

# Camera View-based American Football Video Analysis

Yi Ding and Guoliang Fan \*

School of Electrical and Computer Engineering  
Oklahoma State University, Stillwater, OK, 74078  
{yi.ding, guoliang.fan}@okstate.edu

## Abstract

We present a top-down statistical modeling approach to explore the semantic structure in the American football video. First, a semantic space is defined where the video semantic structure is characterized by semantic units, a dynamic model over semantic units, and an observation model for mapping the semantic units with the visual features. Then, a new hidden Markov model (HMM)-based video generative model is proposed for American football video analysis, where semantic units are defined as latent or hidden states corresponding to four different camera views in the football field. A set of relevant visual features are selected based on the information gain for HMM training and two kinds of state emission function, Gaussian or the Gaussian mixture model (GMM), which characterize the observation density function associated with each latent state, are tested in the proposed HMM for camera view-based video analysis. Experimental results on several real football videos manifest the effectiveness of the proposed algorithm. It is shown that the HMM with GMM emission shows advantages over the Gaussian-based one in terms of the classification accuracy of video shots.

There are mainly two methodologies for content-based video analysis, i.e., *data/feature-driven bottom-up methods* and *concept-driven top-down approaches*. As stated in [3], “a balanced interaction between both methods could be a solution to a generic model for video representation”. In this paper, we propose a new semantic video representation framework and apply it to American football videos, where prior information and high-level knowledge are embedded into a hidden Markov model (HMM)-based video generative model. We argue that the top-down statistical video modeling approaches play an important role in the structured video representation that could be further enriched by feature-driven bottom-up methods for detailed content-based video analysis tasks.

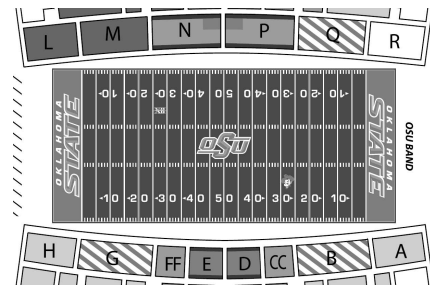


Figure 1. The OSU football field example.

## 1. Introduction

Recently, content-based video analysis has emerged as an prevalent and interesting research topic that is driven by numerous multimedia applications where effective access to video data based on its content is essential [1]. This research is also boosted by the fast development of advanced computation power. Moreover, recent content-based video analysis research is moving toward the discovery of semantic-level structure [2]. However, due to the semantic gap between low-level features and high-level concepts, content-based video analysis remains to be one of most challenging research issues in the field of multimedia.

In broadcast sport video, there are a variety of underlying rules that control the game procedure as well as a certain way of video recording. For example, in Fig. 1, we show a typical football field that covers the play view in most video shots. Usually, cameras are installed around the field to shoot the game <sup>1</sup> Specifically, the camera view reflects the play location in the sport field of each shot, and the camera view transition between two consecutive shots will indicate the types of plays in the game, e.g., advancing the ball or change of possession. The camera view analysis in football video will serve as a case study in this paper.

\*This work is supported in part by the National Science Foundation (NSF) under Grant IIS-0347613 (CAREER).

<sup>1</sup>In this work, we assume that a sport video is composed of a set of play shots and other shots such as breaks and commercials have been removed.

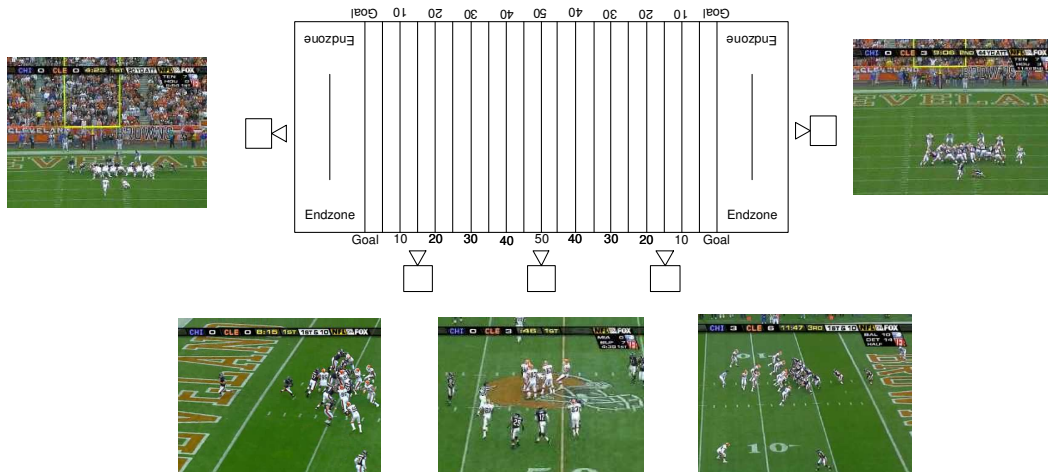


Figure 2. Typical camera setup in a football field and a corresponding shot in each camera.

## 2. Related Work

Sport video analysis is also referred to sport video mining, i.e., the discovery of semantic structures in the video data. There are two major kinds of approaches in this research area: deterministic rule-based (data/feature driven, bottom-up) methods [4, 5, 6] and statistical model-based (concept driven, top-down) methods [7, 8, 9, 10].

Due to plenty of sport rules and a given video recording setup, rule-based methods were proposed to represent semantic events in the sport video. In [4], camera motion related criteria are defined to classify seven different football plays for a given video shot. By evaluating ball trajectories, Yu et. al. [5] proposed an algorithm to perform the play-break segmentation and event detection, such as a goal. The major advantages of rule-based methods are that they are usually goal-oriented and can support specific video search tasks. However, they are not adequate to deal with the uncertainty and ambiguity in the sport video, and their flexibility is limited by the task-specificity.

Another trend of sport video mining is how to use statistical models to discover underlying semantic structures in sport videos, such as play/break or highlights. Descriptive models [11] and generative models [8, 7, 9] are two typical statistical approaches. Compared with descriptive models, generative models are more effective for video mining by involving a dynamic model that is defined on the latent states, i.e., semantic units. In [7], a soccer video is segmented into plays and breaks by involving a Hierarchical HMM (HHMM). The Gaussian mixture model (GMM) was proposed in [8] to find the clusters in a video that correspond to different video scenes or sport plays. In [9], Coupled HMM (CHMM) was used to detect highlights in a sport video. Some interesting semantic structures can be explored by using statistical video models where high-level

knowledge is associated with low-level features. However, the latent states of these models are not directly or explicitly associated with semantic structures in a video, or their semantic meaning has to be interpreted after model training.

In this paper, we propose a novel top-down video representation framework for football video analysis. There are two major contributions. (1) A specific semantic space is defined for football videos where a new HMM-based video generative model is proposed for camera view-based video analysis. (2) Model learning and decoding results not only indicate the play location of each shot, but also may reveal play types via the state transition across multiple shots. This work can be extended to the HHMM architecture or combined with rule-based methods to explore more detailed semantic structures within each shot or between shots. To our best knowledge, few previous work directly and explicitly defines latent states as semantic events in a video.

## 3. Problem Formulation

### 3.1. Semantic Space

In order to bridge the semantic gap for sport video mining, we first construct a semantic space for video representation, as shown in Fig. 3, where the semantic units and their dynamic relationship are defined to characterize the semantic structure in a video. In general, semantic units represent basic entities to represent semantic structures, such as “play” and “break” in a sport video, and the dynamic relationship between them can be characterized by a dynamic model that governs the transition among semantic units. From the previous analysis, we argue that a clear specification of the semantic space for video modeling is very helpful for us to represent and capture the semantic structure existing in the video data.

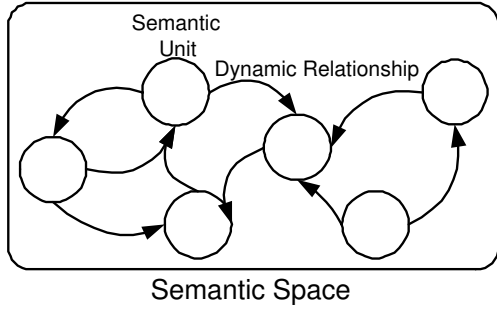


Figure 3. Semantic space.

Given a football field as shown in Fig. 2, most plays we watch in a broadcast football video are captured by five cameras fixed around the field. During the play time, each play or shot is usually captured by only one camera, and the camera in duty is changing from time to time according to the location of the play in a game. Then naturally we can set up a semantic space for football videos that consists of a set of camera views and their transition relationship. Specifically, semantic units in this semantic space are directly related to four camera views, i.e., the left view, the central view, the right view and the end zone view (two end zone views are combined together). In the following section, we will show that the dynamic relationship between semantic units, i.e., the camera views, in this semantic space can be represented by a first-order Markov chain.

### 3.2. A New Video Generative Model

Usually, a generative model involves two basic elements, the latent/hidden states and the observations. Also, there should be a certain dynamic model to govern the state transition and an observation/emission model to characterize the conditional density function of each latent state. Therefore, the key to develop an informative video generative model is to identify the latent states and their associated observations. According to the semantic space we defined before, we can naturally assume the camera view to be the latent state in the generative model. Moreover, shots captured by different cameras exhibit distinct visual effect, such as yard lines and color distributions. Therefore, we consider relevant visual features to be extracted from video shots/frames as the observations, as shown in Fig. 4.

There are four latent states corresponding to four camera views (two end-zone cameras are combined into one camera view), i.e.,  $\{S_t = k, |t = 1, \dots, T; k = 1, 2, 3, 4\}$ . By the Markov assumption, we can have a state transition matrix, i.e.,  $\{a_{kj} = p(S_t = j | S_{t-1} = k) | k, j = 1, 2, 3, 4\}$ , to govern the state dynamics over time, as shown in Fig. 5.

Based on the proposed generative model, the video observations  $\{\mathbf{O}_t | t = 1, \dots, T\}$ , i.e., relevant visual features

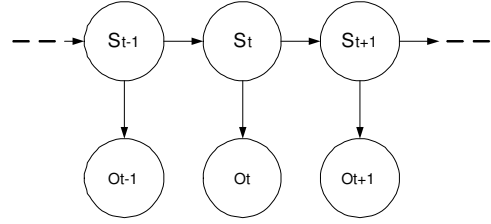


Figure 4. A new video generative model.

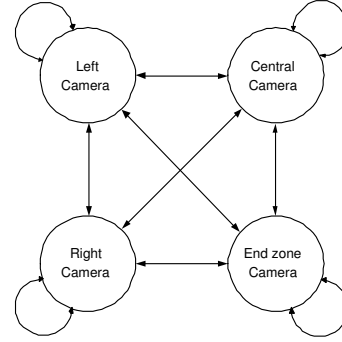


Figure 5. State dynamics of camera views.

extracted shots/frames, are assumed to be drawn from certain emission functions associated with the latent states. The often used emission functions include Gaussian or the Gaussian mixture model (GMM). Then we can define two kinds of observation models as

$$p(\mathbf{O}_t | S_t = k) = \mathcal{N}(\mathbf{O}_t | \mu_k, \Sigma_k) \quad (1)$$

where  $\mu_k$  and  $\sigma_k$  are the mean and variance of the Gaussian for latent state  $k$ , or

$$p(\mathbf{O}_t | S_t = k) = \sum_{n=1}^N (\alpha_n \mathcal{N}(\mathbf{O}_t, \mu_{nk}, \Sigma_{nk})), \quad (2)$$

where  $\{\alpha_n, \mu_{nk}, \Sigma_{nk} | n = 1, \dots, N\}$  parameterizes the  $N$ -order GMM for latent state  $k$  and  $\sum_{n=1}^N \alpha_n = 1$ .

Now, two key issues are left. (1) Can we find salient visual features from video shots/frames that can reflect the distinction between latent states? (2) Can we use aforementioned emission functions to characterize their densities?

### 3.3. Characteristics of Visual Features

When watching American football, the viewer can easily identify the camera view of each shot from the yard numbers along the side line of the field. However, it is very hard for a computer vision algorithm to automatically detect and recognize the yard numbers from the video shot due to two facts: (1) the locations of yard numbers are arbitrary in a

frame or may not even appear in many frames and (2) the significant camera motion and object occlusion make number recognition very difficult. Therefore, we need more robust and accessible visual features to characterize the latent states of the proposed generative video model. In this research, we identify the spatial color distribution and the angle of yard lines to be relevant features. For example, in the center field, there is usually a logo of the host team in the field center that has a strong contrast with the green color of the play ground, and all yard lines are almost vertically oriented. In the following, we will first discuss how to extract salient visual features, including the color distribution and the angle of yard lines, and then we show the correlation between the camera views and the extracted features.

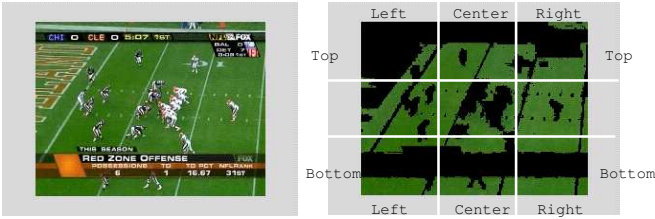


Figure 6. Play ground detection result.

In this work, we use the concept of dominant color and estimate the spatial color distribution by dividing a frame image into multiple regions and computing the ratios of dominant color between different regions. Specifically, we use the Robust Dominant Color Region Detection algorithm in [12] to extract the dominant color region, i.e., the play ground. Fig. 6 shows some results of dominant color extraction. To obtain yard lines, we can simply use Canny edge detection and the Hough transform to detect the yard lines in the region of play ground as shown in Fig. 7.

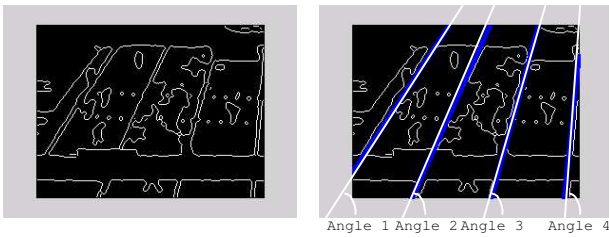


Figure 7. Yard line detection result.

### 3.4 Feature Selection and Extraction

In order to estimate the color distribution, we segment a frame into five regions as shown in Fig. 6.  $W$  is the total pixel number in a frame, and  $W_f$  is the total pixel number in

the dominant color region. Moreover,  $W_{left}$ ,  $W_{right}$ ,  $W_{top}$ ,  $W_{bottom}$ ,  $W_{center}$  are the pixel numbers of the dominant color region in the left part, the right part, the top part, the bottom part, and the central part, respectively. As the the angles of yard lines, we define  $\theta_l$  as the angle of the  $l$ th detected line and  $L$  is the number of detected lines as shown in Fig. 7.  $\theta_{last}$  as the average angle in the last frame in a shot,  $\theta_{first}$  as the average angle in the first frame in a shot. Then, based on the detected play ground region and yard lines, we define six relevant features as follows:

- Ratio of the dominant color region:  $R_f = W_f/W$ ;
- Ratio difference of dominant color between the left/right and center parts:  $R_d = (|W_{left} - W_{center}| + |W_{right} - W_{center}|)/W$ ;
- Ratio difference of dominant color between the left and right parts:  $R_{lr} = (W_{left} - W_{right})/W$ ;
- Ratio difference of dominant color between the top and bottom parts:  $R_{tb} = (W_{top} - W_{bottom})/W$ ;
- Average angle of all yard lines:  $D_{ave} = \sum_{l=1}^L \theta_l/L$ ;
- Angle difference between the first and last frames in a shot:  $D_{shot} = \theta_{last} - \theta_{first}$ .

We compute shot-wise features by averaging all frame-wise ones in a shot. As two examples, we demonstrate the distributions of  $R_{lf}$  and  $D_{ave}$  with respect to four camera views in Fig. 8. There are two observations: (1) both  $R_{lf}$  and  $D_{ave}$  are salient features to distinguish between four camera views; (2) it seems that the GMM-emission model defined in (2) is appropriate for latent state modeling. We have a quantitative evaluation for all six features in next section.

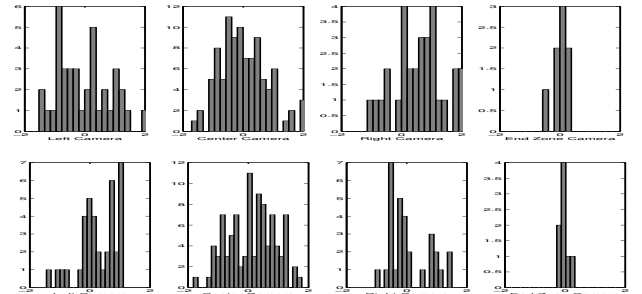


Figure 8. Distributions of  $R_{lf}$  (top) and  $D_{ave}$  (bottom) with respect to four camera views.

### 3.5 HMM-based Video Representation

The HMM is well-suited to the proposed video generative model, where the latent states are just the hidden states

in the HMM and the state emission function should be the GMM for each state. Given a series of video observations  $\{\mathbf{O}_t|t = 1, \dots, T\}$ , the HMM be parameterized by  $\Gamma = \{\mathbf{S}, \pi_k, a_{k,j}, \alpha_n, \mu_{nk}, \Sigma_{nk}|k = 1, \dots, 4; n = 1, \dots, N\}$ :

1.  $\mathbf{S} = \{1, 2, 3, 4\}$ : the hidden states represent the left, central, right, and end-zone camera views.
2.  $\pi_k = p(S_1 = k)$  and  $k \in \mathbf{S}$ : initial state probabilities vector, which represents the initial probabilities of four cameras;
3.  $A = \{a_{kj} = p(S_t = j|S_{t-1} = k)|t = 1, \dots, T; k, j \in \mathbf{S}\}$ : are the state (camera) transition probabilities between states (cameras)  $i$  and  $j$ ;
4.  $p(\mathbf{O}_t|S_t = k) = \sum_{n=1}^N (\alpha_n \mathcal{N}(\mathbf{O}_t, \mu_{nk}, \Sigma_{nk}))$ : is the  $N$ -order GMM emission function of hidden states.

Then we can use the Expectation Maximization (EM) algorithm [13] to obtain the maximum likelihood-based parameter estimate of this HMM as follows:

$$\Gamma^* = \arg_{\Gamma} \max P(\mathbf{O}_{1:T}|\Gamma), \quad (3)$$

where

$$P(\mathbf{O}_{1:T}|\Gamma) = \sum_{S_{1:T}} \pi_{S_1} \prod_{t=1}^T p(S_t|S_{t-1})p(\mathbf{O}_t|S_t = k).$$

After the EM training, we can use the Viterbi decoding algorithm to estimate the optimal state sequence  $S_{1:T}^*$ ,

$$S_{1:T}^* = \arg_{S_{1:T}} \max P(S_{1:T}|\mathbf{O}_{1:T}, \Gamma^*), \quad (4)$$

which corresponds to the camera view classification results for all video shots  $\mathbf{O}_{1:T}$ . On the one hand,  $S_{1:T}^*$  indicates the location of each play. On the other hand, the state (camera) transition between two adjacent shots may indicate certain play, such as advancing the ball or change of possession. Thus the proposed HMM-based video representation framework can explicitly explore semantic structures in American football videos.

In practice, we use K-mean clustering to initialize the emission functions of the HMM prior to EM training, and  $N = 3$  was found to be a reasonable choice for the GMM emission function. For the prior probability  $\pi_k$ , we assume the football game always starts from the central camera ( $S_1^* = 2$ ), i.e.,  $\pi_2 = 1$  and  $\pi_1 = \pi_3 = \pi_4 = 0$ . We also initialize the state transition matrix  $A$  that is an important parameter in the HMM by computing the frequency of actual state transitions and averaging over a few games.

## 4. Experimental Results

We will first address the issue of feature evaluation and then discuss the HMM-based shot classification results.

### 4.1. Feature evaluation

To quantitatively evaluate the saliency of six features, we have manually performed camera view-based shot classification for one football video whose ground truth state sequence is denoted as  $S_{1:T}^g$ , then we employ the information gain [14, 7] to evaluate feature's contribution to HMM-based classification. Given feature  $z \in 1, 2, \dots, 6$ ,  $\Gamma_z$  is learned and the optimal state sequence  $S_{1:T}^z$  can be obtained. The information gain is computed as

$$Info(S_{1:T}^g|S_{1:T}^z) = H(P_{S^z}) - \sum_j P_{S^g} \cdot H(P_{S^z|S^g=j}) \quad (5)$$

where  $H$  is the entropy function, and

$$P_{S^g}(k) = \frac{\#(S_t^g = k|t = 1, \dots, T)}{T},$$

$$P_{S^z}(k) = \frac{\#(S_t^z = k|t = 1, \dots, T)}{T},$$

$$P_{S^z|S^g}(k, j) = \frac{\#(S_t^z = k \text{ and } S_t^g = j|t = 1, \dots, T)}{\#(S_t^g = k, t = 1, \dots, T)},$$

where  $\#$  is the counter. After ranking six features according to  $Info(S_{1:T}^g|S_{1:T}^z)$  in Table. 1, we select the top four features to construct the observation vector of the HMM, i.e.,  $\mathbf{O}_t$ , including  $D_{ave}$ ,  $R_f$ ,  $D_{shot}$ , and  $R_{lr}$ .

Index	$z = 1$	$z = 2$	$z = 3$	$z = 4$	$z = 5$	$z = 6$
Feature	$D_{ave}$	$R_f$	$D_{shot}$	$R_{lr}$	$R_d$	$R_{tb}$
I.G.	0.62	0.41	0.37	0.35	0.16	0.11

**Table 1. Rank of information gain(I.G.)**

### 4.2. Shots classification and discussion

Based on these selected features, we tested our algorithm on the two 30-minute football videos. Video A contains 156 shots, and Video B 171 shots. After K-mean initialization, we use the EM algorithm for HMM training with two different emissions: Gaussian (HMM-G) and GMM (HMM-GMM). Finally, we implement Viterbi decoding to obtain the optimal state sequence. The experimental results are shown as Table. 2. We can see the performance of HMM-GMM is much better than that of HMM-G. It is mainly because the GMM can better characterize the densities of visual features than the Gaussian.

In Table. 3, we also show the classification result concerning each camera based on HMM-GMM. We find that most errors are from the misclassification of left or right camera views as the central camera view. This is mainly because the central camera usually cover the major part of

Data	K-mean	HMM-G	HMM-GMM
Video A	41.03 (64/156)	55.77 (87/156)	72.44 (113/156)
Video B	41.52 (71/171)	62.57 (107/171)	75.44 (129/171)

**Table 2. Shots classification results**

the field, and the number of shots captured by the central camera is usually more than by others. To improve those shortcomings of our current approach, we are exploring the frame-wise features and the temporal dependency cross frames in a short.

Data	Left	Central	Right	End zone
Video A	72.5 (29/40)	73.4 (67/94)	61.54 (8/13)	100 (9/9)
Video B	65.79 (25/38)	78.95 (75/95)	70 (21/30)	100 (8/8)

**Table 3. Single shot classification results**

## 5. Conclusion and future work

We have presented a top-down HMM-based video representation framework for football video analysis. The major contribution of this work is a new video generative model that is defined on a semantic space specified for football videos. Specifically, the latent states in the HMM are associated with the camera views in the football field, and salient visual features extracted from video data are considered as the observations of latent states. Via this generative model, all play shots can be classified into four groups according to their camera views. The preliminary results indicate that the proposed model is very promising in finding the underlying semantic structure in the broadcast football video. It is also computationally efficient because only a low-dimensional feature vector is involved for HMM training.

However, we find the classification result cannot be further improved when we utilize the richer frame-based features for HMM training and decoding. It is likely due to the reason that the strong temporal dependence across frames was not considered in a shot. Therefore, we are working on the development of a multi-level HMM where two-layer latent states capture different semantic units in a football game. Specifically, the first layer is about camera views of video shots and the second layer is the orientation of the camera. We expect that the video generative model which is amenable to a well defined semantic space is important for effective and informative video mining.

## References

- [1] P. Aigrain, H.J. Zhang, and D. Petkovic, "Content-based representation and retrieval of visual media: A state-of-the-art review," *Multimedia Tools and Applications*, vol. 3, no. 3, pp. 179–202, 1996.
- [2] S.-F. Chang, "The holy grail of content-based media analysis," *IEEE Multimedia Magazine*, vol. 9, no. 2, pp. 6–10, 2002.
- [3] J. Calic, N. Campbell, S. Dasiopoulou, and Y. Kompatsiaris, "An overview of multimodal video representation for semantic analysis," in *Proc. of IEE European Workshop on the Integration of Knowledge, Semantics and Digital Media Technologies (EWIMT 2005)*, December 2005.
- [4] M. Lazarescu and S. Venkatesh, "Using camera motion to identify types of american football plays," in *Proc. 2003 International Conference on Multimedia and Expo*, July 2003, pp. II–181–4.
- [5] X. Yu et al, "Trajectory-based ball detection and tracking with applications to semantic analysis of broadcast soccer video," in *Proc. eleventh ACM international conference on Multimedia*, 2003, pp. 11–20.
- [6] T.-Y. Liu, W.-Y. Ma, and H.-J. Zhang, "Effective feature extraction for play detection in american football video," in *Proc. 11th International Multimedia Modelling Conference*, 2005, pp. 164–171.
- [7] L. Xie, S. Chang, A. Divakaran, and H. Sun, "Unsupervised mining of staistical temporal structures in video," in *Video Mining*, D. Doremann A. Rosenfeld, Ed., chapter 10. Kluwer Academi Publishers, 2003.
- [8] Y.-P. Tan and H. Lu, "Model-based clustering and analysis of video scenes," in *Proc. IEEE International Conference on Image Processing*, 2002, vol. 1, pp. 22–25.
- [9] Z. Xiong, "Audio-visual sports highlights extraction using coupled hidden markov models," *Pattern Analysis & Applications*, vol. 8, no. 1, pp. 62–71, 2005.
- [10] F. Wang, Y.F. Ma, H.J. Zhang, and J.T. Li, "A novel moving object segmentation algorithm based on spatiotemporal Markov random field," in *Image Processing, 2004. ICIP '04. 2004 International Conference on*, 2004, pp. 633–636.
- [11] S. C. Zhu, "Statistical modeling and conceptualization of visual patterns," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 25, no. 6, pp. 691–712, 2003.
- [12] A. Ekin, A. Tekalp, and R. Mehrotra, "Automatic soccer video analysis and summarization," *IEEE Trans. on Image Processing*, vol. 12, no. 7, pp. 796–807, 2003.
- [13] J. Bilmes, "A gentle tutorial on the EM algorithm and its application to parameter estimation for Gaussian mixture and hidden Markov models," *ICSI-Technical Report-97-021*, 1997.
- [14] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, John Wiley and Sons, 1991.