

# Sports Video Mining via Multi-channel Segmental Hidden Markov Models

Yi Ding, *Student Member, IEEE*, Guoliang Fan, *Senior Member, IEEE*

**Abstract**—We study sports video mining as a machine learning and statistical inference problem. We focus on mid-level semantic structures that can serve as building blocks for high-level semantic analysis. Particularly, we are interested in how to infer multiple co-existent structures jointly. We present a new multi-channel segmental hidden Markov model (MCSHMM) that is a unique probabilistic graphical model with two advantages. One is the integration of both hierarchical and parallel dynamic structures that offers more flexibility and capacity of capturing the interaction between multiple Markov chains. The other is the incorporation of the segmental HMM (SHMM) to deal with variable-length observations. In addition, we develop a maximum *a posteriori* (MAP) estimator to optimize the model structure and parameters simultaneously. The proposed MCSHMM is used for American football video analysis. The experiment result shows that the MCSHMM outperforms existing HMMs and has potential to be extended for other video mining tasks.

**Index Terms**—Video mining, hidden Markov models, semantic structures, sports video analysis

## I. INTRODUCTION

Driven by the ever increasing needs of numerous multimedia and online database applications, there is a growing interest in video mining research. The goal of video mining is to discover knowledge, patterns, and events called *semantic structures* in the video data to facilitate the user's data access [1], [2]. There are three main research issues for video mining, i.e., summarization/abstraction [3], browsing/skimming [4], and indexing/retrieval [5], [6]. The deliverables are similar to the two tools in a book, i.e., *Table of Contents* (TOC) and *Index*, which provide an overview and a quick method for a reader to find information of interest. Due to different production and editing styles, videos can be classified into two categories: the scripted and the non-scripted [7]. Usually, they are associated with different video mining tasks. *Scripted videos* are produced or edited according to a pre-defined script or plan. Usually, news and movies are highly scripted videos that are composed of pre-defined segments or episodes. Building a TOC [8] is more suitable for the scripted videos that is able to support efficient browsing or indexing. On the other hand, the events in *non-scripted videos* happen spontaneously and usually in a relatively fixed setting, such as meeting, sports, and surveillance videos. How to detect the highlights or events of interests is very useful for non-scripted videos.

Sports video analysis has been widely studied due to its great commercial value [9], [10], [11], [4]. Although sports videos (non-edited) are considered as non-scripted, they usually have a relatively well-defined structure (such as the field scene) or repetitive patterns (such as a certain play type), which could help us to enhance the *scriptedness* of sports videos for more versatile and flexible access. In general, there are two kinds of video mining approaches: the *event-based* [12], [13] and *structure-based* [9], [14], [15]. The event-based approaches detect the events-of-interest or highlights, e.g., goals in a soccer game, which improve the semantic understanding of the video content. However, this kind of approach is usually task-dependent and for a specific purpose. In contrast, the structure-based approaches can parse a long video sequence into individual segments, e.g., the play/break, which can be used as a mid-level video content representation. Although these approaches provide limited semantics, they can facilitate further high-level semantic analysis. Due to their complementary nature, researchers have proposed to integrate two kinds of approach in one video mining framework. For example, a mid-level representation framework was proposed for semantic sports video analysis where both temporal structures and event hierarchy were involved [11]; a mosaic-based generic scene representation was developed from video shots and used to mine both events and structures [1].

Machine learning is considered as one of the most effective approaches for semantic video analysis [4]. Particularly, Hidden Markov models (HMMs) have been the most popular one. In this work, our goal is to establish a general mid-level video representation that has the capability of both event and structure analysis, and supports a low-level to high-level video content description. Our approach involves *explicit semantic modeling* and *direct semantic computing* that are effective to bridge the semantic gap. Specifically, we are interested in exploring the interaction between multiple co-existent semantic structures, i.e., the *play type* (what happened?) and *camera view* (where did it happen?). They are the two common structures in most field-based sports and useful for high-level semantic analysis. We advance a new multi-channel segmental HMM (MCSHMM) to explore the two semantic structures jointly. We also address simultaneous structure learning and parameter learning to fully take advantage of its potential. The proposed MCSHMM demonstrates a new way of capturing and utilizing the coupling effect between two Markov chains with different state space configurations. It has better flexibility and capability than the existing HMMs, and can be applied to other time series analysis applications involving multiple data streams of different semantics.

Y. Ding and G. Fan are with the School of Electrical and Computer Engineering, Oklahoma State University, Stillwater, OK, 74078 USA e-mail: {yi.ding,guoliang.fan}@okstate.edu. This paper was presented in part in the 2008 ACM Multimedia Conference, Vancouver, Canada, Oct. 2008. This work is supported by the National Science Foundation (NSF) under Grant IIS-0347613 and the 2009 Oklahoma NASA EPSCoR Research Initiation Grant.

## II. PREVIOUS WORKS

### A. Machine Learning for Video Mining

The machine learning approaches to video mining are generally classified into two categories. One is the discriminative approach that involves a parametric model to represent the posterior probabilities of hidden states given the observations. The other is the generative approach that estimates posterior probabilities of hidden states by involving conditional densities (likelihood) for observations given some state priors.

The discriminative approaches, like the conditional random field (CRF) [16] and the support vector machines (SVM) [17], tend to directly estimate the posterior probability of latent states given an observation sequence. They are expressive due to their flexibility of statistical modeling, and appreciated by their simplicity and efficiency in training and inference. However, they may be sensitive to noise or limited in some cases where a large and complex data set is involved. Generative approaches are usually preferred for a large data set due to their better generality, and they rely on a dynamic model on the state and an observation likelihood function. As one kind of generative model, the Dynamic Bayesian Network (DBN) provides a unified probabilistic framework to represent various graphical structures and the HMM is the simplest DBN that has been widely used in many video analysis applications.

Various extended HMMs were proposed that aim to improve the *dynamic model* and *observation model* [18]. There are two kinds structures of dynamic models, the *parallel* and *hierarchical* ones. The former one is usually needed when multi-source observations are involved or the information fusion is needed at either the feature or decision level, e.g., the Factorial HMM [19], and the CHMM [20]. The latter one is another important mechanism that was inspired by the multi-scale structure in many natural sequences [21], like the Hierarchical HMM [22]. Additionally, the observation model (or emission) also plays an important role in a HMM. Traditional HMMs normally represent each state by a single observation either by a Gaussian or Gaussian Mixture Model (GMM). However, in some cases, a state may emit a variable-length observation sequence (such as a video shot of multiple frames). An interesting segmental HMM was proposed in [23] where a segmental model was used to characterize the variable-length observations for speech signals.

Model learning is another important issue, especially for a HMM with a complex state space or variable-length observations. Generally, there are two learning issues, *structure learning* [24] and *parameter learning* [25]. The former one tries to obtain a compact state space and condensed state dynamics. The latter one focuses on the optimal model parameters given a certain structure. It is imperative to have simultaneous structure learning and parameter learning for complex DBNs/HMMs to ensure their effectiveness and efficiency in practice. In [26], the concepts of entropic prior and parameter extinction were proposed to simplify and optimize the model structure by a state trimming process. Another example is the reverse-jump Markov Chain Monte Carlo (RJMCMC) that was used to learn the hierarchical HMMs (HHMMs) for unsupervised sports video mining in [22].

### B. Our Previous Research

Our previous research [27], [28], [29] is briefly reviewed in three aspects: problem formulation, feature extraction, and machine learning.

1) *Problem Formulation*: The main challenge of video mining is the *semantic gap* between semantic concepts and low-level features. We believe that explicit semantic modeling and direct semantic computing are effective to bridge the semantic gap for sports video mining. Therefore, we define a set of mid-level semantic structures as rudimentary semantic building blocks. These building blocks should be frequent, repeatable and relatively well-defined, and their states can be governed by certain dynamics, e.g., *camera views* and *play types* that are two common structures in most field-based sports. For example, in American football videos, we can define four camera views (central, left, right, end-zone views), and four play types (the long play, short play, kick, and field goal play). Other structures (e.g., possession) are also possible. A combination of several semantic structures can specify some high-level semantics. For example, a touchdown highlight can be represented by a long play to the end-zone followed by a field goal. This approach supports not only highlight detection but also customized events-of-interest, increasing the usability and interactivity of video data.

2) *Features Extraction*: We are mainly interested in two semantic structures, i.e., the camera view and play type, in this work. Specifically, we used the color distribution and the yard line angle to characterize the camera view [29]. We adopted the concept of dominant color [30] to estimate the spatial color distribution, and Canny edge detection followed with the Hough transform to detect the yard lines and to compute their angles. Also, there are two main types of camera motion, i.e., panning and tilting, which effectively characterize different play types. For example, a long play is usually associated with a strong panning effect, while a short play is normally reflected by a weak panning effect. We chose the optical flow based method [31] to compute two kinds of camera motion between two adjacent frames. Also, the frame indices are included as a temporal feature. By using these features, video mining is formulated as an inference problem where we want to infer the *mid-level semantic structures* from the *visual features*.

3) *Statistical Inference*: The HMMs naturally fit in our problem formulation where the hidden states represent mid-level semantic structure (camera views or play types) defined at the shot level and the observation is the visual features extracted at the frame level. However, traditional HMMs can hardly handle variable-length observations, and an easy treatment is to average frame observations in a shot to compute the shot-wise feature [27]. To take advantage of the rich statistics of frame-wise features in a shot, we proposed an embedded HMM-GMM in [28] that was inspired by the generative models proposed [32]. This model imposes a virtual observation layer represented a GMM. During data generation, a GMM is first drawn for a state defined on a shot, from which a set of frame-wise observations are generated. This two-layer observation model can cope with the variable-length observation. It shows promising results for play-based and view-based shot classification than traditional HMMs.

### III. SEGMENTAL HMM-BASED APPROACH

The embedded HMM-GMM ignores the temporal dependency across frames in a shot. Therefore, we adopted the segmental HMM (SHMM) to handle variable-length observations in a more systematic way [29]. The SHMM was first introduced in [23] for speech recognition that involves a segmental observation model, as shown in Fig. 1. Each state in the SHMM can emit a variable-length observation sequence called a segment. It is assumed that all observations are conditionally independent given the mean of that segment, leading to a closed-form likelihood function.

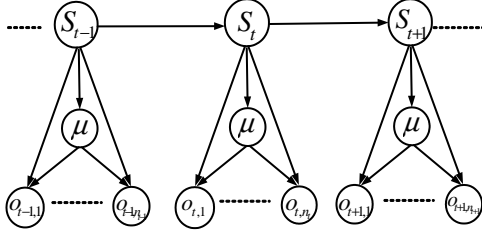


Fig. 1. The SHMM proposed in [23].

A SHMM of  $M$  states can be characterized by

$$\Theta^* = \arg \max_{\Theta} P(\mathbf{O}_{1:T} | \Theta), \quad (1)$$

where  $\Theta = \{\pi_m, a_{m,n}, \mu_{\mu,m}, \Sigma_{\mu,m} | m, n = 1, \dots, M\}$ .  $\pi_m$  is the initial probability, and  $a_{m,n}$  the transition probability.  $\mu_{\mu,m}$  and  $\Sigma_{\mu,m}$  characterize the mean of the segment of state  $m$ , and  $\Sigma_{\mu,m}$  the variance of the segment. Given a shot in time  $t$  with  $n_t$  observations, i.e.,  $\mathbf{O}_t = \{O_{t,k} | k = 1, \dots, n_t\}$ , the conditional likelihood of  $\mathbf{O}_t$  is defined as:

$$p(\mathbf{O}_t | S_t = m, \Theta) = \int p(\mu | S_t = m, \Theta) p(O_{t,1:n_t} | \mu, S_t = m, \Theta) d\mu, \quad (2)$$

Under the conditionally independent assumption, we have:

$$p(\mathbf{O}_t | S_t = m, \Theta) = \int p(\mu | S_t = m, \Theta) \prod_{k=1}^{n_t} p(O_{t,k} | \mu, S_t = m, \Theta) d\mu, \quad (3)$$

where  $p(O_{t,k} | \mu, S_t = m, \Theta) = \mathcal{N}(O_{t,k} | \mu, \Sigma_m)$  and  $p(\mu | S_t = m, \Theta) = \mathcal{N}(\mu | \mu_{\mu,m}, \Sigma_{\mu,m})$ . Compared with traditional HMMs, the new term is the Gaussian prior of mean  $\mu$  that was originally introduced to capture the characteristics of a segment in a speech signal, i.e., the phone or model level, which are variable due to different speakers or stress conditions [23]. If we set  $\Sigma_{\mu,m} = 0$ , (3) can be reduced to:

$$p(\mathbf{O}_t | S_t = m, \Theta) = \prod_{k=1}^{n_t} p(O_{t,k} | S_t = m, \Theta), \quad (4)$$

where  $p(O_{t,k} | S_t = m, \Theta) = \mathcal{N}(O_{t,k} | \mu_{\mu,m}, \Sigma_m)$  that is equivalent to the Gaussian emission. Then it is similar to our previously proposed embedded HMM-GMM [28] that involves a two-layer observation model. In practice, the Gaussian prior of  $\mu$  was found useful to capture the temporal dependency of all frame-wise observations in a shot, as shown by the significant improvement of SHMM over the traditional HMMs on play-based and view-based shot classification [29].

### IV. PROPOSED MULTI-CHANNEL SHMM

All HMM-based approaches we discussed so far can only detect one semantic structure at a time. Usually, a sports video contains multiple mid-level semantic structures in parallel, such as play types and camera views. More interestingly, there are some interactions among them. Specifically, it was observed that camera views and play types are quite related to each other during the game. For example, after a shot of the central view accompanied by a short play, the camera view in the next shot is likely to remain in the same view; while if it is a long play, the next camera view might be switched to the other camera views, either right or left camera view. In this work, we are interested in exploring the interaction between multiple semantic structures with the aim to improve the overall mining performance.

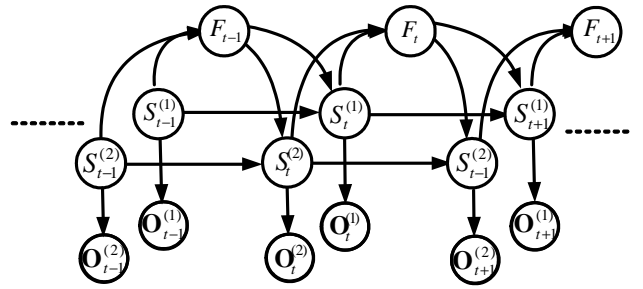


Fig. 2. In the MCSHMM, the first layer includes two SHMMs, and the second layer captures the interaction between two channels.

It is tempting to invoke the Coupled HMM (CHMM) in [20] that can support multiple Markov channels simultaneously. However, the CHMM usually assumes relative strong interaction between two Markov chains that may overweight or even corrupt the Markovian property within each chain if the assumption is not true, as observed in our practice. Therefore, we advance a new multi-channel SHMM model (MCSHMM) that involves two *parallel* SHMMs and a two-layer *hierarchical* Markovian structure, as shown in Fig. 2. In the MCSHMM, on the one hand, two SHMMs are equipped by the powerful segmental observation model to mine individual semantic structures, and on the other hand, the MCSHMM can explore multiple semantic structures with a hybrid parallel-hierarchical dynamic model.

#### A. Model Specification

The unique feature of the MCSHMM is the integration of both parallel and hierarchical structures in the dynamic model and the incorporation of a segmental observation model, which allow the MCSHMM have more flexibility, capacity, and functionality than the CHMM, SHMM and HHMM. In the view of generative models, both the dynamic model (among hidden states) and the observation model in the MCSHMM have a two-layered structure that greatly enhance its capability of learning and inference. Specifically, at the first-layer of the dynamic model,  $\mathbf{S} = \{S_t^{(j)} | t = 1, \dots, T; j = 1, 2\}$  denotes the state sequence of two channels where  $S_t^{(j)}$  denotes the state of shot  $t$  in channel  $j$ , and at the second-layer of the dynamic

model,  $\mathbf{F} = \{F_t = (S_t^{(1)}, S_t^{(2)}) | t = 1, \dots, T\}$  represents the state sequence at the second layer where each state consists of two current states at the first layer. At the observation layer  $\mathbf{O}_t^{(j)} = \{O_{1:n_t}^{(j)} | t = 1, \dots, T; j = 1, 2\}$  indicates observations of shot  $t$  with  $n_t$  frames in channel  $j$ . In this research, all shots are pre-segmented, and the  $n_t$  is known.

Therefore, the MCSHMM's parameter set  $\Theta$  includes following components  $\Theta = \{\Pi, \mathbf{A}, \Omega\}$ :

- Initial probabilities:

$$\Pi = \{P(S_1^{(1)}), P(S_1^{(2)}), P(F_1 | S_1^{(1)}, S_1^{(2)})\}; \quad (5)$$

- Transition probabilities:

$$\mathbf{A} = \{A_w, w = 1, 2, 3\}, \quad (6)$$

where

$$A_1 = \{P(S_t^{(1)} = m | S_{t-1}^{(1)} = n, F_{t-1} = l)\}, \quad (7)$$

$$A_2 = \{P(S_t^{(2)} = m | S_{t-1}^{(2)} = n, F_{t-1} = l)\}, \quad (8)$$

$$A_3 = \{P(F_t = l | S_t^{(1)} = m, S_t^{(2)} = n)\}; \quad (9)$$

- Observation density functions:

$$p(\mathbf{O}_t^{(j)} | S_t^{(j)} = m, \Omega) = \int \mathcal{N}(\mu | \mu_{\mu,m}^{(j)}, \Sigma_{\mu,m}^{(j)}) \prod_{k=1}^{n_t} \mathcal{N}(O_{t,k}^{(j)} | \mu, \Sigma_m^{(j)}) d\mu, \quad (10)$$

where  $\Omega = \{\mu_{\mu,m}^{(j)}, \Sigma_{\mu,m}^{(j)}, \Sigma_m^{(j)} | j = 1, 2; m = 1, \dots, N^{(j)}\}$  specifies the two segmental models,  $N^{(j)}$  is the number of states in channel  $j$ , in this paper,  $N^{(1)} = N^{(2)} = 4$ , and  $O_{t,k}^{(j)}$  denotes the observation of frame  $k$  in shot  $t$  and channel  $j$ . The Gaussian prior of mean  $\mu$  is used to capture temporal dependency among observations in a shot by assuming conditional independency under a given  $\mu$ . In addition, the closed-form likelihood function well represents the variable-length observations in different shots. However, the integration in (19) has to be approximated during EM learning [23].

Given the dual-channel observations of  $T$  shots,  $\mathbf{O} = \{\mathbf{O}_t^{(j)} | t = 1, \dots, T, j = 1, 2\}$ , the joint likelihood is defined as:

$$p(\mathbf{S}, \mathbf{F}, \mathbf{O} | \Theta) = P(S_1^{(1)})P(S_1^{(2)})P(F_1 | S_1^{(1)}, S_1^{(2)}) \prod_{t=1}^T P(F_t | S_t^{(1)}, S_t^{(2)}) \prod_{t=2}^T \prod_{j=1}^2 P(S_t^{(j)} | S_{t-1}^{(j)}, F_{t-1}) \prod_{t=1}^T \prod_{j=1}^2 p(\mathbf{O}_t^{(j)} | S_t^{(j)}). \quad (11)$$

The Expectation Maximization (EM) algorithm is often used for learning a DBN/HMM, such as the MCSHMM, which finds the optimal parameters by maximizing the likelihood function (e.g., (11)) through iterative Expectation (the E-step) and Maximization (the M-step).

However, the direct maximum likelihood (ML) learning of the MCSHMM could be problematic due to the complex state space specified by  $\mathbf{S}$  and  $\mathbf{F}$  that leads to a large set of model parameters, such as  $\mathbf{A}$  given in (6). In addition, we expect

that the useful state space could be much smaller than the one covering all possible cases. Therefore, we are hoping to find an effective learning algorithm that can optimize the parameter set as well as the model structure (the state space and state dynamics) simultaneously.

### B. Learning of the MCSHMM

As we mentioned before, there are two aspects in model learning, *structure learning* and *parameter learning*. The former one aims at finding a compact and effective model structure by simplifying the state space and reducing the parameter set, and the latter one tries to optimize the model parameters given a pre-defined model structure. More advancingly, we will propose a new unsupervised learning algorithm in which two learning processes could be unified into one framework where the model structure and parameters can be optimized simultaneously [22], [26]. In this work, we can pre-define a coarse model structure that includes every possible configuration of the semantic structure, then we will use the ideas of entropic prior and parameter extinction proposed in [26] that result in a maximum *a posteriori* (MAP) estimator.

According to (10), for each segment in each channel, we can obtain the likelihood by using the log-likelihood:

$$\begin{aligned} \log p(\mathbf{O}_t^{(j)} | S_t^{(j)} = m, \Theta) = \\ \log((R_m^{(j)})^{n_t} R_{\mu,m}^{(j)} R_{t,m}^{(j)}) + \frac{1}{2} \mu_{\mu,m}^{(j)} (\Sigma_{\mu,m}^{(j)})^{-1} (\mu_{\mu,m}^{(j)})' + \\ \frac{1}{2} \sum_{k=1}^{n_t} O_{t,k}^{(j)} (\Sigma_m^{(j)})^{-1} (O_{t,k}^{(j)})' - \frac{1}{2} \mu_{t,m}^{(j)} (\Sigma_{t,m}^{(j)})^{-1} (\mu_{t,m}^{(j)})', \end{aligned} \quad (12)$$

where we define some auxiliary parameters as follows:

$$\mu_t^{(j)} = \sum_{k=1}^{n_t} O_{t,k}^{(j)}, \quad (13)$$

$$(\Sigma_{t,m}^{(j)})^{-1} = (\Sigma_{\mu,m}^{(j)})^{-1} + n_t (\Sigma_m^{(j)})^{-1}, \quad (14)$$

$$\mu_{t,m}^{(j)} = (\Sigma_{t,m}^{(j)} ((\Sigma_{\mu,m}^{(j)})^{-1} (\mu_{\mu,m}^{(j)})' + (\Sigma_m^{(j)})^{-1} (\mu_t^{(j)})'))', \quad (15)$$

$$R_m^{(j)} = \frac{1}{(2\pi)^{\frac{n_t}{2}} |\Sigma_m^{(j)}|^{\frac{1}{2}}}, \quad (16)$$

$$R_{\mu,m}^{(j)} = \frac{1}{(2\pi)^{\frac{n_t}{2}} |\Sigma_{\mu,m}^{(j)}|^{\frac{1}{2}}}, \quad (17)$$

$$R_{t,m}^{(j)} = \frac{1}{(2\pi)^{\frac{n_t}{2}} |\Sigma_{t,m}^{(j)}|^{\frac{1}{2}}}. \quad (18)$$

The entropic prior is a bias to the compact model with good determinism. In our work, this MAP-based estimator can be incorporated into the M-step of the EM algorithm which will encourage a maximally structured and minimally ambiguous model. This is accomplished by trimming the weakly supported parameters and states, leading to a compact model with good determinism.

In our work, the MAP estimator focuses on  $\mathbf{A}$  defined in (6) that has many possible state transitions due to a large number of possible states of  $\mathbf{F}$  ( $N_1 \times N_2 = 16$  here). In other words, we want to simplify  $\mathbf{A}$  by keeping only important state transitions,

that would effectively reduce the number of useful states in  $\mathbf{F}$  and balance the incoming and outgoing transitions between the two layers. Consequentially, the MAP-based EM estimator find the optimal parameter by

$$\Theta^* = \arg \max_{\Theta} P_e(\Theta|\mathbf{O}), \quad (19)$$

where

$$P_e(\Theta|\mathbf{O}) \propto P(\mathbf{O}|\Theta)P_e(\Theta) = \left( \sum_{S,F} p(\mathbf{S}, \mathbf{F}, \mathbf{O}|\Theta) \right) P_e(\Theta), \quad (20)$$

where  $p(\mathbf{S}, \mathbf{F}, \mathbf{O}|\Theta)$  is given in (11) and  $P_e(\Theta)$  is the entropic prior of the model corresponding to parameter set  $\Theta$  that, in this work, depends on  $\mathbf{A}$  as

$$P_e(\Theta) \propto \exp \left( \sum_w \sum_p \sum_q P_{p,q}^w \log P_{p,q}^w \right), \quad (21)$$

where  $P_{p,q}^w$  denotes a transition probability in  $A_w$ . Accordingly, in the M-step of the EM algorithm, we will update the transition probabilities by maximizing the entropic prior as,

$$\mathbf{A}^* = \arg \max_{\mathbf{A}} \{ \log(P_e(\Theta|\mathbf{O})) + \sum_w \sum_p \lambda_p^w (\sum_q P_{p,q}^w - 1) \}, \quad (22)$$

where  $\lambda_{p,q}^w$  is the Lagrange multiplier to ensure  $\sum_q P_{p,q}^w = 1$ . Using a similar optimization technique discussed in [26], the MAP estimate of  $\Theta$  can be achieved by setting the derivative of log-posterior given in (22) to zero.

$$\frac{\partial \{ \log(P_e(\Theta|\mathbf{O})) + \sum_w \sum_p \lambda_p^w (\sum_q P_{p,q}^w - 1) \}}{\partial P_{p,q}^w} = 0. \quad (23)$$

Then, we can obtain

$$\frac{\xi_{p,q}^w}{P_{p,q}^w} + \log P_{p,q}^w + 1 + \lambda_{p,q}^w = 0, \quad (24)$$

where

$$\xi_{p,q}^w = p(\mathbf{S}, \mathbf{F}|\mathbf{O}, \Theta). \quad (25)$$

And this equation can be solved by the Lambert W function (Appendix A), and finally we can easily obtain

$$P_{p,q}^w = \frac{-\xi_{p,q}^w}{W(-\xi_{p,q}^w e^{1+\lambda_{p,q}^w})} \quad (26)$$

After  $\mathbf{A}$  is optimized, other parameters in  $\Theta$  can also be updated in the M-step correspondingly. We resort to the junction tree algorithm in [33] to implement the MAP-based EM algorithm. The junction tree is an auxiliary data structure that can convert a Directed Acyclic Graph (DAG), such as the MCSHMM, into an undirected graph by eliminating cycles in the graph, so that belief propagation can be effectively performed on the modified graph for inference and learning. By applying the backward-forward algorithm, we can obtain

$$\alpha_{m,t}^{(j)} = p(\mathbf{O}_1^{(j)}, \dots, \mathbf{O}_t^{(j)}, S_t^{(j)} = m^{(j)}|\Theta), \quad (27)$$

$$\beta_{m,t}^{(j)} = p(\mathbf{O}_{t+1}^{(j)}, \dots, \mathbf{O}_T^{(j)}|S_t^{(j)} = m^{(j)}, \Theta), \quad (28)$$

$$\gamma_{m,t}^{(j)} = \frac{\alpha_{m,t}^{(j)} \beta_{m,t}^{(j)}}{\sum_{m^{(j)}=1}^K \beta_{m,t}^{(j)} \beta_{m,t}^{(j)}}. \quad (29)$$

Then, we can obtain the parameter set for each each channel as follows:

$$\hat{\mu}_{\mu,m}^{(j)} = \frac{\sum_{t=1}^T \gamma_{m,t}^{(j)} \frac{\mu_t^{(j)}}{n_t^{(j)}}}{\sum_{t=1}^T \gamma_{m,t}^{(j)}}, \quad (30)$$

$$\hat{\Sigma}_{\mu,m}^{(j)} = \frac{\sum_{t=1}^T \gamma_{m,t}^{(j)} \frac{(n_t \hat{\mu}_{\mu,m}^{(j)} - \mu_t^{(j)})^2}{n_t^2}}{\sum_{i=1}^T \gamma_{m,t}^{(j)}}, \quad (31)$$

$$\hat{\Sigma}_m^{(j,i)} = \frac{\sum_{t=1}^T \gamma_{m,t}^{(j)} (\sum_{k=1}^{n_t^{(j)}} (O_{t,k}^{(j)})^2 - \frac{(\mu_t^{(j)})^2}{n_t^{(j)}})}{\sum_{t=1}^T \gamma_{m,t}^{(j)} (n_t^{(j)} - 1)}, \quad (32)$$

Consequentially, this learning process enhances the important states in  $\mathbf{F}$  and drives the weakly supported ones towards zero. According to the transitions of small probabilities, the states in  $\mathbf{F}$  that are seldom visited could be found and eventually eliminated. Hereby the state space is optimized by balancing the state transitions between the two layers. Therefore, there are two trimming processes, one is the transition trimming, the other is the state trimming.

For the MCSHMM, we can obtain the transition probabilities in the E-step of the EM algorithm, to find a transition that can be trimmed, we need to find the transition probability for which the gain of the entropic prior outweighs the loss in the likelihood, as shown below:

$$\frac{P_e(\Theta \setminus P_{p,q}^w)}{P_e(\Theta)} \geq \frac{P(\mathbf{O}|\Theta)}{P(\mathbf{O}|\Theta \setminus P_{p,q}^w)} \quad (33)$$

where  $\Theta \setminus P_{p,q}^w$  represents the parameter set  $\Theta$  without the element  $P_{p,q}^w$ .

Now we need to trim the state space based on estimated/trimmed transition probabilities. The purpose is to keep the useful states in  $\mathbf{F}$  which are visited frequently, as expressed by its incoming and exit transitions. Similar to [26], we can detect trimmable states by balancing the prior probability of all its incoming and exit transitions against the probability mass that flow through it, as shown below,

$$\frac{P(\Theta \setminus F_l)}{P(\Theta)} \geq \prod_{m,n,k} (a_{m,n,l}^{(k)})^{a_{m,n,l}^{(k)}} \quad (34)$$

where  $k = \{1, 2, 3\}$ , and

$$a_{m,n,l}^1 = P(S_t^{(1)} = m | S_{t-1}^{(1)} = n, F_{t-1} = l); \quad (35)$$

$$a_{m,n,l}^2 = P(S_t^{(2)} = m | S_{t-1}^{(2)} = n, F_{t-1} = l); \quad (36)$$

$$a_{m,n,l}^3 = P(F_t = l | S_t^{(1)} = m, S_t^{(2)} = n). \quad (37)$$

The above state trimming occurs in each iteration in the M-step, and each iteration may have some transitions/states trimmed, leading to a fast learning process. In practice, we need to decide the size of expected state space. Here we keep 6 of out 16 states in  $\mathbf{F}$ . After model learning, the Viterbi algorithm can be used to estimate the optimal state sequence for both channels at the first layer, i.e.,  $\mathbf{S}$ , which encodes two semantic structures, i.e., camera views and play types. Our EM algorithm implementation was based on the Bayes Net Matlab Toolbox developed by KP Murphy.<sup>1</sup>

<sup>1</sup><http://www.cs.ubc.ca/~murphyk/Software/BNT/bnt.html>

Test Videos	Semantics	Supervised GMM	HMM <sup>(1)</sup> [27]	HMM <sup>(2)</sup> [27]	HMM <sup>(3)</sup> [28]	CHMM <sup>(1)</sup> [20]	SHMM [29]	CHMM <sup>(2)</sup> [20]	MCSHMM
Video-1 (156 shots)	Play Type	26.92%	43.59%	60.90%	77.56%	41.03%	80.13%	78.85%	82.05%
	Camera View	35.90%	55.77%	72.44%	76.28%	58.33%	79.49%	81.41%	84.62%
Video-2 (163 shots)	Play Type	30.67%	49.07%	61.34%	70.56%	53.37%	77.91%	76.69%	81.59%
	Camera View	41.10%	61.35%	67.48%	79.14%	65.03%	84.05%	80.98%	85.28%
Video-3 (167 shots)	Play Type	23.35%	44.91%	49.10%	58.08%	52.10%	68.86%	70.06%	75.45%
	Camera View	37.12%	53.29%	58.68%	61.08%	58.08%	74.85%	77.25%	81.44%
Video-4 (168 shots)	Play Type	36.90%	58.33%	64.67%	70.66%	67.86%	77.98%	69.62%	82.74%
	Camera View	47.61%	64.88%	70.24%	73.21%	69.05%	79.17%	75.60%	84.52%
Video-5 (168 shots)	Play Type	29.17%	50.59%	61.90%	72.02%	70.24%	74.40%	70.83%	80.95%
	Camera View	40.47%	56.55%	65.48%	73.81%	60.71%	73.21%	64.88%	77.38%
Video-6 (170 shots)	Play Type	38.25%	54.11%	60.59%	70.59%	76.47%	80.59%	70.59%	84.71%
	Camera View	41.17%	52.35%	64.12%	70.59%	74.70%	80.00%	71.18%	84.11%
Video-7 (170 shots)	Play Type	37.05%	64.70%	71.17%	70.59%	72.35%	75.88%	76.47%	83.53%
	Camera View	52.35%	68.24%	72.35%	75.29%	67.06%	81.18%	71.76%	87.06%
Video-8 (171 shots)	Play Type	40.93%	56.14%	65.50%	72.51%	68.82%	78.95%	82.94%	84.80%
	Camera View	50.88%	62.57%	75.44%	77.78%	76.02%	80.11%	81.18%	88.30%
Video-9 (173 shots)	Play Type	39.88%	63.01%	67.05%	69.36%	67.63%	75.14%	69.36%	82.08%
	Camera View	43.93%	66.07%	68.21%	71.68%	69.94%	77.46%	73.41%	84.97%
Average Accuracy		43.59%	57.44%	62.62%	71.85%	63.54%	77.42%	75.08%	82.92%

TABLE I  
SHOT-BASED CLASSIFICATION RESULTS OF EIGHT ALGORITHMS FOR NINE 30-MIN NFL VIDEOS.

## V. EXPERIMENTAL RESULTS

### A. Experimental Settings

The MCSHMM was tested on nine 30-minute NFL football games (352×240, 30fps) that have been pre-segmented into a series of consecutive play shots by removing commercials and replays. Each game has 150-170 shots and each shot has 300-500 frames. Additionally, by using the Bayes Net Matlab Toolbox, we generated four two-channel synthetic data sequences each of which has 200 segments of variable-length observations (200-400 samples) according to MCSHMMs learned from four real videos. Using both synthetic and real data allows us to have comprehensive algorithm validation and evaluation. Our algorithm was implemented in Matlab 7.0 and tested on a PC with 3.2GHz CPU and 1GB memory.

Seven previous methods are involved for comparison, as shown in Table II, including a supervised GMM (order 3), the HMM with Gaussian emission (HMM<sup>(1)</sup>) and the HMM with GMM emission (HMM<sup>(2)</sup>) [27], and the embedded GMM-HMM (HMM<sup>(3)</sup>) in [28] and the SHMM in [29]. We also implemented two CHMM-based methods [20], among which CHMM<sup>(1)</sup> and CHMM<sup>(2)</sup> uses a GMM and a segmental observation model respectively. The first five explore two semantic structures (plays and views) independently and separately, while the last three estimate both jointly.

Methods	State number	Observation model	Transition matrix
GMM	4	3-order GMM	N/A
HMM <sup>(1)</sup>	4	Gaussian	4 × 4
HMM <sup>(2)</sup>	4	3-order GMM	4 × 4
HMM <sup>(3)</sup>	4	two-layer GMM	4 × 4
SHMM	4	segmental model	4 × 4
CHMM <sup>(1)</sup>	4, 4	3-order GMM	16 × 16
CHMM <sup>(2)</sup>	4, 4	segmental model	16 × 16
MCSHMM	4, 4	segmental model	4 × 4, 24 × 4 × 2

TABLE II  
THE MODEL SETTINGS FOR ALL TESTING METHODS.

### B. Model Learning and Classification Results

The initialization is important in EM learning. We adopted a coarse-to-fine initialization strategy that uses the training result of a simpler model to initialize a more complex one. Specially, we first use K-mean (4-class) to obtain a coarse classification, and this result can be utilized to initialize HMM<sup>(1)</sup> whose training result can be used to initialize HMM<sup>(2)</sup>, and so on. The training result of SHMM was used to initialize MCSHMM. Moreover, since the first shot is always in the central view, the initial probability of central view is set to be 1. We can initialize the transition matrix by computing the frequencies of different state transition in a couple of real games. Simultaneous structure learning and parameter learning are crucial for the application of MCSHMM. We found the traditional ML estimator cannot fully take advantage of the MCSHMM due to its large state space (before trimming). The entropic prior-based MAP estimator is capable of finding the optimal model structure (a trimmed state space) and parameters jointly. It was mentioned in [26] that this MAP estimator can accelerate learning, and rescue EM from local probability maxima. Fig. 3 and Table I shows the experimental results on both synthetic data and real videos. It is clearly shown that the proposed MCSHMM outperforms all other algorithms with significant improvements.

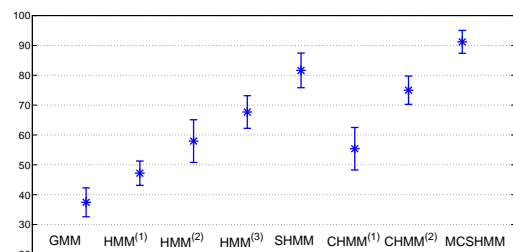


Fig. 3. Classification results on synthetic data for eight algorithms.

### C. More Discussion

The reason that SHMM is better than other HMMs (HMM<sup>(1)</sup>, HMM<sup>(2)</sup>, HMM<sup>(3)</sup>) is because of the segmental model used. The improvement of MCSHMM over SHMM and two CHMMs lies in the new hierarchical-parallel dynamics. CHMMs usually assume relative strong correlation between two Markov chains that is not true in our case. MCSHMM is able to balance the dependency both within each channel and across the two channels via the hybrid hierarchal-parallel dynamics. To prove that, we compared the learned transition matrices of both models with the ground-truth that has a sparse nature. We observed that the learning results in MCSHMM are much closer to the ground-truth than that in CHMM.

However, the current MCHSMM implementation (without program optimization) has the highest computational load (about 50 minutes for one video), while other methods are between 2-20 minutes. It is mainly due to the large number of initial states (most of them will be trimmed during training) at the second-layer and the complexity of the segmental observation model. More efficient learning schemes are needed for fast state trimming. It is also possible to extend the MCSHMM to the cases where more than two channels are involved. The major challenge will be how to condense the initial state space effectively and efficiently. A possible strategy could be “incremental growing” rather than “progressive trimming”.

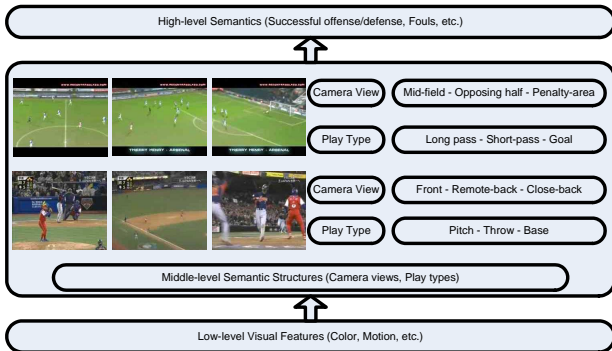


Fig. 4. A proposed semantic analysis paradigm for sports video mining.

The proposed MCSHMM could be applied to other field-based sports, where the two mid-level semantic structures, play types and camera views, can be well specified. After extracting some useful visual features that may be composed of color distribution, camera motion, or relevant visual landmarks, we can invoke MCHSMM to estimate play types and camera views jointly, which can be further used to infer certain high-level semantics by involving some domain knowledge. For example, as shown in Fig. 4, in a soccer game, we can define several views, such as the *whole field*, *mid-field*, *opposing-half*, *field-corner*, and *penalty area*, and a few plays, such as *kickoff*, *pass*, *free kick*, *corner kick*, and *shoot*, etc. A goal highlight can be specified by a *shoot play* in the *penalty area* followed by a *kickoff play* in the *mid-field*. We can have a similar semantic analysis paradigm for video mining in the baseball game, as shown in Fig. 4.

## VI. CONCLUSION

We have presented a new MCSHMM for sports video mining that incorporates ideas from CHMMs, SHMMs, and HHMMs. The new model offers more flexibility, functionality and capacity than its precedents in two aspects. The first is the segmental model that can effectively deal with variable-length observations, and the second is the hierarchical-parallel dynamics that involve decision-level fusion to capture interaction between two Markov chains. Moreover, we developed a MAP-based estimator to optimize the model structure and parameters simultaneously. We showed the usefulness of MCSHMM for exploring two semantic structures in American football video. This new model could also be applied to other video mining applications involving multiple data or information sources.

### APPENDIX A THE LAMBERT W FUNCTION

The Lambert W function is the inverse function of

$$f(w) = we^w, \quad (38)$$

where  $e^w$  is the natural exponential function and  $w$  is any complex number. For any complex number  $z$ , we have

$$z = W(z)e^{W(z)}. \quad (39)$$

So, we have

$$\log W(z) + W(z) = \log z, \quad (40)$$

which equals to

$$-\log W(z) - W(z) + \log z = 0. \quad (41)$$

If we set  $z = e^x$ , we can obtain

$$-\log W(e^x) - W(e^x) + x = 0. \quad (42)$$

For any complex number  $y$ ,

$$\log \frac{y}{W(e^x)} - \frac{y}{W(e^x)} + x - \log y = 0. \quad (43)$$

If we let

$$x = 1 + \lambda_{p,q}^w + \log y, \quad (44)$$

and

$$y = -\xi_{p,q}^w, \quad (45)$$

then the (43) can be expressed by

$$\log \frac{-\xi_{p,q}^w}{W(e^{1+\lambda_{p,q}^w+\log(-\xi_{p,q}^w)})} + \frac{\xi_{p,q}^w}{W(e^{1+\lambda_{p,q}^w+\log(-\xi_{p,q}^w)})} + 1 + \lambda_{p,q}^w = 0. \quad (46)$$

And we can easily find the solution of  $P_{p,q}^w$  by comparing (46) with the (24).

### ACKNOWLEDGEMENTS

The authors also thank the anonymous reviewers for their valuable comments and suggestions that improved this paper.

## REFERENCES

- [1] T. Mei, Y. Ma, H. Zhou, W. Ma, and H. Zhang, "Sports video mining with mosaic," in *Proc. the 11th IEEE International Multimedia Modeling Conference*, 2005, pp. 164–171.
- [2] S.-F. Chang, "The holy grail of content-based media analysis," *IEEE Multimedia Magazine*, vol. 9, no. 2, pp. 6–10, 2002.
- [3] C.W. Ngo, Y.F. Ma, and H.J. Zhang, "Video summarization and scene detection by graph modeling," *IEEE Trans. Multimedia*, vol. 15, no. 2, pp. 296–305, Feb. 2005.
- [4] A. Kokaram, N. Rea, R. Dahyot, M. Tekalp, P. Boutheimy, P. Gros, and I. Sezan, "Browsing sports video: trends in sports-related indexing and retrieval work," *IEEE Signal Processing Magazine*, vol. 23, no. 2, pp. 47–58, March 2006.
- [5] M. Worring and G. Schreiber, "Semantic image and video indexing in broad domains," *IEEE Trans. Multimedia*, vol. 9, no. 5, pp. 909–911, Aug. 2007.
- [6] J. Fan, X. Zhu, A.K. Elmagarmid, W.G. Aref, and L. Wu, "Classview: Hierarchical video shot classification, indexing, and accessing," *IEEE Trans. Multimedia*, vol. 6, no. 1, pp. 70–86, 2004.
- [7] Z.Y. Xiong, X.S. Zhou, Q. Tian, Y. Rui, and T.S. Huang, "Semantic retrieval of video - review of research on video retrieval in meetings, movies and broadcast news, and sports," *IEEE Signal Processing Magazine*, vol. 23, no. 2, pp. 18–27, March 2006.
- [8] Y. Rui, T. Huang, and S. Mehrotra, "Constructing table-of-content for video," in *Proc. ACM Multimedia conference*, 1998.
- [9] D. Zhong and S.-F. Chang, "Structure analysis of sports video using domain models," in *Proc. IEEE Conference on Multimedia and Expo*, Tokyo, Japan, 2001.
- [10] Y. Gong, L. T. Sin, C. H. Chuan, H. J. Zhang, and M. Sakauchi, "Automatic parsing of tv soccer programs," in *Proc. International Conference on Multimedia Computing and Systems*, 1995, pp. 167–174.
- [11] L.Y. Duan, M. Xu, T.S. Chua, Q. Tian, and C.S. Xu, "A mid-level representation framework for semantic sports video analysis," in *Proc. IEEE International Conference on Multimedia and Expo*, 2003.
- [12] J. Assfalg, M. Bertini, C. Colombo, A.D. Bimbo, and W. Nunziati, "Semantic annotation of soccer videos: automatic highlights identification," in *Computer Vision and Image Understanding*, 2003.
- [13] J. Assfalg, M. Bertini, A.D. Bimbo, W. Nunziati, and P.Pala, "Soccer highlights detection and recognition using HMMs," in *Proc. IEEE International Conference on Multimedia and Expo*, 2002.
- [14] W. Hua, M. Han, and Y. Gong, "Baseball scene classification using multimedia features," in *Proc. IEEE International Conference on Multimedia and Expo*, 2002.
- [15] L. Xie, S.-F. Chang, A. Divakaran, and H. Sun, "Structure analysis of soccer video with hidden Markov models," in *Proc. International Conference on Acoustic, Speech and Signal Processing*, Orlando FL, 2002.
- [16] J. Lafferty, A. McCallum, and F. Pereira, "Conditional random fields: Probabilistic models for segmenting and labeling sequence data," in *Proc. the Eighteenth International Conference on Machine Learning (ICML)*, 2001.
- [17] D. Sadlier and N. Connor, "Event detection in field-sports video using audio-visual features and a support vector machine," *IEEE Trans. Circuits Syst. Video Technology*, vol. 15, no. 10, pp. 1225–1233, Oct. 2005.
- [18] G. Bouchard and G. Celeux, "Selection of generative models in classification," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 28, no. 4, pp. 544 – 554, 2006.
- [19] Z. Ghahramani and M. Jordan, "Factorial hidden Markov models," in *Proc. Conf. Advances in Neural Information Processing Systems, NIPS*, 1996.
- [20] M. Brand, "Coupled hidden Markov models for modeling interactive processes," Tech. Report, MIT Media Lab, 1997.
- [21] S. Fine, Y. Singer, and N. Tishby, "The hierarchical hidden Markov model: analysis and applications," *Machine Learning*, vol. 32, no. 1, pp. 41–62, 1998.
- [22] L. Xie, S. Chang, A. Divakaran, and H. Sun, "Unsupervised mining of statistical temporal structures in video," in *Video Mining*, D. Doremann A. Rosenfeld, Ed., chapter 10. Kluwer Academic Publishers, 2003.
- [23] M. Gales and S. Young, "The theory of segmental hidden markov models," Technical Report CUED/F-INFENG/TR 133, Cambridge University, 1993.
- [24] N. Friedman and D. Koller, "Being bayesian about network structure. a bayesian approach to structure discovery in bayesian networks," *Machine Learning*, vol. 50, no. 1, pp. 95–125, Jan. 2003.
- [25] Z. Ghahramani, "Graphical models: parameter learning," in *The Handbook of Brain Theory and Neural Networks*, M.A. Arbib, Ed. MIT Press, 2002.
- [26] Matthew Brand, "Structure learning in conditional probability models via an entropic prior and parameter extinction," *Neural Computation*, vol. 11, no. 5, pp. 1155–1182, 1999.
- [27] Y. Ding and G. Fan, "Camera view-based american football video analysis," in *Proc. the Eighth IEEE International Symposium on Multimedia*, 2006, pp. 317–322.
- [28] Y. Ding and G. Fan, "Two-layer generative models for sport video mining," in *Proc. IEEE International Conference on Multimedia and Expo*, 2007.
- [29] Y. Ding and G. Fan, "Segmental hidden Markov models for view-based sport video analysis," in *Proc. IEEE International Conference on Computer Vision and Pattern Recognition*, 2007.
- [30] 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.
- [31] M. Srinivasan, S. Venkatesh, and R. Hosie, "Qualitative estimation of camera motion parameters from video sequences," *Pattern Recognition*, vol. 30, pp. 593–606, 1997.
- [32] N. Petrovic, A. Ivanovic, and N. Jovic, "Recursive estimation of generative models of video," in *Proc. IEEE International Conference on Computer Vision and Pattern Recognition*, 2006, vol. 1, pp. 17–22.
- [33] K. Murphy, *Dynamic Bayesian Networks: Representation, Inference and Learning*, Ph.D. thesis, UC Berkeley, 2002.



approaches for video mining.

**Yi Ding (S'06)** received the B.S. degree in communication engineering from Xi'an University of Technology, China, and the M.S. degree in communication engineering from Xidian University, China, in 2002 and 2005 respectively. He is currently working toward the Ph.D. degree in the School of Electrical and Computer Engineering, Oklahoma State University, Stillwater, OK. His research interests include multimedia content analysis, pattern recognition, image processing and computer vision. His current research focuses on machine learning



**Guoliang Fan (S'97, M'01, SM'05)** received his B.S. degree in Automation Engineering from Xi'an University of Technology, Xi'an, China, M.S. degree in Computer Engineering from Xidian University, Xi'an, China, and Ph.D. degree in Electrical Engineering from University of Delaware, Newark, DE, USA, in 1993, 1996, and 2001, respectively. From 1996 to 1998, he was a graduate assistant in the Department of Electronic Engineering, Chinese University of Hong Kong. Since 2001, Dr. Fan has been with the School of Electrical and Computer Engineering, Oklahoma State University (OSU), Stillwater, OK, where he is currently an Associate Professor. He was awarded the First Prize in 1997 IEEE Hong Kong Section Postgraduate Student Paper Contest and the First Prize in 1997 IEEE Region 10 (Asia-Pacific) Postgraduate Paper Contest. Dr. Fan is a recipient of the National Science Foundation (NSF) CAREER award (2004). He received the Halliburton Excellent Young Teacher Award (2004), Halliburton Outstanding Young Faculty award (2006) from the College of Engineering at OSU and Outstanding Professor Award (2008) from OSU-IEEE. His research interests include image processing, computer vision and machine learning.