Semantic Analysis of Soccer Video Using Dynamic Bayesian Network

Chung-Lin Huang, *Senior Member, IEEE*, Huang-Chia Shih, *Student Member, IEEE*, and Chung-Yuan Chao

***Abstract*—Video semantic analysis is formulated based on the low-level image features and the high-level knowledge which is encoded in abstract, nongeometric representations. This paper introduces a semantic analysis system based on Bayesian network (BN) and dynamic Bayesian network (DBN). It is validated in the particular domain of soccer game videos. Based on BN/DBN, it can identify the special events in soccer games such as *goal event*, *corner kick event*, *penalty kick event*, and *card event*. The video analyzer extracts the low-level evidences, whereas the semantic analyzer uses BN/DBN to interpret the high-level semantics. Different from previous shot-based semantic analysis approaches, the proposed semantic analysis is frame-based for each input frame, it provides the current semantics of the event nodes as well as the hidden nodes. Another contribution is that the BN and DBN are automatically generated by the training process instead of determined by ad hoc. The last contribution is that we introduce a so-called *temporal intervening network* to improve the accuracy of the semantics output.**

***Index Terms*—Dynamic Bayesian network (DBN), temporal intervening network (TIN), video semantic analysis.**

I. Introduction

In THE past decade, a large amount of digital media data including image, audio, video, streaming video clips, panorama images, and three-dimensional (3-D) graphics have been delivered to audience. We need a flexible and scalable way to manage these rich media of which the digital video has been widely accepted as the most accessible one. The MPEG-7 has tried to standardize the content-based media access methods. For example, the video indexing and retrieval are useful query tools for us to access the media, which consists of automatic classification, summarization and understanding of video shot.

Several research efforts have been undertaken by using domain knowledge to facilitate extraction of high-level concepts directly from features. Some approaches use stochastic methods that often exploit automatic learning capabilities to derive knowledge, such as hidden Markov models (HMMs) [1]–[3]. Ekin [4] proposes a fully automatic and computationally efficient framework for sports video analysis and summarization by using low-level video processing algorithms. Recently, automatic detection of the principal highlights of sports video has become popular. Snoek *et al.* [5] utilize the time interval maximum entropy (TIME) to classify the event in multimodal video. In [6], [7], Gong *et al.* use three aspects of feature design in soccer video indexing. They propose a maximum-entropy method to choose the features with more distinguished power which is applied to detect and classify baseball highlights for soccer indexing.

Recently, the Bayesian network (BN) [8] has been applied for semantic analysis. In [9], Sun *et al.* uses BN for scoring event detection in soccer video based on using six different low-level features including gate, face, audio, texture, caption, and text. Shih *et al.* [10] develop the so-called multilevel semantic network (MSN) to interpret the highlights in baseball game video. Another highlight detection method [11] exploits visual cues estimated from the video stream, the currently framed playfield zone, player's position, and the colors of players' uniforms.

The low-level features are used for semantic analysis to identify the highlight [14], i.e., object, color and texture features are employed to represent the highlight. Xu *et al.* [12] propose an effective algorithm for soccer video, which detects the plays and breaks in soccer games by motion and color features. Wan *et al.* [13] detect and track important activities such as ball possession in soccer video that is highly correlated to the camera's field-view.

Rule-based video analysis and indexing systems using the mixture of cinematic and object descriptors are proposed in [15] and [16]. A content-based video categorizing method focusing on broadcasted sports videos using camera motion parameters has been developed in [17]. A combination of the speech-band energy tracking in audio domain and the color dominance pattern recognition in video domain provides a useful contribution to event detection for football video [18]. A knowledge-based semantic inference scheme for events recognition in sports video has been presented by three-layer semantic inference scheme [19].

The dynamic Bayesian network (DBN) [20] is based on the BNs and their extensions, it tries to unify temporal dimension with uncertainty. DBN is a useful tool for representing complex stochastic processes. Recent developments in inference and learning in DBN [20]–[24] have been applied to many real-world applications. In [20], they propose a robust audiovisual feature extraction scheme and text detection and recognition method. Their system provides automatic indexing of sports videos based on speech and video analysis. They focus on the use of DBN and demonstrate how they can be effectively applied for fusing the evidence obtained from different media information sources.

Here, we introduce an innovative, high-level, semantics-based content description analysis for reliable media access and navigation service based on the DBN. Given a video

Manuscript received January 4, 2005; revised August 15, 2005. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Jie Yang.

C.-L. Huang and H.-C. Shih are with the Electrical Engineering Department, National Tsing-Hua University, Hsinchu 30043, Taiwan, R.O.C. (e-mail: clhuang@ee.nthu.edu.tw; hc.shih@ieee.org).

C.-Y. Chao is with Magic Pixel, Inc. Hsinchu, Taiwan, R.O.C.

Digital Object Identifier 10.1109/TMM.2006.876289

in a specific domain, our system may extract the low-level evidences and then generate the high-level semantic meaning. Specific domains contain rich spatial and temporal transitional structures for the semantic interpretation process. In sports, the events that unfolded are governed by the rules of the game, so that they contain a recurring temporal structure. The rules of production of sports video have also been standardized. In soccer videos, there are only a few recurrent views, such as close-up and global-view, whereas in baseball videos, there are pitching, close-up, home plate, battering, and crowd etc.

Sports programs are usually lengthy. The content provider needs to extract and present the highlights for the viewers. The video information of various sports programs can be significantly different, such as the rules of the game, the player action, the camera motion, and strategy of the game. This paper presents the automatic interpretation of the highlights in the soccer game video. The BN/DBN is constructed by linking subnets to a root node of which the status indicates the certainty of the specific event.

Different from previous semantic analysis approaches, the proposed semantic analysis is frame-based instead of shot-based. For each input frame, it provides the current semantics of the event nodes as well as the hidden nodes. The second major contribution is that the BN and DBN are automatically generated in the training process rather than determined by *ad hoc*. The last contribution is that we introduce a so-called *temporal intervening network* to improve the accuracy of the semantic analysis. Our method may identify the highlight events in soccer video including *goal event*, *corner kick event*, *penalty kick event*, and *card event*.

II. Dynamic Bayesian Network

The BN [8] encodes the conditional dependence relationships among a set of random variables in the form of a graph. A linkage between two nodes denotes a conditional dependence relation, which is parameterized by a conditional probability model. The structure of the graph encodes the domain knowledge, such as the relationship between the observation nodes and the hidden states, while the parameters of the conditional probability models can be learned from training data.

However, BN does not provide the direct mechanism for representing temporal dependencies. We need to add temporal dimension into the BN model as "temporal" or "dynamic." DBN is used to model a temporally changing system. This model will enable users to monitor and update the system as time proceeds, and even predict further behavior of the system.

DBN is usually defined as the special case of singly connected BN specifically aimed at time series modeling. All the variables, arcs, and probabilities that form static interpretation of a system is similar to BN. Variables can be denoted as the states of a DBN, because they include a temporal dimension. The states satisfy the Markov condition, it is defined as follows: the state of a DBN at time t depends only on its immediate past, i.e., its state at time $t-1$. In DBN, we allow not only intra-slice connections (i.e., within time slices) but also the interslice connections (i.e., between time slices). The inter-slice connections incorporate condition probabilities between variables from different time slices. Each state in a dynamic model at one time instance may depend on one or more states at the previous time instance or/and on some states in the same time instance. So, the state at time t may depend on the system states at time $t-1$ and possibly on current states of some other variables of DBN structure at time t.

To completely specify a DBN, we need to define three sets of parameters: 1) State transition probability $\mathrm{P}(x_t|x_{t-1})$, that specifies time dependency between the states. 2) Observation probability $\mathrm{P}(y_t|x_t)$, that specifies dependency of observation nodes regarding to the other nodes at time slice t. 3) Initial probability $\mathrm{P}(x_0)$, that brings the *priori* probability distribution in the beginning of the process.

From the input video, the video analyzers may find the possible existence of certain low-level evidences. To generate a DBN for sports video, we develop the following steps. 1) Formulate problem in terms of creating a set of variables representing the distinct elements of the situation being modeled. 2) Assign the set of mutually exclusive states or outcomes of each variable. 3) Generate the priori probabilities for each variable and their conditional probabilities based on the training data. 4) Determine the causal dependency relationships between these two variables. This involves creating direct edges linking from the parent (influencing) nodes to the child (influenced) nodes, and from the previous time slice nodes to the current time slice nodes.

III. Low-Level Evidence Extraction

Most of the semantic analysis methods rely on the low-level evidence in the scene. Here, we briefly describe the methods to identify the probability of the existence of the low-level evidence including *dominant color region*, *short-term motion*, *texture intensity*, *logo*, *parallel lines*, *score board*, *black object*, *audio energy*, and *long-term static scene*. They are essential for the inference process in DBNs to generate high level semantic interpretation.

A. Dominant Color Region

In soccer video, there are two different scenes in bird's-eye view or close-up view. A bird's-eye view captures the entire soccer field, whereas a close-up view shows the detail interactions among the players and/or the referee. We use the similar method in [4] to find one dominant color (i.e., green) in the soccer video, however, it may vary from stadium to stadium, different weather, and lighting conditions. Normally, the dominant color region indicates the soccer field. The dominant color is described by the peak value of each color component.

The color image is in RGB space with the color histogram of each component defined as $\mathrm{H}[i]$ (i.e., $\mathrm{H}[i]$ indicates the color histogram of R, G, or B component). For each component, we determine the peak index, i_{peak}, for $\mathrm{H}[i]$, and then find an interval $[i_{min}, i_{max}]$ with $i_{min} \leq i_{peak} \leq i_{max}$, where i_{max} and i_{min} satisfy the conditions: $\mathrm{H}[i_{min}] \geq k\mathrm{H}[i_{peak}]$, $\mathrm{H}[i_{min}-1] < k\mathrm{H}[i_{peak}]$, $\mathrm{H}[i_{max}] \geq k\mathrm{H}[i_{peak}]$, $\mathrm{H}[i_{max}+1] < k\mathrm{H}[i_{peak}]$, with $0 \leq k \leq 1$. The conditions define the minimum (maximum) index as the smallest (largest) index to the left (right), including the peak that has a predefined number of pixels, i.e., $k = 0.2$. Then, we convert the peak of each color component in RGB to HSI.

Fig. 1. Dominant color region detection in bird's-eye view, and close-up view. (Color version available online at http://ieeexplore.ieee.org.)

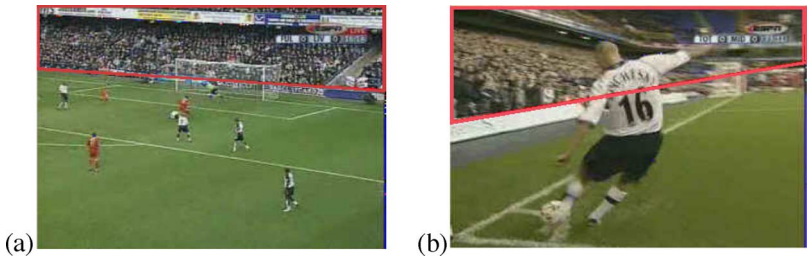


Fig. 2. Audience region detection. (Color version available online at http://ieeexplore.ieee.org.)

For color pixels in each frame, we calculate the distance from each pixel j to the peak color (i.e., $d_{cylindrical}$) by the cylindrical metric. We assign pixel j to the dominant color region if it satisfies the constraint $\mathrm{d_{cylindrical}} < \mathrm{T_{color}}$ where $\mathrm{T_{color}}$ is a predefined threshold which is video dependent. The existence of the dominant color region is based on the ratio of the area of the pixels in dominant color and the area of the entire frame. The higher area ratio indicates the higher probability of the existence of the dominant color region. Fig. 1 illustrates the results of dominant color region detection for two views. We use the area ratio to define the probability of the existence of a dominant color region.

B. Short Term Motion

We can calculate the camera motion between two consecutive frames (or *short-term motion*) by using two one-dimensional (1-D) projections of two consecutive frames with $m \times n$ picture elements (pixels). The pixel motion can be obtained by analyzing the characteristics of the vertical and horizontal projections respectively. First, we calculate the vertical projection $f_x(i)$ for each frame. Second, we calculate the *sum of absolute difference* (SAD). We divide the 1-D projection into small slices with N pixels width $(\mathrm{N} = 16)$. From two consecutive frames a and b, we take a slice of frame a, slide it overlap frame b, and calculate the SAD value of the two slices as

$$SAD(n_0, s_x) = \sum_{i=n_0-N/2}^{n_0+N/2} |f_{x,a}(i) - f_{x,b}(i+s_x)|$$

where n_0 is the index of the center position of the slice from frame a and s is the displacement value. Third, for each slice, we find the horizontal displacement vector s_x that generates the minimum SAD values.

For each frame, we may have a set of displacement vectors $\{s_x\}$. Similarly, we may have the horizontal projection, $f_y(j)$, and use the similar method to find a set of vertical displacements $\{s_y\}$. The magnitude of the average of the displacement vectors $\{s_x\}$ and $\{s_y\}$ indicate the priori probability of short-term motion.

C. Texture Intensity

The two major background regions in the soccer scene are the audience region and grass field region. With difference texture features, they can be differentiated by the texture density information. To analyze texture (edge) intensity, we segment each frame into $h \times w$ blocks, and let $D_{(m,n)}$ represent the edge density in each block (located at (m,n)) which is defined as

$$D_{(m,n)} = \frac{1}{h*w} \sum_{i=0}^{h-1} \sum_{j=0}^{w-1} E(m+i, n+j)$$

where $h \times w$ is the block size, and $E(i,j) = 1$ if pixel (i,j) is an edge pixel, otherwise $E(i,j) = 0$. If the edge density or texture density of a block is large enough, we say that this block belongs to the audience region. When there are many blocks of audience region, we merge these blocks as the audience region. As shown in Fig. 2(a) and 2(b), the bounded region is detected as the audience region.

D. Logo

In broadcast sports video, replays provide the viewers another chance to watch the interesting events. The replays can be utilized for efficient navigation, indexing, and summarization of the sports video programs. The replay segment finding method identifies the replay via detecting the logos that sandwich the replay [25] (see Fig. 3).

To check the color and luminance differences between two consecutive frames, we apply the *histogram-based* scene cut detection algorithm [26]. We apply the distance measure for the difference between two consecutive frames X and Y as

$$d(X,Y) = \sum_{j=1}^{K} \begin{cases} \frac{(H_X(j)-H_Y(j))^2}{\max(H_X(j),H_Y(j))}, & \text{if } (H_X(j) \neq 0) \cup (H_Y(j) \neq 0) \\ 0, & \text{otherwise} \end{cases}$$

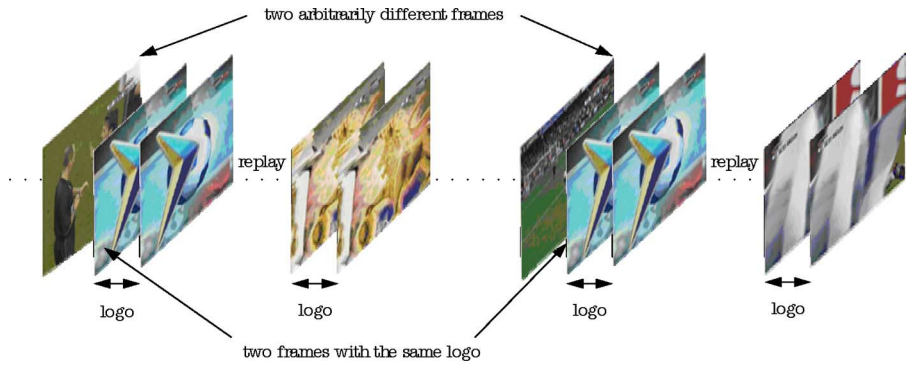


Fig. 3. The corresponding frames in different logo transitions contain the same logo. (Color version available online at http://ieeexplore.ieee.org.)

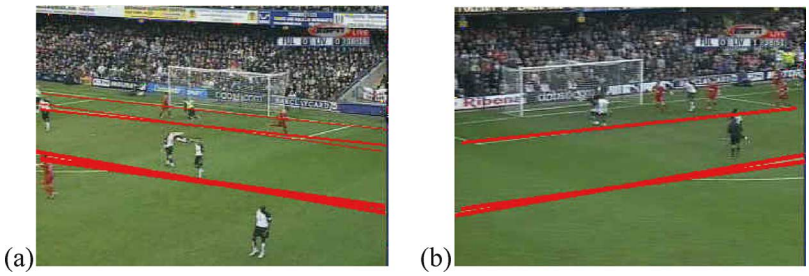


Fig. 4. Gate region detection. (Color version available online at http://ieeexplore.ieee.org.)

where $H_X(\mathrm{j})$ and $H_Y(\mathrm{j})$ are the bin value of the color or luminance histogram of frames X and Y, and K is the overall number of bins. A logo transition will be detected if $d(X,Y) > \theta_{large}$ where frames X and Y are in different sessions. The logo usually appears in less than 1 s, and the logo image sequence is a set of a freeze logo pattern which can be verified as $d(X,Y) < \theta_{small}$ where frames X and Y are in the same logo session and $\theta_{small} \ll \theta_{large}$.

The logo detection algorithm is based on the scene cut detection which may find the scene change between the logo session and the replay with slow motion or in regular speed. Its performance is also insensitive to slow-motion video segments that are captured with high-speed camera. The experimental results show that the accuracy of our algorithm is more than 98% (i.e., accuracy $=$ the number of correct detections/ total number of logos)

E. Parallel Lines

The appearance of two or three parallel field lines in a bird's-eye view can be used to indicate the occurrence of the gate. The appearance of gate and parallel field lines are highly co-related. The gate is visible when the players appear close to or within one of the penalty boxes. This information of parallel lines which indicates the penalty box is very useful for gate detection. The information of the parallel is more reliable than the information of the gate post from the video scene, since the gate post detection may fail due to the cluttered background pixels.

Here, we use edge detector and Hough Transform to detect the parallel lines. As shown in Fig. 4(a), the parallel lines are detected, and their slope angles range from 140° to 170°, Fig. 4(b) illustrates another example of parallel lines, and their angles range from 10° to 40°. When parallel lines tilt to left, it implies a right goal, otherwise a left goal.

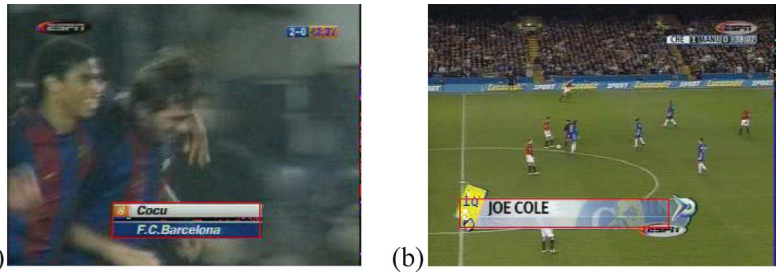
F. Score Board

The score board is a caption region distinguished from the surrounding region, which provides the information about the score of the game or the status of the players. Here, we combine the dynamic and static properties to detect the caption region. We make use of the fact that the caption often appears at the bottom part of image frame for a short while and then disappears. So 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 detects the four edge segments (which enclose a rectangle box) to locate the caption precisely. In Fig. 5, the rectangular box with red border is detected as the caption region. The position and size of the rectangle indicates different possibility of the existence of the board.

G. Black Object

In soccer video, the persons of interest for semantics interpretation are the referee and the players. The information of referee is useful for the event detection, e.g., yellow/red card events. The events may also involve close-up frames of the participating players. The referee identification is robust because the variation of their shirt colors is limited. We assume that the referee is dressed in black, we may use the color segmentation to find the referee. After color segmentation, we may use a compact rectangle to enclose the black region which is called the minimum bounding rectangle (MBR).

The existence of *black object* depends on two size-invariant properties of MBR: 1) the ratio of the area of the MBR to the frame area, and 2) MBR aspect ratio (width/height). Different area ratio and aspect ration indicates different prior probability



(a) (b)

Fig. 5. Different kinds of board region detection. (Color version available online at http://ieeexplore.ieee.org.)

Fig. 6. Referee detection. (Color version available online at http://ieeexplore.ieee.org.)

of MBR. The MBR with ratio of the area inside (0.05, 0.75) interval and the aspect ratio values inside (0.2, 1.8) interval has a higher prior probability of *black object* (as shown in Fig. 6).

H. Audio Energy

The soccer video program is always accompanied with voice, which conveys crucial information of the game. When a goal event occurs, the excited announcer and the audience will make a very loud cheering voice. For other incidents, the announcer will also raise his voice indicating certain ongoing highlights. The higher energy of the voice indicates the higher occurrence of an event. Mostly the high voice intensity occurs during the goal event.

I. Long-Term Static Scene

In the soccer game video, the camera keeps tracing the ball, so that the continuous camera panning motion stops only when a particular event occurs, such as a penalty kick. To detect long-term static scene, we find 1) very small global motion displacement $s \approx 0$ and 2) no logo session (a very small duration of static scene). Normally, the static video sequence lasts for more than 10 s time interval. The existence of long-term static scene depends on the duration of the static scene. The longer duration indicates the higher probability of the existence of the long-term static scene.

IV. Semantic Analysis Using BN and DBN

BN and DBN are powerful semantic analysis tools which have been applied to model the high-level semantic information embedded in the video data. In sports, the high-level semantics are the highlight events containing recurring temporal structure. Here, we use BN/DBN to model the semantic highlights of soccer game such as goal event, corner kick event, penalty kick event, and card event. The BN/DBN is automatically generated after the following training process rather than determined by *ad hoc*.

A. Training Phase

Based on the extractable features and their causality in the soccer video, we define three types of nodes: 1) the *event nodes*, such as goal, corner, penalty, and card; 2) the *hidden nodes*, such as replay, board, close-up, audio, audience, gate, panning, static camera, and referee; 3) the *evidence nodes*: such as energy, logo, texture, motion, parallel lines, and dominant color. Initially, every node in the network is not connected. In the training phase, the human observers count the number of the appearance of each node or the joint appearance of two nodes in the training video sequences. Training can be categorized into two kinds: qualitative (structural training) and quantitative training (parameter training). The former generates the DBN network structure of the model, whereas the latter determines the specific conditional probabilities.

1) Quantitative Training: In quantitative training, the dependence between the nodes and the occurrence possibility of each node in the network will be determined. Nodes are the graphical representation of the evidence of the events in the video which are usually termed as variables or states. The training procedure can be divided into three phases. In the first training phase, we compute all the conditional probabilities between event-hidden nodes or the hidden-hidden nodes by counting the number of times that the joint appearance of the event-hidden node pair (e.g., *goal* and *close-up*) is true and the number of times that the appearance of the event node is true. We can also count the number of times that hidden-hidden node pair (e.g., *replay* and *close-up*) is true and the hidden node (e.g., *replay*) is True. Given these counts (with sufficient statistics), we can calculate the conditional probability as $\mathrm{P}(close{-}up = Y | goal = Y) = \mathrm{P}(close{-}up = Y, goal = Y)/\mathrm{P}(goal = Y)$.

The second training phase is applied for all the temporal dependency for each event-event pair, event-hidden pair, or hidden-hidden node pair at two consecutive time slices. Every two nodes have certain temporal relationship which can be described in terms of the conditional probability. Given two nodes (e.g., $goal_t$ and $goal_{t-1}$), we can count the number of joint appearance of $goal_t$ and $goal_{t-1}$, or the single appearance of $goal_t$, and compute the conditional probability as $\mathrm{P}(goal_t = \mathrm{Y} | goal_{t-1} = \mathrm{Y}) = \mathrm{P}(goal_t = \mathrm{Y}, goal_{t-1} = \mathrm{Y})/\mathrm{P}(goal_{t-1} = \mathrm{Y})$.

The third training phase is applied to generate the conditional probability of the existing link between the evidence nodes and

the hidden nodes. The appearance of the hidden node is obtained by human observers, and the appearance of evidence node is obtained by feature extraction process. We count the number of times that the gate and the parallel lines appear simultaneously, and the number of times that the gate appear, and compute the conditional probability as $\mathrm{P}(parallel{-}line = Y|gate = Y) = \mathrm{P}(parallel{-}line = Y, gate = Y)/\mathrm{P}(gate = Y)$.

2) Qualitative Training: After the quantitative training, every two nodes in the network are somehow related. If the directional relationship (i.e., conditional probability) of any two nodes is large enough, the linkage between those two nodes is established. 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 $p(n_{e_i}|n_{c_j})$ represent the conditional probability relating the cause node n_{c_j} to the effect nodes n_{e_i} where $j = 1, \ldots J$ and $i = 1, \ldots I$, where I is the number of effect nodes. After the quantitative training, we normalize the conditional probability relating the cause node n_{c_j} to the effect node n_{e_i} as $p(n_{e_i}|n_{c_j}) = p(n_{e_i}|n_{c_j})/\Sigma_j p(n_{e_i}|n_{c_j})$.

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 (event-hidden node pair or hidden-hidden node pair) and $\{P(x)\} = \{p(n_{e_i}|n_{c_j})\}$ be the original distribution after training. M^* is the candidate network with $\{P^*(x)\} = \{p^*(n_{e_i}|n_{c_j})\}$ as the distribution after *thresholding*. We define (1) the *Size* of M^*, Size(M^*), which is the number of entries in P^* that $p^*(n_{e_i}|n_{c_j}) > t$, (2) the *Cross Entropy Distance*, $\mathrm{Dist}(P, P^*) = P(x)\Sigma_j \log P(x)/P^*(x)$.

There is a trade-off between $Size(M^*)$ and $Dist(P, P^*)$. If we have a larger threshold t then the $Size(M^*)$ will be smaller, and the cross entropy distance $Dist(P, P^*)$ will become larger, and vice versa. Therefore, we define the Acceptance Measure as $Acc(P, M^*) = Size(M^*) + kDist(P, P^*)$. Then, we use the Langrage method to choose Langrage multiplier k and the threshold t that minimize the $Acc(P, M^*)$. Finally, we use the Bayes' rule to obtain the posteriori probability $\mathrm{p(n_{c_i}|n_{e_j})}$. In the training phase, 500,000 frames are used to generate a reliable DBN.

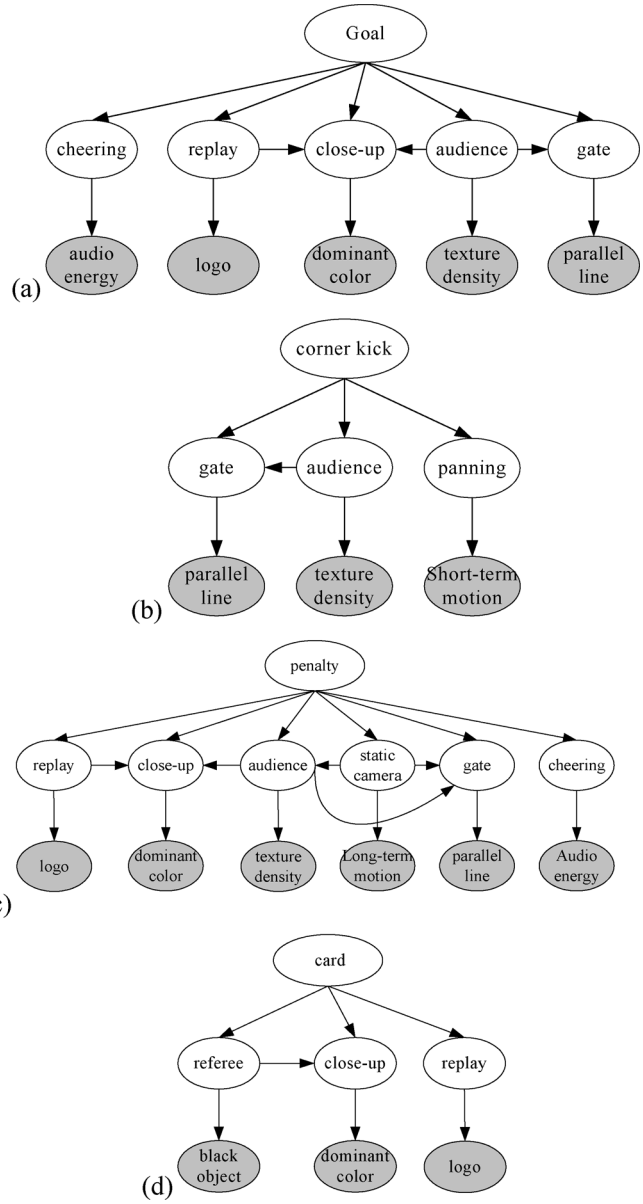


Fig. 7. Different structure of BN network for detecting the (a) goal event; (b) corner kick event; (c) penalty kick event; (d) card event.

B. DB and DBN Model

After the training processes, we generate the BN/DBN an inference of unobservable concepts based on their relevance with the observable evidences. Given the evidences as the input, the BN/DBN may infer certain high-level semantics. In Fig. 7, four BNs are illustrated: (a) Goal event; (b) corner kick event; (c) penalty kick event; and (d) card event.

For soccer videos, the evidences with shaded nodes applied to infer the goal event are parallel line, energy, and dominant color, etc. These evidence nodes are the input to the network, as shown in Fig. 7(a). These domain specific features are specified by the evidence nodes, i.e., the logo, dominant color, parallel line, and texture density. They are somehow related and the relationships among them are through the hidden nodes. In Fig. 7(b), the evidences applied to infer the corner kick event are parallel line, texture density, and motion. In the corner event, the ball moves with high velocity, so the motion is very important information.

In Fig. 7(c), the evidences (the shaded nodes) required to infer the penalty kick event are dominant color, parallel line, audio energy, etc. We find that the posteriori probability of the static camera and penalty kick event is larger in the penalty kick BN than the other BNs. In Fig. 7(d), the evidences applied to infer the card event are black objects, dominant color, and logo. Since the referee is always involved in the card event, the black object is very important information. The inference propagation in the DBN generates the occurrence possibility of the root node. After a simple decision making, we may decide whether the event exists or not.

After the training, for different event, we have developed a corresponding DBN for each BN as shown in Fig. 8. Some

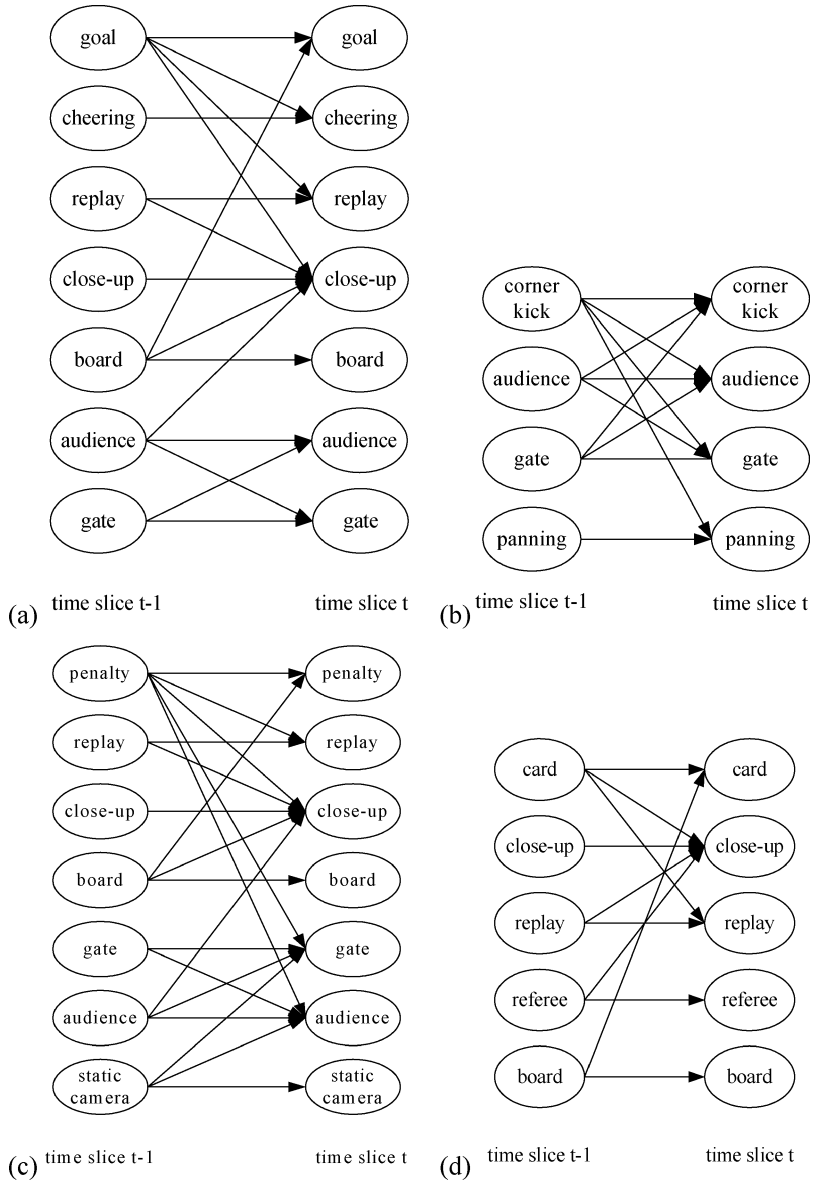


Fig. 8. Different DBN networks for (a) goal event; (b) corner kick event; (c) penalty kick event; (d) card event.

hidden nodes appear in the BN, but not in the corresponding DBN, on the other hand, some hidden nodes may be found in DBN, but not in the corresponding BN. For instance, we found no "board" in the BN [Fig. 7(a)], but it appears in the corresponding DBN [Fig. 8(a)]. It is because after the second phase of the quantitative training, the temporal causality between score board and the goal is stronger than its spatial causality. We also find the cheering node appears in BN [Fig. 7(c)] but not in DBN [Fig. 8(c)]. It is because the spatial causality of the cheering-event node is stronger than their temporal causality.

C. Propagation in Bayesian Network

Here, we apply the algorithm of probability updating in Bayesian networks. The algorithm does not work directly on the Bayesian network, but on a so-called junction tree which is a tree of clusters of variables. The clusters are also called cliques because they are cliques in a *triangulated graph*, which is a special graph constructed over the network. Each clique holds a table over the configurations of its variables, and probability propagation consists of a series of operations on these tables. The operations of propagations are rather complicated of which the details are mentioned in [8].

After the inference propagation of DBN, there are two types of decision-making: *test-decisions* and *action-decisions* [8]. The former is a decision that requires more evidence to be entered into the model if the test of the results leads to uncertainty. Whenever the decision leads to uncertainty, the system requires more information or evidence (since it is not free), it is called the test-decision. The latter is a decision that requires certain actions to change the states of the model. In our system, we consider the action-decision.

With a complete set of evidences, the final inference propagation of DBN will lead to the action-decision which can further be divided into two types: *intervening actions* and *non-intervening actions*. During inference propagation, the *inter-*

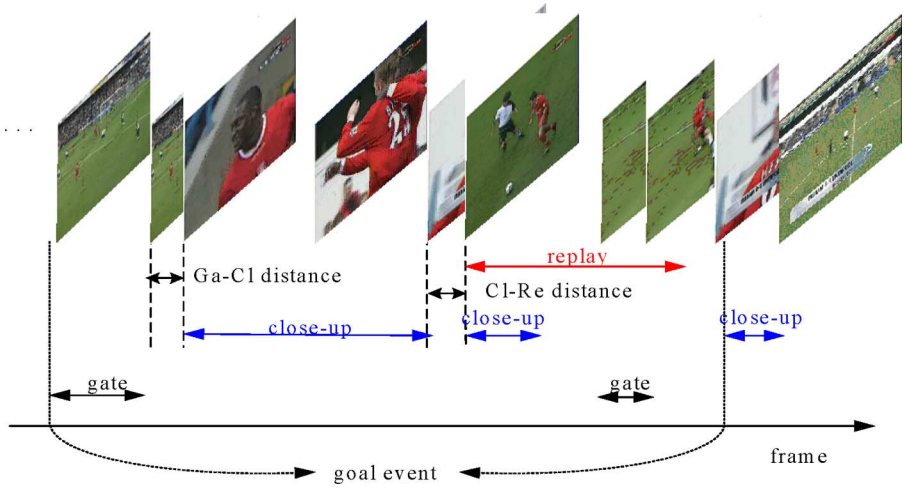


Fig. 9. The occurrence of replay, close-up, and gate in a goal event. (Color version available online at http://ieeexplore.ieee.org.)

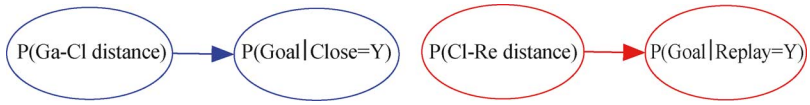


Fig. 10. TIN for goal event detection. (Color version available online at http://ieeexplore.ieee.org.)

vening action changes the *posteriori probability* distribution of the model, whereas, the nonintervening action has no impact on the model. In our system, we introduce the so-called *temporal intervening network* for taking the intervening action. With evidence propagation procedure of DBN, and the posteriori probability increment via the *temporal intervening network*, we can improve the identification detection rate of the high level semantic meaning in the soccer video.

D. Temporal Intervening Network

After the inference propagation, the confidence of the event is equi-probable, e.g., $P(Goal = Y) \approx P(Goal = N)$, and the posteriori probability of the event node and the hidden node may also be similar, e.g., $\mathrm{P}(\mathrm{Goal} = Y|Close{-}up = Y) \approx \mathrm{P}(Goal = N|Close{-}up = Y)$. However, the events in soccer video have certain regularity which can be used to differentiate the *posteriori* probabilities. 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., Fig. 9), we always find the appearance of gate. When the gate disappears, the first close-up will appear in less than 20 frames time interval. After the first 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.

We define the following abbreviations: $Goal(Go)$, $Card(Ca)$, $Gate(Ga)$, $Close{-}up(CL)$, and $Referee(Ref)$, and compute the $Gate - Close{-}up$ distance and the $Close{-}up - Replay$ distance as follows: $Ga - Cl\ distance = (Frame\ \#\ of\ 1^{st}\ close{-}up\ appears) - (Frame\ \#\ of\ gate\ disappears)$ and $Cl - Re\ distance = (Frame\ \#\ of\ 1^{st}\ replay\ appears) - (Frame\ \#\ of\ close{-}up\ disappears)$.

TABLE I
THE CONDITIONAL PROBABILITY OF THE EXISTENCE OF THE CLOSE-UP AFTER THE GATE ($\mathrm{close{-}up} = \mathrm{Y}$)

Ga-Cl distance<20	Ga-Cl distance>20	Go
0.907	0.093	Y
0.260	0.740	N

Once the regularity (i.e., $Ga{-}Cl\ distance < 20$) is found, the so-called *temporal intervening network* (TIN) is activated. When the close-up disappears, the replay will appear in less than 20 frames time interval. Once the $Close{-}up - Replay\ distance < 20$, we may also use the *temporal intervening networks* to increase $\mathrm{P}(Goal = Y|Replay = Y)$. For different events, we may also apply the TIN to change the *posteriori probabilities* for some linkages in the DBN to improve the accuracy of the final inference results as shown in Fig. 10.

1) Goal Event Example: We focus on the first close-up because the appearance of the close-ups is sequential, so we only need to know whether the gate appears and then disappears for 20 frames time interval before the first close-up. Given $close{-}up = Y/N$ and $goal = Y/N$, we find the probability of the existence of the gate in front of the close-up. Based on the training data, we can generate the conditional probabilities, i.e., $\mathrm{P}(Ga - Cl\ distance < 20|Go = \mathrm{Y}, Cl = \mathrm{Y}) = 0.907$ and $\mathrm{P}(Ga - Cl\ distance < 20|\mathrm{Go} = \mathrm{N}, \mathrm{Cl} = \mathrm{Y}) = 0.260$, as shown in Table I.

Now, at an arbitrary instance of the testing video, suppose the close-up appears, and the probability of the appearance of goal is $\mathrm{P}(Goal = Y|Close{-}up = Y) = 0.6$. If we assume $Ga - Cl\ distance < 20$, we may have a new *posterior* probability $P'(Go = Y|Cl = Y)$, as shown in the equa-

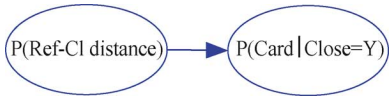


Fig. 11. TIN for card event. (Color version available online at http://ieeexplore.ieee.org.)

TABLE II
THE CONDITIONAL PROBABILITY BETWEEN THE CLOSE-UP AND THE REFEREE ($\mathrm{close-up} = \mathrm{Y}$)

Ref-Cl distance<20	Ref-Cl distance>20	Ca
0.784	0.216	Y
0.197	0.803	N

tion at the bottom of page, and $P'(Go = N|Cl = Y) = 1 - 0.839 = 0.161$. Otherwise, if the $Ga - Cl\ distance > 20$, the new *posterior* probability becomes $P'(Go = Y|Cl = Y) = P((Go = Y|Cl = Y)|Ga - Cl\ distance > 20) = 0.159$. If $Ga - Close - up$ distance < 20, the goal event occurrence probability is increased from 0.6 to 0.839. Otherwise, if $Ga{-}Close{-}up$ distance > 20, the goal event probability is reduced from 0.6 to 0.159. However, this network is not activated when close-up does not appear, the original posterior probability does not change.

2) Card Event Example: Similarly, another TIN (as shown in Fig. 11) can also be used to increase the *posteriori* probability $\mathrm{P}(Ca = Y|Cl = Y)$. In the card event, the referee appears in the global event and then appears in a close-up event. Here, we define the $Ref - Cl\ distance = (Frame\#of\ 1^{st}\ close{-}up\ appears) - (Frame\#of\ referee\ appears)$. Suppose we have the conditional probability and $Close{-}up = True$ (see Table II). During the card event, the probability of $Ref - Cl - distance < 20$ is 0.784, i.e., $P(Ref - Cl\ distance < 20|Ca = Y, Cl = Y) = 0.784$, and the probability of $Ref - Cl - distance > 20$ is $1 - 0.784 = 0.216$, i.e., $P(Ref - Cl - distance > 20|(Ca = Y, Cl = Y)) = 0.216$. If $close{-}up = True$, during the card event, we always find the referee appear before the close-up.

Now, suppose $close{-}up = True$ and $P(Ca = Y|Cl = Y) = 0.6$, if the $Ref - Cl\ distance < 20$, then we can have a new posterior probability $P'(Ca|Cl = Y) = 0.857$ and $P'(Ca = N|Cl = Y) = 1 - 0.857 = 0.143$. Otherwise, if the $Ref - Cl\ distance > 20$, the new posterior probability $P'(Ca|Cl = Y) = 0.287$. The advantage of using TIN is to improve the card event probability from 0.6 to 0.857, When $Close{-}up = True$ and the referee appears before these $close{-}ups$, the card event probability is reduced from 0.6 to 0.287.

3) Penalty Kick Event Example: We can also develop the temporal intervening network (as shown in Fig. 12) for the DBN of penalty kick event. In the penalty event, the gate always appears before the close-up. When the gate disappears, the close-up will appear in less than 20 frames time interval. In Fig. 12(b), the temporal intervening network can be also applied to DBN when $Replay = True$. In the penalty event, we often find the replay after the close-up. When the close-up disappears, the replay will appear in less than 20 frames time interval.

V. EXPERIMENTAL RESULTS

Here, we show some experimental results to illustrate the system performance. Our system is frame-based event detection which is different from shot-based event detection in that ours can identify the semantics of the video sequence at every frame instance. We have tested the proposed algorithms based on a data set of seven soccer video games for more than 11 hours from two TV broadcast stations as shown in Table III. Five soccer video programs of England Premier League from TV station one, whereas the other two soccer video programs of UEFA Cup from TV station 2. The shooting styles of the soccer videos from two TV stations are similar. The video source is MPEG-1 clips in 320 × 240 resolution at 30 frames/s. Audio is sampled at 44 kHz with 16 bits per sample.

In the experiments, we do the frame-base and shot-based event detection. The former does the inference propagation for each input frame, and makes the event detection for each time slice, whereas, the latter accumulates more evidence in the hidden nodes before the inference propagate to the event nodes for event detection of each video shot.

1) Frame-Based Event Detection: For frame-base event detection, given an input frame, the system generates the low-level evidence, initiates the inference network propagation, and then makes a semantic analysis for each frame. Each game lasts about 95 min, including the first half and the second half. To measure the performance of our system, we compute the detection rate and false alarm rate by comparing the semantics output with human observers. The detection rate and false alarm rate are defined as

$$detection\ rate = \frac{\#\ of\ frames\ identified\ as\ event\ A\ correctly}{\#\ of\ frames\ of\ event\ A}$$

$$false\ alarm\ rate = \frac{\#\ of\ frames\ misindentified\ as\ event\ A}{\#\ of\ frames\ of\ nonevent\ A}$$

$$\begin{aligned} P'(Go = Y|Cl = Y) &= P\left((Go = Y|Cl = Y)|Ga - Cl\ distance < 20\right) \\ &= \frac{P\left(Ga - Cl\ distance < 20|(Go = Y|Cl = Y)\right) \times P(Go = Y|Cl = Y)}{P(Ga - Cl\ distance < 20)} \\ &= 0.907 \times 0.6/(0.907 \times 0.6 + 0.260 \times 0.4) = 0.839 \end{aligned}$$

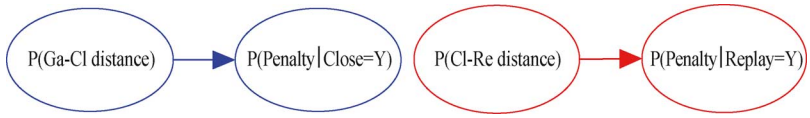


Fig. 12. TIN for the penalty kick event. (Color version available online at http://ieeexplore.ieee.org.)

TABLE III
THE TEST SEQUENCES

Sequence	【times】	# frame
England 1	【00:95:04】	171,125
England 2	【00:98:05】	176,555
England 3	【00:94:10】	169,512
England 4	【00:94:28】	170,041
England 5	【00:94:41】	170,438
UEFA 1	【00:95:14】	171,429
UEFA 2	【00:94:07】	169,421
Total	【11:05:49】	1,198,521

TABLE IV
EXPERIMENTAL RESULTS OF USING BN/DBN

	Goal	Corner	Penalty	Card
Detected frames #	178,094	91,351	2,648	5,492
Ground Truth	202,310	107,163	2,842	6,079
Average detection rate	88.03%	85.24%	93.17%	90.35%
Average false alarm rate	21.89%	0.59%	20.73%	30.46%

TABLE V
EXPERIMENTAL RESULTS OF USING BN/DBN/TIN

	Goal	Corner	Penalty	Card
Detected frames #	183,475	91,346	2,654	5,641
Ground Truth	202,310	107,163	2,842	6,079
Average detection rate	90.69%	85.24%	93.37%	92.80%
Average false alarm rate	11.52%	0.59%	0.99%	8.54%

Then, we add the DBN into the system for semantic analysis. Table IV shows the experimental results of seven complete soccer games using the BN/DBN. Finally, we add *temporal intervening network* (TIN) to BN/DBN. Table V shows the experimental results of the same seven complete soccer games by using the additional TIN. We can see that TIN improves the detection rate slightly but reduce the false alarm rate greatly.

Here, we do not apply the TIN for every event-hidden node pair. For instance, the "score board" is not included in the TIN for goal event. Since the causality between score board and the goal is very weak (in most of the cases, the appearance of score board does not necessarily indicate the goal event), so it is no use to use the TIN to increase the *posteriori* probability between the score board and goal.

The false alarm of goal event is due to the appearance of close-up and gate, which do not necessary indicate the occurrence of the goal events. In the corner event video, we always find the panning motion followed by the appearance of gate. These two cues are significant for DBN to distinguish corner event from other events. To improve the penalty and card event detection rate, we introduce the TIN model to overcome the non-expectant situation such as the referee appears in noncard event and the gate disappears. The audience and static camera are very strong cues for BN/DBN to differentiate the penalty event from the goal event.

2) Shot-Based Event Detection: We may extend the frame-based event detection to shot-based event detection. Table VI shows another statistics of the experimental results of goal event detection. The reason why the false alarm rate in Table VI is larger than the missed rate is that the offside is misidentified 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 useful information: which team has dominated the game. Similarly in England 5, the number of detected left goals of the first team is more than the number of right goals of the second team. On the other hand, in England 3, the two teams are well-matched.

For each network structure, we compute the precision and recall which are defined as

$$Precision = \frac{N_c}{N_c + N_f} \times 100\% \quad Recall = \frac{N_c}{N_c + N_m} \times 100\%$$

where N_c is the correct detection, N_m is the number of miss, N_f is the number of false alarm, $N_c + N_m$ is the number of existing events, and $N_c + N_f$ is the number of overall declaration.

Table VII shows another statistics of the 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. Left/right goal is useful information to analyze the soccer game. It provides the different performance statistics of two teams in the game. Comparing the precision rate in Tables VI and VII, we find that the precision rate in Table VII 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 a higher chance of false alarm.

The reason of the false alarm rate in Tables VI and VII is larger than the miss rate is because 1) an offside is often misindentified as the goal event, or 2) a long passing from the corner of the field is also easily misindentified as corner event. Here, we do not provide the event-detection statistics for the penalty and the card events. It is because the penalty and the card events rarely occur in our test video sequence and the experimental results do not provide sufficient data statistics.

TABLE VI
USING BN/DBN AND TIN FOR GOAL EVENT

Sequence	England 1		England 2		England 3		England 4		England 5		UEFA 1		UEFA 2	
	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 VII
USING BN/DBN AND TIN FOR CORNER EVENT

Sequence	England 1		England 2		England 3		England 4		England 5		UEFA 1		UEFA 2	
	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

VI. CONCLUSIONS

We have proposed a video program understanding system. Given an input sequence, the system will collect the low-level evidence, and applies the inference engine in BN/DBN to infer high-level semantic concepts that interpret the semantic content of video sport program. The main contribution of this paper is to add the temporal intervening network to DBN to improve the semantic interpretation accuracy. We have demonstrated that our system can understand the semantic concepts effectively.

REFERENCES

[1] L. Xie, S.-F. Chang, A. Divakaran, and H. Sun, "Structure analysis of soccer video with Hidden Markov Models," in *Proc. ICASSP*, Orlando, FL, May 2002, vol. 4, pp. 4096–4099.

[2] P. Chang, M. Han, and Y. Gong, "Extract highlights from baseball video game video with Hidden Markov Model," in *Proc. IEEE ICIP*, 2002.

[3] G. Xu, Y. F. Ma, H. J. Zhang, and S. Yang, "A HMM based semantic analysis framework for sports game event detection," in *Proc. IEEE ICIP*, 2003, vol. 1, pp. 25–28.

[4] A. Ekin, A. M. Tekalp, and R. Mehrotra, "Automatic soccer video analysis and summarization," *IEEE Trans. Image Process.*, vol. 12, no. 7, pp. 796–807, Jul. 2003.

[5] C. G. M. Snoek and M. Worring, "Time interval maximum entropy based event indexing in soccer video," in *Proc. IEEE Int. Conf. Multimedia and Expo (ICME)*, Baltimore, MD, 2003.

[6] M. Han, W. Hua, T. Chen, and Y. Gong, "Feature design in soccer video indexing," in *PCM 2003 Conf.*, Singapore, Dec. 15–18, 2003.

[7] Y. Gong, M. Han, W. Hua, and W. Xu, "Maximum entropy model-based baseball highlight detection and classification," *Comput. Vis. Image Understand.*, vol. 26, pp. 181–199, 2004.

[8] F. V. Jensen, *An Introduction to Bayesian Networks*. New York: Springer, 1996.

[9] X. Sun, G. Jin, M. Huang, and G. Xu, "Bayesian network based soccer video event detection and retrieval," in *Multispectral Image Processing and Pattern Recognition*, Beijing, China, Oct. 2003.

[10] H. C. Shih and C. L. Huang, "MSN: Statistical understanding of broadcasted sports video using multilevel semantic network," *IEEE Trans. Broadcast.*, vol. 51, no. 4, pp. 449–459, Dec. 2005.

[11] J. Assfalg and M. Bertini, "Semantic annotation of soccer videos: automatic highlights identification," *Comput. Vis. Image Understand.*, vol. 91, no. 3, 2003.

[12] P. Xu, L. Xie, and S.-F. Chang, "Algorithms and system for segmentation and structure analysis in soccer video," in *Proc. IEEE ICME*, Tokyo, Japan, 2001, pp. 721–724.

[13] K. Wan, J.-H. Lim, C. Xu, and X. Yu, "Real-time camera field-view tracking in soccer video," in *Proc. ICASSP*, Hong Kong, 2003, vol. 3, pp. 185–188.

[14] H. C. Shih and C. L. Huang, "Detection of the highlights in baseball video program," in *Proc. IEEE ICME*, Taipei, Taiwan, R.O.C., Jun. 2004, vol. 1, pp. 595–598.

[15] D. Zhong and S.-F. Chang, "Structure analysis of sports video using domain models," in *Proc. IEEE ICME*, Tokyo, Japan, Aug. 2001, pp. 713–716.

[16] W. Zhou, A. Vellaikal, and C.-C. J. Kuo, "Rule-based video classification system for basketball video indexing," in *ACM Multimedia Conf.*, Los Angeles, CA, Nov. 2000.

[17] S. Takagi and S. Hattori, "Sports video categorizing method using camera motion parameters," in *Proc. IEEE ICME*, Baltimore, MD, Jul. 2003, vol. 2, pp. 461–464.

[18] D. A. Sadlier, N. O'Connor, S. Marlow, and N. Murphy, "A combined audio-visual contribution to event detection in field sports broadcast video," in *IEEE Int. Symp. Signal Processing and Information Technology*, Darmstadt, Germany, 2003.

[19] C. Wu, Y.-F. Ma, H.-J. Zhang, and Y.-Z. Zhong, "Events recognition by semantic inference for sports video," in *Proc. IEEE ICME 2002*, Lausanne, Switzerland, 2002, vol. 1, pp. 805–808.

[20] M. Petkovic, V. Mihajlovic, W. Jonker, and S. Kajan, "Multi-model extraction of highlights from formula 1 programs," in *Proc. IEEE ICME*, Lausanne, Switzerland, 2002.

[21] A. Garg, V. Pavlovic, and J. M. Rehg, "Boosted learning in dynamic bayesian networks for multimodal speaker detection," *Proc. IEEE*, vol. 91, no. 9, pp. 1355–1369, Sep. 2003.

[22] J. Forbes, T. Huang, K. Kanazawa, and S. J. Russell, "The BATmobile: toward a Bayesian automated taxi," in *Proc. IJCAI*, 1995.

[23] J. N. Hwang and Y. Luo, "Automatic object-based video analysis and interpretation: a step toward systematic video understanding," in *IEEE ICASSP*, Orlando, FL, May 2002, vol. 4, pp. 4084–4087.

[24] V. Mihajlovic and M. Pekovic, Dynamic Bayesian Networks: A State of the Art CS Dept., Univ. Twente. Enschede, The Netherlands, 2001.

[25] H. Pan, B. Li, and M. Sezan, "Automatic detection of replay segments in broadcast sports programs by detection of logos in scene transitions," in *Proc. IEEE ICASSP*, Orlando, FL, May 2002, vol. 4, pp. 3385–3388.

[26] C. L. Huang and B. Y. Liao, "A Robust scene-change detection method for video segmentation," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 11, no. 12, pp. 1281–1288, Dec. 2001.

Chung-Lin Huang (SM'04) received the B.S. degree in nuclear engineering from the National Tsing-Hua University, Hsinchu, Taiwan, R.O.C., in 1977, the M.S. degree in electrical engineering from National Taiwan University, Taipei, in 1979, and the Ph.D. degree in electrical engineering from the University of Florida, Gainesville, in 1987.

From 1987 to 1988, he was with Unisys, Mission Viejo, CA, as a Project Engineer. Since August 1988, he has been with the Electrical Engineering Department, National Tsing-Hua University, Hsinchu. Currently, he is a Professor in the same department. His research interests are in the area of image processing, computer vision, and visual communication.

Dr. Huang received the Distinguish Research Awards from the National Science Council, Taiwan, in 1993 and 1994. In November 1993, he received the best paper award from the ACCV, Osaka, Japan, and in August 1996, he received the best paper award form the CVGIP Society, Taiwan. In December 1997, he received the best paper award from IEEE ISMIP Conference held Academia Sinica, Taipei. In 2002, he received the best paper annual award from the Journal of Information Science and Engineering, Academia Sinica, Taiwan.

Huang-Chia Shih (S'03) was born in Changhua, Taiwan, R.O.C., in 1978. He received the B.Sc. degree (with first place awards) in electronic engineering from the National Taipei University of Technology, Taipei, Taiwan, in 2000 and the M.S. degree in electrical engineering from the National Tsing Hua University (NTHU), Hsinchu, Taiwan, in 2002. He is currently pursuing the Ph.D degree in electrical engineering at NTHU.

His research interests are content-based video summarization, video indexing and retrieval, applications of statistical models in multimedia processing. During summer 2002, he was a Summer Intern at Computer & Communications Research Labs, Industrial Technology Research Institute, Taiwan.

Mr. Shih has received several awards and prizes, including the Awards of Excellent Engineering Student from the Chinese Institute of Engineers, in 2000 and Awards of the Outstanding College Youth from China Youth Corps in 2000. In 2006, he is recognized as the ambassadorial scholar from Rotary International. He has also served on the program committee of several international conferences and workshops.

Chung-Yuan Chao was born in Taipei, Taiwan, R.O.C., in 1980. He received the B.S. degree in electrical engineering from Chung Yuan Christian University, Chungli, Taiwan, in 2002, and the M.S. degree in electrical engineering from the National Tsing Hua University, Hsinchu, Taiwan, in 2004.

Currently, he is an Engineer with Magic Pixel, Inc., Hsinchu. His research interests including sports video analysis, content-based video retrieval, and semantic analysis.