SEMANTICS-BASED HIGHLIGHT EXTRACTION OF SOCCER PROGRAM USING DBN

Chung-Yuan Chao, Huang-Chia Shih, and Chung-Lin Huang
Institute of Electrical Engineering
National Tsing Hua University
HsinChu, Taiwan,
E-mail: clhuang@ee.nthu.edu.tw

ABSTRACT

This paper proposes a novel semantics-based content analysis system for reliable media highlight extraction using Dynamic Bayesian Network (DBN). It extracts the low-level evidences and then converts the input video to high-level semantic meaning. Specific domains contain rich spatial and temporal transitional structures that help the transformation process. We introduce a robust audio-visual low-level evidence extraction scheme, and develop the so-called *temporal intervening network* to improve the performance of our system. In the experiments, we show that our system can detect the soccer events such as goal event, corner kick event, penalty kick event, and card event effectively.

1. INTRODUCTION

Sports video has been broadcasted to large audiences for their daily life entertainment. We need a flexible and scalable way to manage the sports video, for instance automatic and real time sports video summarization. Obviously, the main gap between low-level media features and high-level concepts needs to be bridged. To solve this problem, several research efforts have been undertaken by using domain knowledge to facilitate extraction of high-level concepts directly from features. The goal of this paper is to develop a system for automatic indexing of sports videos based on speech understanding and video analysis.

Some approaches use stochastic methods with automatic learning capabilities to derive knowledge, such as Hidden Markov Models (HMMs) [1], [2], Bayesian Networks (BNs) [3]. Sun et al. [4] scoring event is detected based on BN form six kinds of features including gate, face, audio, texture, caption and text. Ekin [5] propose a fully automatic and computationally efficient framework for analysis and summarization of sport videos using low-level video processing algorithms. Assfalg *et al.* [6] present a system that performs automatic annotation of the principal highlights of soccer video. A DBN-based system has been proposed using audio-visual feature extraction scheme and text detection and recognition for content-based video retrieval. Takagi et al. [8] proposed a content-based video categorizing method focusing broadcasted sports videos using camera motion parameters. A combination of the audio-band energy and the color dominance pattern recognition has been proposed for event detection of the football video [9].

In this paper, we focus on the highlight extraction of soccer program. Based on DBNs, we can find soccer events such as goal event, corner kick, penalty kick event, and card event, etc. Given a video in specific domain, our system can extract the low-level evidence and interpret the input video in terms of high-level semantic. Our system will extract and present the meaningful and relevant information for the viewers.

2. DYNAMIC BAYESIAN NETWORK

DBN is a model used to describe a dynamically changing system. All the variables, arcs, and probabilities that form static interpretation of a system is similar to BN. Variables here can be denoted as the state of a DBN, because they include a temporal dimension. The states of any system described as a DBN satisfy the Markov condition, it is defined as follows: the state of a system at time t depends only on its immediate past, i.e., its state at time $t-1$. In DBN, we can allow not only intra-slice connections (*i.e.*, the connections within time slices) but also the inter-slice connections (i.e., the ones between time slices). The inter-slice connections incorporate condition probabilities between variables from different time slices. Previous works [1, 2] demonstrated the power of these models in fusing video and audio cues with contextual information and expert knowledge. To completely specify a DBN, we need to define three sets of parameters:

1. State transition *pdfs* $P(x_t \mid x_{t-1})$, that specifies time dependency between the states.
2. Observation *pdfs* $P(y_t|x_t)$, that specifies dependency of observation nodes regarding to the other nodes at time slice t.
3. Initial state distribution $P(x_0)$, that brings the initial probability distribution in the beginning of the process.

Inference is the process of updating probabilities of outcomes based upon the relationships in the model and the evidence about the situation at hand. Since in the DBN, only a subset of states can be observed at each time slice, we have to calculate all the unknown states in the network. The problem of inference in DBN can be represented as the problem of finding $P(x_t \mid y_t)$ with a finite set of T consecutive observations as $\boldsymbol{y}_t=\{y_0, y_2, \ldots y_T\}$ and $\boldsymbol{x}_t$ as the set of the corresponding hidden variables, $\boldsymbol{x}_t=\{x_0, x_2, \ldots x_T\}$. Having constructed a DBN, we need to determine various probabilities of interest from the model by running inference procedure which gives the observations and evidences (*i.e.*, low-level media features).

3. LOW-LEVEL EVIDENCE EXTRACTIONS

The feature extraction process provides low-level evidences based on different media components of the soccer game video. These low-level evidences are essential for the DBNs.

1. **Close-up view.** We classify soccer frames into two classes: bird's-eye view, and close-up view. The former displays the entire soccer field, whereas the latter shows the detail interactions among the players and/or the referee. We assume the existence of the dominant color indicating the soccer field. The dominant field color is described by the peak value of each color component. The computation involves the determination of the peak index, i_{peak}, for color histogram. First, we determine the peak index, i_{peak}, for each histogram. Then, we find an interval $[i_{min}, i_{max}]$ with $i_{min} \le i_{peak} \le i_{max}$, where i_{max} and i_{min} satisfy the conditions: $H[i_{min}] \ge k H[i_{peak}]$ and $H[i_{max}] < k H[i_{peak}]$, with k=0.2, and H refers to the color histogram. We convert the peak of each color component in RGB to HSI. Each pixel is measured by the distance to the peak color (*i.e.*, $d_{cylinder}$) by the cylindrical metric. We assign the pixel to the dominant color region if it satisfies the constraint $d_{\text{cylinder}} < T_{\text{color}}$ where T_{color} is a pre-defined threshold which is video dependent.When the number of dominant color pixels of a frame exceeds a certain threshold value, we say that the frame is bird's-eye view, otherwise it is a close-up view.

2. **Camera Motion.** In the video, the camera motion consists of zooming and panning. We can calculate the camera motion

from the motion of the two-dimensional m×n picture elements using two one-dimensional vectors. First, we calculate the vertical projection for each frame, and the horizontal projection. Second, we find the displacement vector *s* by removing a small distance of the slices from one image and compare with those of the next frame that makes the minimum difference. Having found the displacement value *s* in vertical and horizontal directions, we may interpret the global horizontal motion of the video caused by camera panning motion.

3. **Audience region.** We find that the audience region and grass field region have difference texture, thus we may relate the edge density information to the audience information. we segment each frame into 16×16 blocks. We measure $D(m, n)$ which represents the edge density in each block. If the edge density or texture density of a block is large enough, we say that this block $D(m, n)$ belongs to audience region. When there are many blocks identified as audience regions, we connected these blocks as the audience region.

4. **Replay.** We proposed replay detection schemes that checkup the frames containing logos in the special scene transitions that sandwich the replays. Because the temporal transition between logo and replay is always large, and with special editing effects are applied to the logos, we call the transition as "logo transition". We can use Hue and Intensity differences between two consecutive to find these logo transitions. If the Hue difference is more than 20 and I difference is more than 35 and contain the same structure, we label this frame as starting (or ending) logo transition.

5. **Gate.** We treat the gate finding problem as the searching problem for two or three parallel field lines. In Fig. 1(a), an overhead view of the whole soccer field is shown, and two or three parallel field lines are also shown in bold face in Fig. 1(b). The gate becomes visible when players appear around or within one of the penalty boxes. This information of penalty box yields a robust hint for gate detection. As shown in Fig. 1(c), the parallel lines and the angle ranges from 140 to 170 degree, Fig. 1(d) is another example of parallel and the angle ranges from 10 to 40 degree. To detect goal event, we need to extract the information of parallel lines. When parallel lines tilt to left we may identify it as a right goal, otherwise it is a left goal.

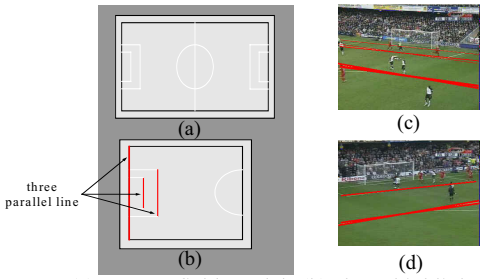


Figure. 1(a) Soccer field model, (b) three highlighted parallel lines around goal area, (c)(d) parallel line detection.

6. **Board.** Bard is a caption region distinguished from the surrounding region, on which some text information about the scoring or team is displayed. We make use of the fact that the caption is often at the image bottom part, arranged horizontally, and appears or disappears abruptly in some frame. Therefore, the abrupt intensity change at the bottom part of the adjacent frames is used to detect the appearance and disappearance of the caption. Our method based on detecting the edge connectivity can be used to locate the caption precisely.

7. **Referee.** In soccer video, the persons of interest (POI) are referee and players. We only search the close-up frames for persons of interest (POI) detection. We describe the shirt colors of referee by black or yellow color, when the ratio of the black color region exceeds a certain threshold, we label the block as the referee region. We call the block as the minimum bounding rectangle (MBR). The decision about the existence of the referee in the current frame is based on the following size-invariant shape descriptors.
   i) If the ratio of the area of the MBR_{ref} to the frame area is outside (0.05, 0.75) interval, then it is considered as outliers.
   ii) Frames with aberrant MBR_{ref} aspect ratio values outside (0.2, 1.8) interval is also considered as outliers

8. **Audio.** The game video is always accompanied with voice and in different stage the voice presents different information in the game. When a scoring event occurs, the excited announcer and audience make a very loud cheering voice, i.e., the voice intensity difference is very large and can be used to help us to determine whether the audio energy is large or small.

4. SEMANTIC ANALYSIS OF SOCCER VIDEO PROGRAMS USING DBN

In sports, the events that unfolded are governed by the rules of sports hence they contain recurring temporal structure. For example, in soccer goal event video, there are recurrent views, such as gate, close-up, and replay etc. DBNs are used to model semantic feature of soccer game such as goal event, corner kick event, penalty kick event, and card event.

4.1 Training

Training can be categorized into two kinds: qualitative (structural training) and quantitative training (parameter training). Qualitative training concerns the network structure of the model and quantitative training determines the specific conditional probabilities.

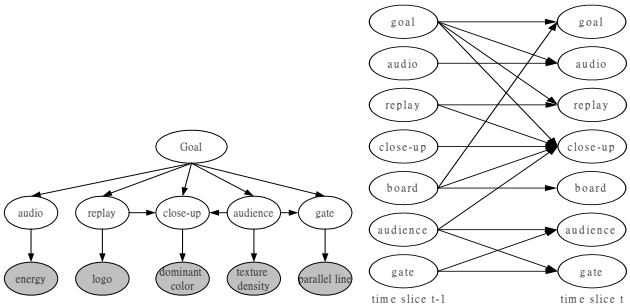
A) **Quantitative Training**. In quantitative training, the dependence between the nodes and the occurrence possibility of each node in the network will be determined. The training procedure can be divided into three phases.
(1) Given root nodes and hidden nodes (*e.g.,* goal and close-up), we count the times of joint appearance is *True* and the times that the appearance of node goal is *True*. Then, we calculate the conditional probability $P(close\text{-}up=Y \mid goal=Y)$.
(2) The dynamic component of DBN which is similar to the 1st training phase. Given two nodes representing the status of the appearance of goal in two consecutive time frames (*e.g., $goal_t$* and $goal_{t-1}$), we count the times of $goal_t$ and $goal_{t-1}$ appear simultaneously, and the times of $goal_t$ exist, and then find the conditional probability $P(goal_t=Y \mid goal_{t-1}=Y)$.
(3) Finding the conditional probability between the given evidence nodes and the hidden nodes. The appearance of hidden node is obtained by human observation, and the appearance of evidence node is obtained by feature extraction process. For instance, we count the total times that the gate and the parallel lines appear simultaneously, and the number of times that the gate exist, and have $P(parallel\ line=Y \mid gate=Y)$.

B) Qualitative Training. After the quantitative training, any causal relation between any two nodes is represented by the directional linkage between them, which leads from the cause (parent) node (*i.e.*, n_c) to the effect (child) node (*i.e.*, n_e). Each effect node may be connected to *J* cause nodes. We let $p(n_{e_i} \mid n_{c_j})$ represent the conditional probability relating the cause node n_{c_j} to

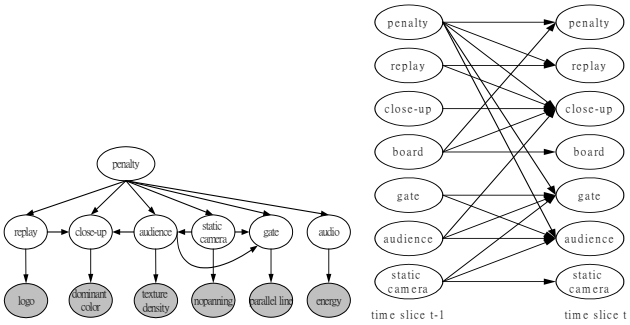
the effect nodes n_{e_j} where $j=1,...J$ and $i=1,...I$. To determine the effective linkages for the network, we let $U = \{(n_{e_i}, n_{c_j})\}$ be the universe of the configuration over a universe of the linkages of every two nodes and $\{P(x)\}=\{p(n_{e_i} \mid n_{c_j})\}$ be the original distribution after training. M is the candidate network with $\{P^*(x)\}=\{p^*(n_{e_i} \mid n_{c_j})\}$ as the distribution after *thresholding*. We define (1) the *Size* of M, Size(M), which is the number of entries in P^* that $p^*(n_{e_i} \mid n_{c_j}) > t$, (2) the *Cross Entropy Distance*, Dist(P, P^*)= $P(x) \Sigma_j \log P^*(x)/P(x)$, and (3) the *Acceptance Measure*, $Acc(P, M)=Size(M)+k$ Dist(P, P^*). Then, we use the Langrage method to choose k and the *threshold* t that minimize the $Acc(P, M)$. Finally, we use the Bayes' rule to obtain the *posteriori* probability $p(n_{c_j} \mid n_{e_i})$. In the training phase, 500,000 frames are used to generate a reliable DBN.

4.2 Dynamic Bayesian Network Model

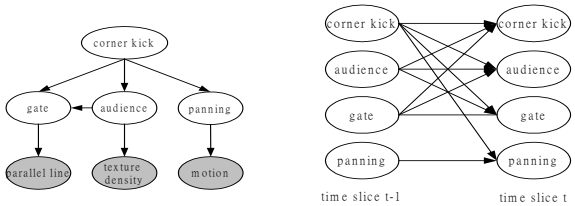
The major objective of our approach is using the DBN to link low-level media features to the high-level concepts. Different networks of event detection is illustrated in Figure 2 of which the one on left-hand side denotes the BN and the one on right-hand side represents the temporal dependency of DBN for (a) goal event; (b) penalty kick event; (c) corner kick event, and (d) card event respectively.



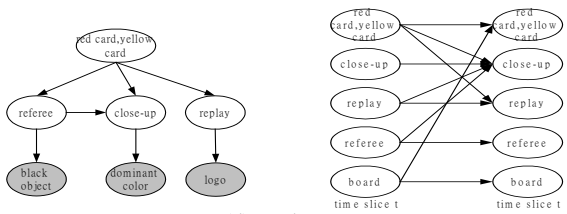
(a) goal event



(b) penalty kick event



(c) corner kick event



(d) card event

Figure 2. Different structure of event DBN network.

4.3 Temporal intervening network

The events in soccer video have certain regularity which can used to increase certain *posteriori* probabilities. For instance, in the goal event, the occurrences of gate, close-up, and replay follow certain rules of causality. In the beginning of a goal event video (*e.g.,* Figure 3), we always find the appearance of gate. When the gate disappears, the 1st close-up will appear in less than 20 frames time interval. After the 1st close-up, the replay segment appears, and there are other close-ups and gate within the replay video segment. Finally, the last close-up and the score board appear, and then the goal event terminates.

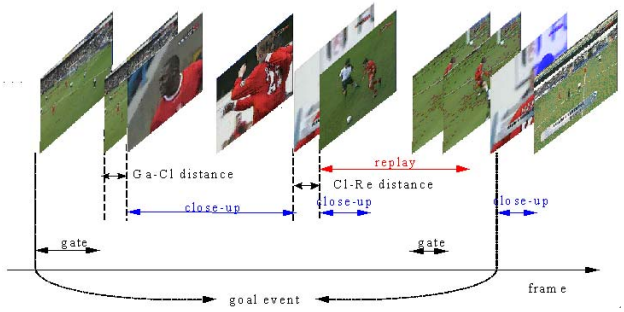


Figure 3. The temporal regularity in a goal event.

We find that in the goal event, before close-up, the gate appears. Once the gate disappears, the close-up will appear immediately or in a time interval of 20 frames. This regularity can be used to improve the accuracy of event detection. In goal event, we introduce two *temporal intervening networks* as shown in Figures 4(a) and (b). The temporal intervening network changes the *posteriori probability* distribution of the DBN during the model inference when Close-up=*True*, it is called the *intervening action* [3]. Similarly, the other temporal intervening network (i.e., Fig. 4(b)) can also be applied to our DBN when Replay=*True*. In the goal event, we always find replay after close-up. When the close-up disappears, the replay will appear immediately or in a time interval of 20 frames.

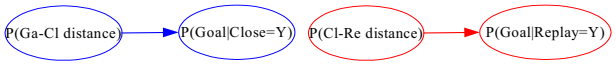


Figure 4. Temporal intervening network for goal event detection.

The close-ups appear throughout the entire video, however, only a part of close-ups will be found in the goal event. Close-ups also appear during the card event or other undefined events. When the close-ups emerge during the goal event, in front of these close-ups, we always find the appearance of gates. Similarly, when the replays are found the goal event, in front of these replays, we may find the existence of close-up.

5. EXPERIMENTAL RESULTS

In the experiments, we have tested more than 10 hours (630 min.) video sequences from five soccer TV programs of England Premier League, two soccer TV programs of UEFA Cup. The data-base is MPEG-1 clips in 320x240 resolution at 30 frames per second. Audio is sampled at 88kHz with 16 bits per sample.

5.1 Frame-based event detection

Table 1 shows the experimental results of the seven complete soccer games with the temporal intervening networks. On the overall accuracy is about 90.5% and false alarm rate about 5.4% for four events.

Table 1. Results of event detection for seven test sequences with DBN

Sequence	Results	goal	corner	penalty	card
England 1 (95:04)	Accuracy	92.91%	84.29%	93.48%	92.75%
	False alarm	11.23%	0.55%	0.94%	7.81%
England 2 (98:05)	Accuracy	91.52%	89.97%	N/H	89.99%
	False alarm	12.41%	0.23%	1.26%	9.61%
England 3 (94:10)	Accuracy	93.58%	83.60%	92.17%	90.69%
	False alarm	8.41%	0.19%	1.28%	8.53%
England 4 (94:28)	Accuracy	88.16%	85.55%	N/H	96.31%
	False alarm	9.66%	0.23%	0.65%	7.74%
England 5 (94:41)	Accuracy	92.09%	85.57%	94.47%	94.59%
	False alarm	13.52%	0.62%	1.66%	7.63%
UEFA 1 (95:14)	Accuracy	88.98%	85.22%	N/H	91.80%
	False alarm	12.07%	0.74%	0.32%	8.86%
UEFA 2 (94:07)	Accuracy	87.61%	84.69%	N/H	93.50%
	False alarm	13.31%	0.71%	0.86%	9.62%
Average	Accuracy	90.69%	85.24%	93.37%	92.80%
	False alarm	11.52%	0.59%	0.99%	8.54%

N/H: denoted non-happen

5.2 Frame-based event detection

Table 2 shows other experimental results of goal event detection. The reason why the number of false alarm rate in Table 2 is more than miss rate is that the offside is miss-identified as the goal event. When the goal event is detected, we may differentiate it as a left or right goal. If the frames of goal event last more than continuous 500 frames, we say that there is one complete goal event. This left/right goal event detection provides an useful information: which team has dominated the game. Similarly in England 5, the number of detected left goal of team 1 is more then the right goal of team2. On the other hand, in England 3, the two teams are well-matched.

Table 2. Using DBN and temporal intervening network for goal event.

Sequence	England 1		England 2		England 3		England 4		England 5		UEFA 1		UEFA 2	
Field	Left	Right	Left	Right	Left	Right	Left	Right	Left	Right	Left	Right	Left	Right
Detected	21	10	23	17	16	20	15	19	19	9	13	15	18	21
False alarm	3	2	5	3	3	6	2	3	5	3	2	3	6	5
Missed	0	1	0	0	1	0	0	0	0	0	0	0	0	0
Precision	87.5	83.3	82.1	85	84.2	76.9	88.2	86.4	79.2	75	86.7	83.3	75	80.8
Recall	100	90.9	100	100	94.1	100	100	100	100	100	100	100	100	100

Table 3 shows other experimental results of corner event detection. When the corner event occurs, we differentiate the left/right corner using the gate information. If the frames of corner event last more than 20 continuous frames, we say that there is one complete corner event. Comparing the precision rate in Tables 2 and 3, we find that the precision rate in Table 3 is worse. It is because the duration of the corner event is shorter than the goal event (*i.e.,* finishes within 20 frames) so that it induces higher chance of False alarm.

Table 3. Using DBN and temporal intervening network for corner event

Sequence	England 1		England 2		England 3		England 4		England 5		UEFA 1		UEFA 2	
Field	Left	Right	Left	Right	Left	Right	Left	Right	Left	Right	Left	Right	Left	Right
Detected	13	10	9	9	6	14	10	11	13	5	10	11	12	12
False alarm	5	4	4	3	2	5	5	3	7	2	3	4	7	5
Missed	1	0	2	0	1	2	0	1	1	0	0	1	1	0
Precision	72.2	71.4	69.2	75	75	73.7	66.6	78.6	65	71.4	76.9	73.3	63.2	70.6
Recall	92.9	100	81.8	100	85.7	87.5	100	91.7	92.9	100	100	91.7	92.3	100

6. CONCLUSIONS

We have proposed a soccer video understanding system based on DBN. Given an input sequence, the system will collect the low-level evidence, and the inference engine in DBN can be applied to infer a high-level semantic concepts. The main contribution of this paper is to develop a DBN and the temporal intervening network to model the relationship between unobservable concepts and observable concepts. The experimental results demonstrate that our system can detect the semantic events of the sports video effectively.

REFERENCE

[1] L. R. Rabiner, "A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition," *Proc.* IEEE Vol. 77, No. 2, Feb. 1989.

[2] L. Xie, S.-F. Chang, A. Divakaran, H. Sun, "Structure analysis of soccer video with Hidden Markov Models," Proc. IEEE ICASSP, 2002.

[3] F. V. Jensen, *An Introduction to Bayesian Networks*, Springer, 1996.

[4] X. Sun, G. Jin, M. Huang, G. Xu, "Bayesian network based soccer video event detection and retrieval," Multispectral IP and PR, Beijing, China, October 2003.

[5] A. Ekin, A. M. Tekalp, and R. Mehrotra, "Automatic Soccer Video Analysis and Summarization," IEEE Trans. on Image Processing, Vol. 12, No. 7, July 2003.

[6] J. Assfalg, and M. Bertini, "Semantic annotation of soccer videos: automatic highlights indentification," *Computer Vision and Image Understanding* 91(3) 2003.

[7] V. Mihajlovic and M. Petkovic, "Automatic Annotation of Formula 1 Races for Content-Based Video Retrieval," Technical Report, TR-CTIT-01-41, 2001.

[8] S. Takagi and S. Hattori, "Sports video categorizing method using camera motion parameters," *Proc.* IEEE-ICME, Baltimore, USA, July 2003.

[9] D. A. Sadlier, N. O'Connor, S. Marlow, and N. Murphy, "A combined audio-visual contribution to event detection in field sports broadcast video. Case study: Gaelic Football," IEEE International Sym. on Signal Processing and Information Technology, Darmstadt, Germany, Dec. 2003.