs in sports event detection for highlight extraction. DBNs extend BNs to time series data modeling by considering the state transition between time slices. Meanwhile, DBNs allow a set of random variables instead of only one hidden state node at each time instance, like HMMs [8].

In our system, three key events in a soccer game are defined as highlights, i.e., shoot, corner kick, and free kick, because viewers usually only care about the events leading to the scoring of a goal. In order to build a completed set for soccer events, the other two additional events are also defined, i.e., play and break. We construct a DBM model for this problem using prior knowledge. In such structure, low-level primitives are defined as basic actions, which constitute high-level events in a probabilistic context. The observation input is a set of statistical features extracted from each frame, which are more robust than object-based features, such as player, ball, or goal gate. Finally, the DBN model infers the maximum likelihood series of events from the given observation sequence.

The rest of paper is organized as follows. In Section 2, the architecture of the DBN model for soccer video analysis is introduced. Feature extraction is discussed in Section 3. Section 4 presents the learning and inference of the DBN model in details. The experimental results are reported in Section 5. Finally, Section 6 concludes the paper.

0-7803-8554-3/04/$20.00 ©2004 IEEE.

2. MODELLING SOCCER EVENTS

A DBN is typically described by two sets of parameters (Λ, Θ). The first set Λ represents the structure of the DBN which includes the number of nodes per time slice, and the topology of the network. The second set Θ quantifies conditional probability distributions (CPDs) associated with the edges in the network, and the probabilities of the initial nodes. In this section, we determine the structure of the DBN model by using prior knowledge. Then, the issues about feature extraction, learning and inference are detailed in Section 3 and Section 4, respectively.

As we define the five soccer events according to human understanding, especially for the first three events which are strictly defined by game rules, those events all contain rich high-level semantics. In fact, the direct mapping from low-level features to high-level semantics has been proved ineffective. Therefore, we convert this problem to an inference problem, in which high-level semantic events are decomposed into a serial of low-level primitives. Through statistical inference, the high-level semantics can be obtained from the combination of low-level primitives. Here, the low-level primitives may have directly mapping to low-level features.

According to the prior knowledge, both editing rules and game rules, we define five scene views as common primitives in soccer videos. As shown in Figure 1, the primitive scene views include: (a) close-up view, (b) medium view, (c) midfield view, (d) forward-field view, and (e) goal view. A close-up view usually shows an above-waist view of one player or a view of coach or audience out of the field. A medium view is a zoom-in view normally focusing on one or several players in the field. Midfield views, forward-field views, and goal views are all taken from a wide (or global) view, but serve for different localization of the field.

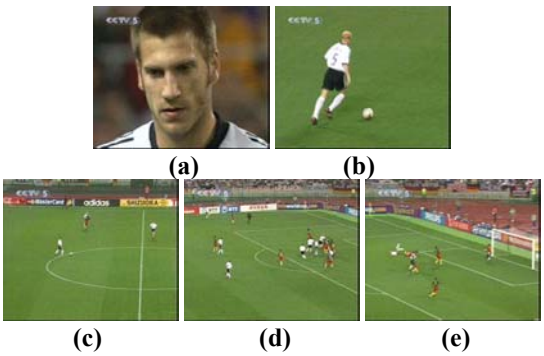


Figure 1. Five types of primitive scene views

As there is a close relationship between events and scene views, a special contextual pattern or transition probability among scene views usually implies an event occurrence. For example, a typical corner kick always has a structural pattern with a goal view near the end line following some close-up views and medium views. Although the exact contents of videos differ from game to game, such production styles and editing patterns are usually followed to help viewers to understand the game.

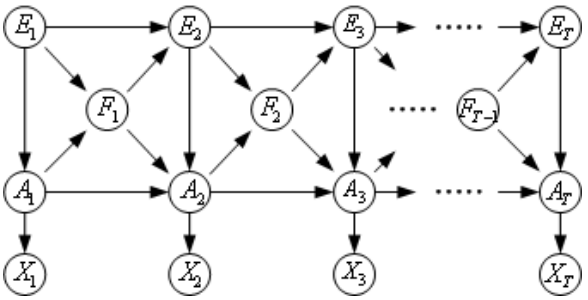


Figure 2. A DBN model for soccer event detection

Based on the above assumptions, we construct a DBN model as shown in Figure 2. There are three levels in the network, including event level, primitive level, and observation level from top to bottom. At each time slice of the sequence, the DBN keeps a set of states which denote all pending derivations. With such representation, low-level features are mapped to high-level semantic events in a way of probabilistic inference. It is worth noted that although our system is tuned for soccer games, the model designed here can be applied to other sports videos with their own specific definitions of primitives and events.

At the top level of this DBN model, the random variable, $E_t = \omega_i$, i = 1…5, specifies the evolution of events in a video sequence. Here, we assume the transitions of events accord with Markov chain.

Each event is composed of a series of common primitives, that is, scene views denoted by node A_t . In addition, note F_t is a binary indicator that can be turned on only if all primitive views in the event E_t are finished. Thus, it is a signal indicating a new event starts.

At the bottom level, features X_t are extracted as the observations from each frame to distinguish the primitive views. Usually, to segment the whole video into shots is the first step in generic video processing. However, it is arguable for sports video due to the following reasons: (1) shots are neither aligned with the views nor consistent with the events; (2) shot detectors tend to give lots of false alarms caused by intense object/camera motion. Rather than the prerequisite of shot boundary detection, we directly analyze features from the video sequence by uniformly sampling.

3. FEATURE EXTRACTION

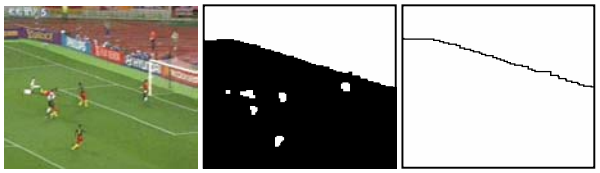
The features we currently use are field color descriptor, player size descriptor, goal area descriptor, and midfield descriptor, in the form of

$$X = [f_{field} \quad f_{player} \quad f_{goal} \quad f_{mid}]^T$$

Field Color Descriptor: A field color is the dominant color in most scenes of a soccer video, which usually indicates the appearance of the playing field, thus important to view classification. As addressed in [6], the field color can be determined with an adaptive field color model. Hence we can define the field color descriptor as:

$$f_{field} = \frac{n_{field}}{n} \tag{1}$$

where n_{field} is the number of field color pixels in the image, and n is the total number of pixels.



(a) (b) (c)

Figure 3. Goal area detection
(a) Original image (b) Binary image (c) Field contour

Player Size Descriptor: For each frame, the image is masked with the field color to achieve a binary image as shown in Figure 3(b). The regions in the field are assumed to be players. The maximum size of the regions is measured as the size of the players in the image:

$$f_{player} = \max(\frac{n_{region}}{n}) \tag{2}$$

Goal Area Descriptor: In our implementation, we convert goal area detection to field contour estimation, which includes side line and end line, as shown in Figure 3(c). If events occur around the goal area, the end line becomes visible while the side line disappears. With such observation, a robust goal area descriptor is defined as:

$$f_{goal} = \frac{n_{end}}{\sqrt{w^2 + h^2}}(1 - \frac{n_{side}}{w}) \tag{3}$$

where w and h are the width and height of frame, while n_{end} and n_{side} denote the number of pixels at the end line and the side line, respectively. To detect the line pixels, we fill the player regions of the filed area in the binary image first. Then the Sobel operator is used to extract edges of the field area. Finally, the lines are detected by the Hough transform with angle and distance constraints.

Midfield Descriptor: This descriptor indicates the presence of the midfield line, defined as:

$$f_{mid} = \frac{n_{mid}}{h} \tag{4}$$

where n_{mid} is the number of pixels at the midfield line. The detection algorithm is similar to field contour detection, except for the operations on the gray image. We also limit the detection in the field region corresponding to the dark region in the binary image (Figure 3(b)). In this way, not only is the Hough transform computation reduced, but also the accuracy is improved by cutting down noises.

4. LEARNING AND INFERENCE

Given the topology of the DBN discussed in Section 2, there are two computational tasks that must be performed to detect the events. The first task is to estimate the parameters of the probabilistic distributions associated with the network. Once the parameters have been learned from training data, the remaining task is inference, i.e. computing the maximum likelihood series of state nodes given the observation sequence of low-level features. A major benefit of DBNs is that they are easy to be interpreted and learned, because the graph is directed and the CPDs of each node can be estimated independently.

Depending on whether the structure is unknown and whether some nodes are hidden, learning methods are different. If the structure is determined and the states are fully observable, such as the DBN we built in this paper, the CPDs may be directly determined by computing statistics from data samples. We define the associated probabilistic functions of the DBN model in Figure 2 as follows. For the feature node X_t , we represent the CPD $p(X_t \mid A_t = \sigma_j)$ using Gaussian Mixture Models (GMMs). In our implementation, we train a GMM for each class of the primitive views by the Expectation Maximization (EM) algorithm. For the primitive nodes, we have

$$P(A_t = \sigma \mid A_{t-1} = \sigma', E_t = \omega, F_{t-1} = 0) = A_\omega(\sigma', \sigma) \tag{5}$$

$$P(A_t = \sigma \mid A_{t-1} = \sigma', E_t = \omega, F_{t-1} = 1) = \pi_\omega(\sigma) \tag{6}$$

where the signal node F_t is turned on by

$$P(F_t = 1 \mid A_t = \sigma, E_t = \omega) = A_\omega(\sigma, end) \tag{7}$$

It is only required to learn the initial and transition probabilities (π, A) among the views in each event separately. At the top level, the associated probabilistic functions of the event nodes are

$$P(E_1 = \omega) = \pi(\omega) \tag{8}$$

$$P(E_t = \omega \mid E_{t-1} = \omega', F_{t-1} = f) = \begin{cases} \delta(\omega', \omega) & f = 0 \\ A(\omega', \omega) & f = 1 \end{cases} \tag{9}$$

where (π, A) is learned among events. During the training period, those probabilities are all estimated from the labeled samples.

As for the inference task, the DBN performs structural rectification of all the candidate sequences $\{E_{1:T}, A_{1:T}\}$ to find a plausible interpretation for the given observation sequence $X_{1:T}$, which has the maximum likelihood. This inference process can be denoted by

$$\{\hat{E}_{1:T}, \hat{A}_{1:T}\} = \arg\max P(X_{1:T}, E_{1:T}, A_{1:T})$$

where $P(X_{1:T}, E_{1:T}, A_{1:T})$ is the joint distribution of the DBN. We use the *Viterbi* algorithm to obtain the optimum state sequence, which is a classical dynamic programming algorithm with the time complexity of $O(T)$.

5. EXPERIMENTAL RESULTS

The experiment composes of two parts. First, we evaluated the performance of the proposed event detection algorithm. Then, we evaluated the soccer highlights generated based on semantic events. About 2 hours of videos are selected from 4 soccer matches as our experimental data, which were segmented into 54 testing clips (from a few minutes to ten minutes). These videos involve different teams, stadiums and cable companies. The ground truth is labeled manually in advance. We use half of the labeled data as training set and the other half as testing data.

Table 1. Event detection results

Event	Hit	False	Miss	Precision	Recall
Corner	27	17	3	61.36%	90.00%
Free	15	7	6	68.18%	71.43%
Shoot	42	17	10	71.19%	80.77%
Play	117	15	11	88.64%	91.41%
Break	55	13	11	80.88%	83.33%
Total	256	69	41	78.77%	86.20%

The event detection results are shown in Table 1. It is promising that the average precision achieves 78.77% with the recall of 86.20%. The highest scores obtained in the events of play and break show that the proposed approach is very robust, because play and break events usually have great variations. The recognition rates of corner kick and free kick are relative low. The reason mainly lies in the similar structure patterns of primitives in the two events. To enhance the performance, finer categories of primitives and more effective features are required.

Table 2. Highlight extraction results

	Hit	False	Miss	Precision	Recall
Frame	50960	10560	6340	82.83%	88.94%
Segment	101	20	2	83.47%	98.06

In the second part of the experiment, we extract three key events as soccer highlight, namely, shoot, corner kick, and free kick. The evaluation is measured both at frame level and segment level by comparing to the ground truth. The results are satisfactory, especially only 2 segments missed. 11.06% (=100%-88.94%) of frames that were missed, however, resulted in only 1.94% (=100%-98.06%) errors in segments. It means that most of missed frames occur at the boundaries of the segments, which are trivial for highlight extraction from the viewer's point. Actually, it is difficult or meaningless for human to give strict ground truth of event boundaries.

6. CONCLUSIONS

In this paper, we propose an effective approach to soccer event detection as well as highlight extraction. The main contributions of this paper are two-folds. First, a DBN based framework of inference is proposed for sports event detection/recognition, which effectively utilizes prior knowledge, including game rules and editing rules. Second, a set of discriminative and robust statistical features are extracted for soccer video analysis. The encouraging experimental results have demonstrated the effectiveness of the proposed DBN based method. In future work, we will focus on multimodal integration in DBNs to further improve the performance.

7. REFERENCES

[1] 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.

[2] Y. Rui, A. Gupta, and A. Acero, "Automatically Extracting Highlights for TV Baseball Programs," *Proc. ACM Multimedia*, 2000.

[3] Y.F. Ma, L. Lu, H.J. Zhang, and M.J. Li, "A User Attention Model for Video Summarization," *Proc. ACM Multimedia*, 2002.

[4] J. Assfalg, M. Bertini, C. Colombo, A.D. Bimbo, and W. Nunziati, "Semantic Annotation of Soccer Videos: Automatic Highlights Identification", *Computer Vision and Image Understanding*, vol. 92, no. 2, 2003.

[5] Y. Gong, *et al.*, "Automatic Parsing of TV Soccer Programs," *Proc. IEEE ICMCS*, 1995.

`

[6] M. Luo, Y.F. Ma, and H.J. Zhang, "Pyramidwise Structuring for Soccer Highlight Extraction," *Proc. IEEE Pacific-Rim Conference on Multimedia*, 2003.

[7] G. Xu, Y.F. Ma, H.J. Zhang, and S.Q. Yang, "A HMM Based Semantic Analysis Framework for Sports Game Event Detection," *Proc. IEEE ICIP*, 2003.

[8] K.P. Murphy, "Dynamic Bayesian Network: Representation, Inference and Learning," *PhD Dissertation*, University of California, Berkeley, 2002.