A Generic Framework for Semantic Sports Video Analysis Using Dynamic Bayesian Networks

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*
Email: {feiwang, jtli}@ict.ac.cn, {yfma, hjzhang}@microsoft.com

Abstract

Automatic detection of semantic events in sport videos is a challenging task. In this paper, we propose a multimodal multilayer statistical inference framework for semantic sports video analysis using Dynamic Bayesian Networks (DBNs). Based on this framework, three instances including factorial hierarchical hidden Markov model (FHHMM), coupled hierarchical hidden Markov model (CHHMM), and product hierarchical hidden Markov model (PHHMM), are constructed and compared. Play-break detection in soccer videos is used as a testbed with hierarchical hidden Markov model (HHMM) as a baseline. Experimental results indicate the superior capability of the PHHMM, because it not only effectively models dynamic interactions between different modalities, but also sufficiently utilizes context constraints in multilayer structures.

Keywords: *event detection, sports video analysis, statistical modeling.*

1. Introduction

The growing amount of multimedia information in the form of digital video is driving the development of content-based video analysis and retrieval technologies in recent years [1]. However, due to a wide variety of video content, it is still difficult to extract automatically high level semantics from raw video data. Thus, most of current works focus on specific domain in order to investigate the roles of domain knowledge. As a great favorite of a large audience over the world, sports video represents an important application domain. Exciting commercial appeals and academic promises have attracted attentions from various disciplines [2-17].

There are two distinct characteristics in sports videos that differentiate them from other video programs (e.g., news and movie). These characteristics represent the domain knowledge in the sports video analysis. First, sports video is composed of repetitive domain-specific events, which are intrinsically multimodal. Different information sources including closed caption text, speech, sound, camera motion, and visual scene, are used by the broadcasters to convey the meaning. Thus, the integration of multimodal information to detect events is an effective solution for semantic sports video analysis [6-12]. Second, most of sports games have tree structures. That is, the relationships among the events follow a set of rules. For example, a tennis game is divided first into sets, then games and serves. It is important to improve the accuracy of individual event detection by sufficiently utilizing such multilayer constraints [13-17].

In this paper, we present a generic framework for sports video analysis based on Dynamic Bayesian Networks (DBNs). This framework not only detects events with multimodal information, but also utilizes multilayer structure among events. Although both issues have been addressed in previous works, to consider the two issues together has not gotten sufficient attentions. Within the unified probabilistic framework proposed in this paper, the two aspects may enhance each other resulting in more robust and effective event detection. From this framework, three instances are derived for the multimodal and multilayer analysis of sports video, including factorial hierarchical hidden Markov model (FHHMM), coupled hierarchical hidden Markov model (CHHMM), and product hierarchical hidden Markov model (PHHMM). For a comparison purpose, the three variations as well as conventional hierarchical hidden Markov model (HHMM) are all applied to play-break

Proceedings of the 11th International Multimedia Modelling Conference (MMM'05)
1550-5502/05 $20.00 © 2005 **IEEE**

detection in soccer videos. Experimental results show that the PHHMM outperforms all the other models.

The rest of paper is organized as follows: In Section 2, we briefly review the related work of sports video analysis. The proposed multimodal multilayer DBN models are presented in Section 3. In Section 4, we apply the three models as well as conventional HHMM to play-break detection in soccer videos. The performance comparison results are reported in Section 5. Finally, Section 6 concludes the paper

2. Related Work

Recently, the integrated use of multimodal data is an emerging trend in semantic sports video analysis. In this section, we first survey the related work on this issue. These multimodal based approaches usually focus on isolated event recognition. Different from these approaches, our previous works also considers relationships between events [16, 17], which are briefly discussed in the second part.

2.1. Multimodal Analysis for Sports Videos

Existing techniques of multimodal integration include feature fusion and decision fusion methods. In feature fusion methods, observation vectors are obtained by concatenation of low-level features from different modalities. The approach proposed in [7] concatenated image, audio and text features from consecutive scene shots into a feature vector, and then a maximum entropy method was used to detect key events of baseball videos. Because the concatenation tends to result in a high dimension feature vector, the feature fusion method is often followed by a dimensionally reduction transform. However, such a transform is argued in [20], because Wang *et al.* found that the correlation between features from different modalities (e.g. audio, color, and motion) was very low. Furthermore, these feature fusion methods, also called early fusion, are not effective for asynchrony and dynamic interactions between different modalities [6].

Decision fusion methods, also called late fusion, are more popular than feature fusion methods in sports video analysis research. The decisions from different modalities are combined at a high level to generate an overall result. Different combination strategies have been proposed, including rule-based reasoning [8-10], linear combination [11] and machine learning methods such as DBNs [12]. However, both rule-based reasoning and linear combination often result in poor generality [4]. First, explicitly setting the inference rules may not be easy for a large number of cues. Second, some hard thresholds have to be chosen. Apart from these deterministic approaches, the effectiveness of probabilistic inference for fusing the evidences from different modalities has been demonstrated by [12], in which multimodal features were fused by DBNs to extract highlights from Formula 1 race videos.

2.2. Multilayer Analysis for Sports Videos

As aforementioned, the multilayer structure among events is another important characteristic of sports video. The fundamental idea behind multilayer analysis is to divide the event detection problem into two stages – recognition of primitives and recognition of structure. For example, a typical home run in a baseball video can be composed of a pitch view followed by an audience view and then a running close-up view. Although the exact contents of video streams differ from game to game, such production styles and editing patterns are followed by the broadcasters to help viewers understand the game. One reason for the multilayer analysis is that 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 context constraints, high-level semantics are inferred from low-level primitives. On the other hand, low-level primitives may be directly mapping to low-level features. In addition, with the multilayer framework, the semantic events may be segmented and recognized simultaneously.

In the literatures, there are two kinds of multilayer analysis techniques, i.e. syntactic-based and statistical-based methods. Motivated by the analogy between natural language and sport video, we introduced Context-Free Grammars (CFGs) and compiler principles to sports video parsing in [17]. Probabilistic aspects of syntactic-based methods were presented in [21], but not for sports video analysis. They used the Stochastic Context-Free Grammar (SCFG) parser to recognize visual activities. However, one of difficulties in syntactic-based methods is to automatically learn grammars and parameters from training data.

Statistical-based methods for multilayer analysis are always built upon probabilistic graphical models, such as HMMs and their variances. These methods combine intuitive graphical representations with efficient algorithms for statistical inference and learning. Xie *et al.* [15] suggested an unsupervised method using hierarchical hidden Markov models (HHMMs) to discover multilevel video structures. Instead of using unsupervised strategy, our previous work [16]

presented a multilayer framework based on HMMs for sports game event detection, in which the models are learned from the transcriptions of pre-defined events.

3. Multimodal Multilayer Model

Both multimodal clues and multilayer constraints are critical for semantic sports video analysis. However, there are very few efforts on a unified statistical solution for both of them. In this paper, we present a multimodal and multilayer framework for this issue using DBNs. Based on the DBN framework, three multimodal multilayer models, including FHHMM, CHHMM and PHHMM, are constructed.

DBNs are directed graphical models, in which hidden states are represented in terms of individual variables or factors. DBNs extend traditional Bayesian Networks (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. As a class of probabilistic graph models, DBNs provide us a feasible and powerful method to model sports events by fusing multimodal data and combining multilayer constraints.

In this section, we first give the DBN representations of these models, and then discuss the algorithms for learning and inference.

3.1. Representation

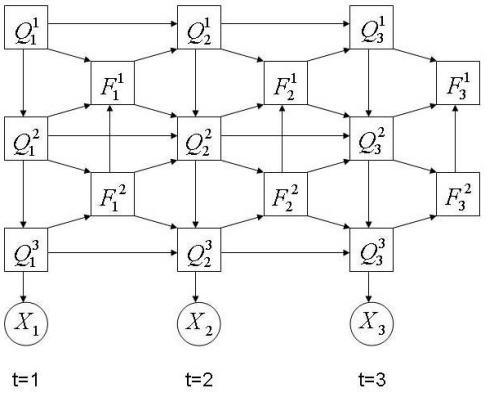


Figure 1. DBN representation of 3-level HHMM.

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.

Figure 1 represents an HHMM as a DBN. The state of the HHMM at level $d = 1 \ldots D$ (from top to bottom) and time t is represented by Q_t^d. The number of the states at level d is denoted as n_d. F_t^d is an indicator variable that is tuned on only if the sub-HMM of Q_t^d has just finished, i.e., entered the end state. Note that if $F_t^d = 1$, then $F_t^{d'} = 1$ for all lower levels ($d' > d$). Besides the topology, the associated CPDs are defined as follows. The top, middle and bottom layers of the hierarchy are considered separately, since they have different local topology.

Top level:

$$P(Q_1^1 = i) = \pi_1^1(i) \tag{1}$$

$$P(Q_t^1 = j \mid Q_{t-1}^1 = i, F_{t-1}^1 = f) = \begin{cases} \delta(i,j) & \text{if } f = 0 \\ a_1^1(i,j) & \text{if } f = 1 \end{cases} \tag{2}$$

Middle level:

$$P(Q_1^d = i \mid Q_1^{d-1} = k) = \pi_k^d(i) \tag{3}$$

$$P(Q_t^d = j \mid Q_{t-1}^d = i, Q_t^{d-1} = k, F_{t-1}^d = f, F_{t-1}^{d-1} = g) = \begin{cases} \delta(i,j) & \text{if } f = 0 \\ a_k^d(i,j) & \text{if } f = 1 \text{ and } g = 0 \\ \pi_k^d(j) & \text{if } f = 1 \text{ and } g = 1 \end{cases} \tag{4}$$

$$P(F_t^d = 1 \mid Q_t^d = k, Q_t^{d+1} = i, F_t^{d+1} = f) = \begin{cases} 0 & \text{if } f = 0 \\ a_k^d(i, end) & \text{if } f = 1 \end{cases} \tag{5}$$

Bottom level:

$$P(Q_t^D = j \mid Q_{t-1}^D = i, Q_t^{D-1} = k, F_{t-1}^{D-1} = 0) = a_k^D(i,j) \tag{6}$$

$$P(Q_t^D = j \mid Q_{t-1}^D = i, Q_t^{D-1} = k, F_{t-1}^{D-1} = 1) = \pi_k^D(j) \tag{7}$$

$$P(F_t^{D-1} = 1 \mid Q_t^{D-1} = k, Q_t^D = i) = a_k^d(i, end) \tag{8}$$

$$P(X_t \mid Q_t^D = i) = N(X_t, u_i, \sigma_i) \tag{9}$$

where π_k^d represents the initial state distribution for the sub-HMM of state $Q_t^{d-1} = k$, a_k^d(i, j) is the state transition probability, and $N(X_t, \mu_i, \sigma_i)$ is the probability of the observation X_t given the ith state of the bottom node. Here, the observation probability is modeled using Gaussian, while other models are possible, such as Gaussian mixture model. As a whole, the parameter set Θ of an HHMM is formed as follows:

$$\Theta = (\bigcup_{d=1}^{D} \bigcup_{k=1}^{n_{d-1}} \{\pi_k^d, a_k^d\}) \cup (\bigcup_{i=1}^{n_D} \{u_i, \sigma_i\}) \tag{10}$$

A two-level HHMM model has been used in [14] to model structure in video. They modeled individual recurring events as HMMs, and the high-level transitions between the events as another level of Markov chain. In [16], we found sports game video is essentially an event sequence with a hierarchical structure, and by the need for a multilayer HHMM

model. However, both the HHMM models cannot be effectively applied for fusing multimodal information, since they take the observation as one data stream.

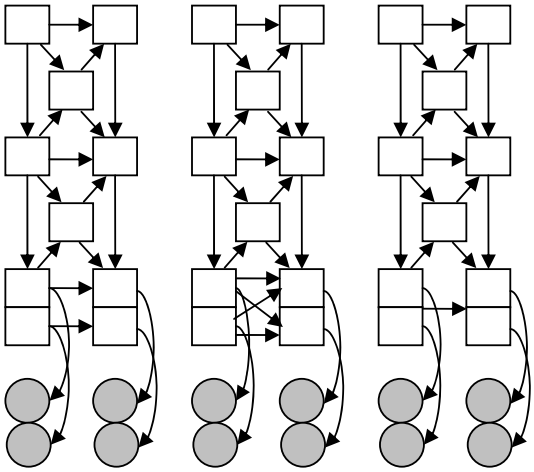


Figure 2. Multimodal multilayer models: (a) FHHMM; (b) CHHMM; (c) PHHMM.

An interesting instance of the DBNs is the so-called coupled hidden Markov model (CHMM), which has been successfully applied in audio-visual speech recognition [19]. Following this perspective, we extend HHMMs for the purpose of multimodal and multilayer sports video analysis. Figure 2 illustrates different fusion strategies of multimodal information under this framework. Compared with HHMMs, this framework has two advantages: (1) avoiding high dimensional feature vectors; and (2) modeling dynamic interactions between different modalities effectively.

An FHHMM as shown in Figure 2(a) generalizes an HHMM by representing the hidden state by a set of variables or factors. These factors are assumed to be independent of each other, but they all contribute to the high layer transitions. Except the bottom layer, an FHHMM is the same as an HHMM. At the bottom layer, the parameters of an FHHMM are described as:

$$P(Q_t^D = \mathbf{j} \mid \begin{matrix} Q_{t-1}^D = \mathbf{i}, Q_t^{D-1} = k \\ F_{t-1}^{D-1} = 0 \end{matrix}) = \prod_s a_k^s(i_s, j_s) \quad (11)$$

$$P(Q_t^D = \mathbf{j} \mid \begin{matrix} Q_{t-1}^D = \mathbf{i}, Q_t^{D-1} = k \\ F_{t-1}^{D-1} = 1 \end{matrix}) = \prod_s \pi_k^D(j_s) \quad (12)$$

$$P(F_t^{D-1} = 1 \mid Q_t^{D-1} = k, Q_t^D = \mathbf{i}) = a_k^d(\mathbf{i}, end) \quad (13)$$

$$P(X_t \mid Q_t^D = \mathbf{i}) = \prod_s N(X_t^s, u_{i_s}, \sigma_{i_s}) \quad (14)$$

where $\mathbf{i} = [i_1, \ldots, i_S]$ is the set of states in each stream at the bottom layer, $a_k^s(i_s, j_s)$ is the state transition to j_s from state i_s in stream s, $\pi_k^s(i_s)$ and $N(X_t^s, \mu_{i_s}, \sigma_{i_s})$ are the initial probability and the observation probability of state i_s in stream s respectively. The FHHMM allows for more flexibility than the HHMM in modeling the state asynchrony from different modalities, while omits the natural correlation along time between them.

A CHHMM adopts a more complex strategy for multimodal integration. It allows the nodes from different streams to interact, and at the same time to have their own observation, as shown in Figure 2(b). The elements of a CHHMM at the bottom layer are described as:

$$P(Q_t^D = \mathbf{j} \mid \begin{matrix} Q_{t-1}^D = \mathbf{i}, Q_t^{D-1} = k \\ F_{t-1}^{D-1} = 0 \end{matrix}) = \prod_s a_k^s(\mathbf{i}, j_s) \quad (15)$$

$$P(Q_t^D = \mathbf{j} \mid \begin{matrix} Q_{t-1}^D = \mathbf{i}, Q_t^{D-1} = k \\ F_{t-1}^{D-1} = 1 \end{matrix}) = \prod_s \pi_k^D(j_s) \quad (16)$$

$$P(F_t^{D-1} = 1 \mid Q_t^{D-1} = k, Q_t^D = \mathbf{i}) = a_k^d(\mathbf{i}, end) \quad (17)$$

$$P(X_t \mid Q_t^D = \mathbf{i}) = \prod_s N(X_t^s, u_{i_s}, \sigma_{i_s}) \quad (18)$$

Different from the FHHMM, the transition probability of each bottom node is computed as joint probability of the set of states at previous time. With the constraint $a_k^s(\mathbf{i}, j_s) = a_k^s(i_s, j_s)$, a CHHMM is reduced to an FHHMM.

A PHHMM can be seen as a standard HHMM, where each state of the bottom node is represented by a set of states, one emitting its own observation. The parameters of a PHMM at the bottom layer are:

$$P(Q_t^D = \mathbf{j} \mid \begin{matrix} Q_{t-1}^D = \mathbf{i}, Q_t^{D-1} = k \\ F_{t-1}^{D-1} = 0 \end{matrix}) = a_k^D(\mathbf{i}, \mathbf{j}) \quad (19)$$

$$P(Q_t^D = \mathbf{j} \mid Q_{t-1}^D = \mathbf{i}, Q_t^{D-1} = k, F_{t-1}^{D-1} = 1) = \pi_k^D(\mathbf{j}) \quad (20)$$

$$P(F_t^{D-1} = 1 \mid Q_t^{D-1} = k, Q_t^D = \mathbf{i}) = a_k^d(\mathbf{i}, end) \quad (21)$$

$$P(X_t \mid Q_t^D = \mathbf{i}) = \prod_s N(X_t^s, u_{i_s}, \sigma_{i_s}) \quad (22)$$

The use of PHHMM is justified because it allows for state asynchrony, since each of the bottom nodes can be in any combination of states from different modalities. Unlike the FHHMM, the PHHMM preserves the natural correlation of these modalities along time. Furthermore, the PHHMM also allows for the correlation at the same time, which is prohibited by the CHHMM.

3.2. Learning and Inference

Given the DBN representations of these models, there are two computational tasks that must be performed to segment and recognize 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. In this paper, the structures of the DBN models are determined, but the states of each node are not fully observable in training data. Based on our multilayer models, the higher-level structure elements usually correspond to observable semantic events, while the bottom-level states represent hidden primitives. These hidden primitives constitute the events and also have directly mapping to low-level features. For example, in the experiments for play-break event detection in soccer videos, we labeled the plays and breaks in the training data, but the primitives below these events were unlabelled. Thus, the DBN models in this situation can be trained using the expectation maximization (EM) algorithm [18].

As for the inference task, we convert the HHMM into an equivalent HMM, and then use the forward-backward algorithm [22]. This is the method of choice, since the forward-backward algorithm is exact, and is very simple to implement. Another class of algorithms are based on the junction tree algorithm for BNs, since A DBN can be unrolled as a BN.

In order to apply the forward-backward algorithm, we need to define the parameters of the equivalent HMM. For example, an equivalent HMM of an HHMM is defined as follows:

$$\pi(\mathbf{i}) = \pi_1^1(i_1)\prod_{d=2}^{D}\pi_{i_{d-1}}^{d}(i_d) \tag{23}$$

$$a(\mathbf{j}\mid\mathbf{i}) = a_{i_{r-1}}^{r}(i_r, j_r)\prod_{d=r+1}^{D}a_{i_{d-1}}^{d}(i_d, end)\pi_{j_{d-1}}^{d}(j_d) \tag{24}$$

$$b(X_t\mid\mathbf{i}) = N(X_t, \mu_{i_D}, \sigma_{i_D}) \tag{25}$$

where the state transition of the whole HHMM is encoded by the vector $\mathbf{i} = [i_1,\ldots,i_D]$ and $\mathbf{j} = [j_1,\ldots,j_D]$, and $i_d = j_d$ if $d < r$. That is, the state transition occurs at the level r, when the states of higher levels ($d < r$) remain and the states of lower levels ($d > r$) exit. Then the Viterbi algorithm is used to decode the optimum state space for event segmentation and recognition.

Similar to the inference of HMMs, the forward variable defined as $\alpha_t(\mathbf{i}) = P(X_1, \ldots, X_t, Q_t = \mathbf{i})$ and the backward variable defined as $\beta_t(\mathbf{i}) = P(X_{t+1}, \ldots, X_T \mid Q_t = \mathbf{i})$ can be computed iteratively with the forward-backward algorithm. The forward and backward variables are then used to re-estimate the parameters of the HHMM with the EM algorithm (also called the Baum-Welch method) as follows.

E step: since the state space is not full observable, we calculate the expected transition and initialization of states on the training data.

$$\begin{aligned}\xi_t(\mathbf{i},\mathbf{j}) &= P(Q_t = \mathbf{i}, Q_{t+1} = \mathbf{j} \mid X_1,\ldots,X_T,\Theta)\\ &= \frac{\alpha_t(\mathbf{i})a(\mathbf{j}\mid\mathbf{i})b(X_{t+1}\mid\mathbf{j})\beta_{t+1}(\mathbf{j})}{\sum_{\mathbf{i}}\sum_{\mathbf{j}}\alpha_t(\mathbf{i})a(\mathbf{j}\mid\mathbf{i})b(X_{t+1}\mid\mathbf{j})\beta_{t+1}(\mathbf{j})}\end{aligned} \tag{26}$$

$$\begin{aligned}\gamma_t(\mathbf{i}) &= P(Q_t = \mathbf{i} \mid X_1,\ldots,X_T,\Theta)\\ &= \sum_{\mathbf{j}}\xi_t(\mathbf{i},\mathbf{j}) = \frac{\alpha_t(\mathbf{i})\beta_t(\mathbf{j})}{\sum_{\mathbf{i}}\sum_{\mathbf{j}}\alpha_t(\mathbf{i})\beta_t(\mathbf{j})}\end{aligned} \tag{27}$$

M step: the parameters of the model can be re-estimated by using above expected values, which leads to a greater likelihood of the training data. Here, the update on the equivalent HMM is mapped back to the original HHMM.

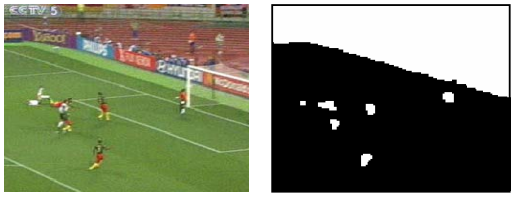
Because the EM algorithm is known to converge to a local maximum of data likelihood, the choice of the initial parameters of the model is a critical issue. In this paper, we adopt an efficient method for the initialization using the K-means and Viterbi algorithm. First, we use the K-means algorithm to cluster samples in each data stream. By this means, each sample in the training set is assigned with a label, which corresponds to the hidden primitives, i.e., the states at the bottom layer. Now, with the full observable states, the CPDs of the model can be directly determined by computing statistics from the segmented data. Then, we use the Viterbi algorithm to obtain the optimal state sequence, which can be used to estimate the new parameters of the CPDs again. This step is repeated until the difference between the observation probabilities of the training sequences at consecutive iterations falls below a convergence threshold.

4. Play-Break Detection in Soccer Videos

We apply the proposed framework to play-break detection in soccer videos, which is also a testbed for our comparison experiment. Play-break detection is an important task for video summary and further semantic analysis. Play events are defined when a soccer ball is in the field and game is going on; break events refer to the compliment set when game is ceased by some reasons, such as a foul or the ball going out of the field. With detected play-break segments, the concise summary consisting plays can be generated. Also, play-break detection is the basic step for high-level event recognition, such as goal detection.

In our implementation, color and motion features are extracted from video first. As stated in [20], color and motion features are derived from different modalities and the correlation between them is very low. On the other hand, for the play-break detection, the two modalities carry complementary information about underlying process. Thus, we seek a framework to model dynamic interactions between individual modalities as well as context constraints in a multilayer state space. Here, we don't use audio clues, because they are not very discriminative for play/break in soccer videos based on our observation.

In [14], motion intensity and dominant color ratio were used in HMMs for play/break. Then classification and segmentation were performed with dynamic programming based on the likelihood outputs from the HMMs. According to shot classification and game rules, a straightforward method of checking shot transition patterns was proposed in [9]. However, these methods are only focused on play-break detection and lacks of generality. Based on their work in [14], Xie *et al.* [15] presented a more generic approach to structure discovery from long video sequence using HHMMs. Their experiments showed that the HHMM based approach was effective for play-break detection. In this paper, we also implemented the conventional HHMM as a baseline for performance comparison.



(a) (b)

Figure 3. Field color map: (a) original image; (b) binary image.

A low level set of color and motion features are employed in our experiments so as to demonstrate the generality of the proposed framework. The color feature X_c is extracted based on the field color. The field color is the dominant color in the scene, which usually indicates the appearance of the field, thus important to characterize sports scenes. The field color is determined by an adaptive field color model by picking up dominant color value throughout a number of frames. Then, for each frame, the image is masked with this field color to achieve a binary map, as shown in Figure 3(b). Geometric moments are computed to represent the shape characteristics of the field color map. Assuming a frame has $M \times N$ pixels, a moment of order (p, q) is given by

$$m_{pq} = \sum_{y=0}^{N-1} \sum_{x=0}^{M-1} x^p y^q f(x, y) \tag{31}$$

Based on these geometric moments, three color-based descriptors are derived, including field area (32), horizontal variance (33), and vertical variance (34).

$$A = \frac{m_{00}}{MN} \tag{32}$$

$$\sigma_x = \frac{1}{M} \sqrt{m_{20} / m_{00} - m_{10}^2 / m_{00}^2} \tag{33}$$

$$\sigma_y = \frac{1}{N} \sqrt{m_{02} / m_{00} - m_{01}^2 / m_{00}^2} \tag{34}$$

The motion feature X_m, representing intensity (35) and entropy (36), is computed with motion vectors in a frame as follows:

$$m_i = \frac{1}{K} \sum \sqrt{v_x^2 + v_y^2} \tag{35}$$

$$m_e = -\sum_{i=0}^{8} \frac{h(i)}{K} Log \frac{h(i)}{K} \tag{36}$$

where K is the total number of the motion vectors $\boldsymbol{v} = [v_x, v_y]$, and $h(i)$ is the number of the motion vectors in one of eight directions ($i = 0$ means still). The motion feature carries complementary information to color-based features. For example, a field view with high motion intensity and low entropy often results from camera pan during a play; while a close-up view of player usually has high entropy.

As a whole, two data streams are produced from different modalities, X_c the color feature and X_m the motion feature. We extract them from the video sequence by uniformly sampling. Then a two-layer DBN model (i.e. FHHMM, CHHMM or PHHMM) is used to fuse the two modalities for inferring the maximum likelihood series of play/break events. The bottom layer represents the hidden primitives at each sample composing the play/break events; the top layer represents the events and the transitions between them.

5. Experiments

The experiments carry two objectives. The first is to evaluate the improvement in information fusion brought by the multimodal multilayer models, i.e. FHHMM, CHHMM and PHHMM. The second is to compare the performance of the proposed fusion architectures with other fusion techniques.

To fulfill the first objective, we built an HHMM based system as the baseline system. The system was first trained and tested only using the color modality, then only using the motion modality. The color-only system is based on the HHMM with two layers, the top layer corresponding to the events and the bottom layer

for the hidden states. The motion-only system has a similar topology. To achieve the second objective, we implemented a common form of the feature fusion approach, i.e. fusion by concatenating the color and motion feature vectors. For the feature fusion scheme, the HHMM based system identical to the color-only system uses the same output distribution for the concatenated feature vectors of a state. Our FHHMM, CHHMM or PHHMM based system uses separate distributions to model the different modalities of a joint state. In the experiments, we tried different state number from 5 to 9 at the bottom layer of each model to find the optimal performance.

The data set used for the evaluation includes 20 video clips (from a few minutes to ten minutes). These clips were randomly selected from 5 soccer matches, which involve different teams, stadiums and cable companies. All video clips are in MPEG-1 format, 352×288 size, and 25 f/s. The color and motion features are computed at interval of 0.5 second. The ground truth was labeled manually in advance. Due to the limited amount of the data, a cross validation scheme was used. Namely, the models were trained on a subset containing 90% of the available sequences and tested on the remaining 10%; this process was repeated until all sequences had been covered in testing.

The evaluation is measured both at frame level and segment level by comparing to the ground truth. At frame level, the evaluation calculates the ratio of the number of correctly annotated frames by the system to the total number of frames in the ground truth, as shown in Table 1. This measurement evaluates the performance at each frame, and is suitable for the case when users want to know the statistics of plays/breaks globally, such as time percent.

The segment-level evaluation, as shown in Table 2, is useful when users want to request to see a specific event segment. It calculates the precision and recall rates for each type of events. We counted each detected event as a hit if it temporally overlapped with an actual event of that type. Multiple detections of the same event were counted at the first time, and the rest hits were taken as false alarms. The F-value metric that summarizes the recall and precision metric is also defined, which is a harmonic average of precision and recall ($F = 2RP/(R + P)$).

From Table 1, our experimental results indicate that the PHHMM based system performs best overall. All the three multimodal multilayer models outperform the systems based on the single modality. They are also better than the HHMM based system using feature fusion, in which the gain is limited, and even slightly worse than the color-only system.

From Table 2, we can see that the multimodal systems, except the PHHMM, always give the high recall rates, while the precision rates are very low. It is caused by over segmentation. The over segmentation caused multiple hits on the same event, and increased the false alarms. The primary reason for the over segmentation is that the primitive detection at the bottom layer is not cooperated with the higher layers very well. If the system only gives emphasis to local detection and does not consider the global constraints from the multilayer integration, the performance will be hurt badly by the over segmentation. Different from those biased models, our preliminary experimental results show that the PHHMM is a viable tool for both modeling the dynamic interaction between different modalities and also preserving the flexibility within the multilayer context.

Table 1. Evaluation results at frame level. HHMM_C is the color-only system; HHMM_M is the motion-only system; other systems use both the features.

Model Type	Accuracy
HHMM_C	78.46
HHMM_M	64.07
HHMM	77.08
FHHMM	81.14
CHHMM	82.60
PHHMM	80.05

Table 2. Evaluation results at segment level.

Model Type	Precision	Recall	F-value
HHMM_C	87.61	71.22	78.57
HHMM_M	50.64	84.40	63.30
HHMM	69.46	84.06	76.07
FHHMM	36.09	100	53.04
CHHMM	36.41	99.30	53.28
PHHMM	73.26	90.00	80.77

6. Conclusions

In this paper, we have presented a generic framework for semantic sports video analysis based on DBNs. Compared with existing works on sports video analysis, the proposed framework provides a unified probabilistic solution for event detection by utilizing both multimodal evidences and multilayer constraints. The multimodal analysis with machine learning techniques leads to more robust and accurate systems in which the evidences from different modalities are fused. At the same time, multilayer analysis based on

DBNs provides a general graphical representation of structural constraints among events, for which a set of efficient statistical inference and learning algorithms are available. In addition, the optimal semantic events are segmented and recognized simultaneously.

Based on this framework, we constructed and compared three multimodal multilayer models, i.e. FHHMM, CHHMM and PHHMM with conventional HHMM on play-break event detection for soccer video. Experimental results indicate that PHHMM is a very attractive choice for semantic sports videos analysis.

The proposed framework using DBNs is open and easy to extend, which may integrate different modalities and domain-specific context constraints. Also, the semantics at different levels are all available from this framework. In the future work, we will further improve this system by 1) using more effective and robust feature representations from different modalities; 2) employing more powerful and discriminative learning methods.

7. References

[1] N. Dimitrova, H.J. Zhang, B. Shahraray, I. Sezan, T. Huang, and A. Zakhor, "Applications of Video-Content Analysis and Retrieval," *IEEE Multimedia*, vol. 9, no. 4, 2002.

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

[3] J. Assfalg, M. Bertini, C. Colombo, and A.D. Bimbo, "Semantic Annotation of Sports Videos," *IEEE Multimedia*, vol. 9, no. 2, 2002.

[4] B. Li and M.I. Sezan, "Semantic Sports Video Analysis: Approach and New Applications," *Proc. IEEE Intl. Conf. on Image Processing*, 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] M. Barnard, J.M. Odobez, and S. Bengio, "Multi-Modal Audio-Visual Event Recognition for Football Analysis," *Proc. IEEE Workshop on Neural Networks for Signal Processing*, 2003.

[7] M. Han, W. Hua, W. Xu, and Y.H. Gong, "An Integrated Baseball Digest System Using Maximum Entropy Method," *Proc. ACM Multimedia*, 2002.

[8] M. Xu, L.Y. Duan, C. Xu, and Q. Tian, "A Fusion Scheme of Visual and Auditory Modalities for Event Detection in Sports Video," *Proc. IEEE Intl. Conf. on Acoustics, Speech, and Signal Processing*, April 2003.

[9] L.Y. Duan, M. Xu, T.S. Chua, Q. Tian, and C.S. Xu, "A Mid-level Representation Framework for Semantic Sports Video Analysis," *Proc. ACM Multimedia*, Nov 2003.

[10] N. Babaguchi, Y. Kawai, and T. Kitahashi, "Event Based Indexing of Broadcasted Sports Video by Intermodal Collaboration", *IEEE Trans. on Multimedia*, vol. 4, no. 1, March 2002.

[11] A. Hanjalic, "Generic Approach to Highlights Extraction from a Sports Video," *Proc. IEEE Intl. Conf. on Image Processing*, 2003.

[12] M. Petkovic, V. Mihajlovic, W. Jonker, and S. Djordjevic-Kajan, "Multi-Modal Extraction of Highlights from TV Formula 1 Programs," *Proc. IEEE Intl. Conf. on Multimedia and Expo*, 2002.

[13] D. Zhong and S.F. Chang, "Structure Analysis of Sports Video Using Domain Models," *Proc. IEEE Intl. Conf. on Multimedia and Expo*, Aug 2001.

[14] L. Xie, S.F. Chang, A. Divakaran and H. Sun, "Structure Analysis of Soccer Video with Hidden Markov Models," *Proc. IEEE Intl. Conf. on Acoustic, Speech, and Signal Processing*, 2002.

[15] L. Xie, S.F. Chang, A. Divakaran, and H. Sun, "Unsupervised Discovery of Multilevel Statistical Video Structures Using Hierarchical Hidden Markov Models," *Proc. IEEE Intl. Conf. on Multimedia and Expo*, 2003.

[16] 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 Intl. Conf. on Image Processing*, 2003.

[17] F. Wang, K.J. Lu, and J. Li, "Automatic Parsing of Sports Videos", *Proc. Joint Conf. on Information Sciences*, North Carolina, Sept 2003.

[18] K. Murphy, *Dynamic Bayesian Network: Representation, Inference and Learning*, Ph.D. Thesis, University of California, Berkeley, 2002.

[19] A.V. Nefian, L. Liang, X. Liu, and K. Murphy, "Dynamic Bayesian Networks for Audio-Visual Speech Recognition," *EURASIP Journal on Applied Signal Processing*, Nov 2002.

[20] Y. Wang, Z. Liu, and J.C. Huang, "Multimedia Content Analysis Using Both Audio and Video Clues," *IEEE Signal Processing Magazine*, 2000.

[21] Y.A. Ivanov, and A.F. Bobick, "Recognition of Visual Activities and Interactions by Stochastic Parsing," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 22, no. 8, Aug 2000.

[22] L.R. Rabiner, "A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition," *Proc. The IEEE*, vol. 77, no. 2, 1989.