Multi-channel Segmental Hidden Markov Models for Sports Video Mining

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 new multi-channel segmental hidden Markov model (MCSHMM) for sports video mining that is a unique probabilistic graphical model with two advantages. One is the integration of both hierarchical and parallel structures that offer more flexibility and capacity of capturing the interaction between multiple Markov chains. The other is the incorporation of the segmental HMM that represents a variable-length sequence of observations. Especially, we develop a maximum a posteriori (MAP) estimator to optimize model structures and model parameters simultaneously. The proposed MCSHMM is used for American football video analysis, where two semantics structures, play types and camera views, are involved. The experiment shows that the MCSHMM outperforms previous HMM-based approaches.

Categories and Subject Descriptors
1.4.8 [Image Processing and Computer Vision]: Scene Analysis

General Terms
Algorithms

Keywords
sports video mining, entropic prior, hidden Markov models

1. INTRODUCTION

1.1 Motivation

Video mining is to discover knowledge, patterns and events, or called semantic structures, in the video data, and its benefits range from efficient summarizing of video content of interest to facilitating video access and retrieval in large databases. In this work, we study sports video mining where relatively definite semantic structures exist and that has tremendous commercial value. However, multiple semantic structures could concur in a sports video, e.g., play types (what happened?) and camera views (where it happened?). Specifically, we are interested in developing powerful generative models that are able to capture the interaction between multiple semantic structures and improve the overall performance. Here we propose a new multi-channel segmental hidden Markov model (MCSHMM) that is inspired by recent progress on graphical model theory. As a case study, we test the new MCSHMM for American football video analysis.

1.2 Related works

There are two major methods in video mining research [8]. One is the discriminative approaches, e.g., [9], and the other is the generative approaches, e.g., [11]. A generative approach is usually preferred when video mining is performed on a large data set due to its better generality. The Dynamic Bayes Network (DBN) [10] provides a unified probabilistic framework to represent various generative models with different graphical structures. Recently, several specific DBNs, i.e., the Factorial HMM [7], the Coupled HMM (CHMM) [1], and the influence model [14], were proposed to explore complex semantic structures underlying multiple data or feature streams. Additionally, it is imperative to have simultaneous structure learning and parameter learning for complex DBNs to ensure their effectiveness and efficiency in practice. In [2], the concepts of entropic prior and parameter extinction were proposed to simplify and optimize the model structure by a state trimming process. In [13], the hierarchical HMMs (HHMMs) were used for unsupervised sports video mining, where the reverse-jump Markov Chain Monte Carlo (RJMCMC) method is embedded in the learning of HHMMs to support model selection and feature selection.

1.3 Our work and contribution

This work is built on our previous research [3, 5, 4], where two semantic structures were considered independently, i.e., play types and camera views. We improved existing HMMs by developing more effective observation models to capture rich statistics of a variable-length sequence of observations. This work is aimed at how to explore the interaction between two Markov chains (i.e., semantic structures). There are two contributions. One is the integration of both hierarchical and parallel structures and the incorporation of a segmental observation model. The other is the application of entropic prior and parameter extinction in the model training to optimize the model structure and model parameters simultaneously. The MCSHMM offers more flexibility, functionality, and capacity compared with existing HMMs, and holds great potential for other video mining applications.
2. PREVIOUS RESEARCH

Sport videos normally have two fundamental semantic structures regarding “what” and “where”. Or in the case of American football videos, they are called “play types” and “camera views” [4, 5], which could be used as building blocks to deliver higher level semantics. Each semantic structure involves multiple semantic events that are assumed to follow a first-order Markov chain over time, as shown in Fig. 1.

![Figure 1: The two semantic structures in the American football videos, camera views (left) and play types (right).](image)

2.1 Problem Formulation

Given a video sequence has been partitioned into $T$ shots of variable-length, each of which corresponds to a specific play and a camera view, the video mining task here can be formulated as the inference problem of a generative model where the latent state indicates the semantic event of a shot, and the observation model characterizes visual features extracted from all frames in a shot with respect to different semantic events. The features used for view analysis includes the color distribution and the angle of yard lines [4], and that of play analysis involves motion vectors extracted by an efficient optical flow technique [12]. How to capture rich statistics of frame-wise observations to support shot-wise states is a key issue for our previous research. A two-layer generative model was proposed in [5] that defines states on a Gaussian mixture model (GMM) (not on the observation), and involves an embedded HMM-GMM structure.

![Figure 2: The segmental HMM (SHMM) where $O_j^{i}$ are the observations of shot $i$ with $\tau_i$ frames.](image)

2.2 Segmental Hidden Markov Models

The SHMM was originally proposed for speech signal processing [6] that involves a segmental model to represent a variable-length sequence of observations, as shown in Fig. 2. In the SHMM, each state emits a sequence of observations that is called a segment. In each segment, all observation are conditionally independent given the mean of the segment, leading to a closed-form likelihood function for the observation model. In [4], the SHMM was shown effective on viewing-based video mining compared with the traditional HMMs and the embedded HMM-GMM in [5]. Hence the SHMM will serve as the building block for our proposed model.

3. THE PROPOSED MCSHMM

This research focuses on exploring the interaction between multiple semantic structures with the aim to improve the overall mining performance. Specifically, it was observed that camera views and play types are quite related to each other during a 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. It is tempting to invoke the Coupled HMM (CHMM) proposed in [1] for our problem. However, the CHMM usually assumes 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 multi-channel SHMM model (MCSHMM) that involves two parallel SHMMs and a two-layer hierarchical Markovian structure, as shown in Fig. 3.

![Figure 3: In the MCSHMM, the first layer includes two SHMMs , and the second layer captures the interaction of two channels at the first layer. $O_j^{i} = \{O_j^{i1}, O_j^{i2}\}$ represents all frame-wise observations of shot $i$ in channel $j \in \{1, 2\}$.](image)

3.1 Model Parameters

The unique feature of the MCSHMM is the integration of both parallel and hierarchical structures on latent states and the incorporation of a segmental observation model, which allow the MCSHMM have more flexibility, capacity, and functionality than the CHMM, SHMM and HHHM. 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, $S = \{S_i^j | i = 1, ..., T; j = 1, 2\}$ denotes the state sequence of two channels where $S_i^j$ denotes the state of shot $i$ in channel $j$, and $O_i^j = \{O_i^{j1}, O_i^{j2}\}$ indicates observations of shot $i$ with $\tau_i$ frames in channel $j$. $F = \{F_i = (S_i^1, S_i^2) | i = 1, ..., T\}$ represents the state sequence at the second layer where each state consists of two states at the lower layer. Therefore, the MCSHMM’s parameter set $\Theta$ includes following components $\Theta = \{\Pi, A, \Omega\}$:

- Initial probabilities:
  \[ \Pi = \{P(S_1^1), P(S_1^2), P(F_1|S_1^1, S_1^2)\}; \] (1)

- Transition probabilities:
  \[ A = \{A_{uw}, w = 1, 2, 3\}, \] (2)

where

\[ A_1 = \{P(S_i^1 = m|S_{i-1}^1 = n, F_{i-1} = l)\}, \]
\[ A_2 = \{P(S_i^2 = m|S_{i-1}^2 = n, F_{i-1} = l)\}, \]
\[ A_3 = \{P(F_i = l|S_i^1 = m, S_i^2 = n)\}; \]
• Observation density functions:
  \[ p(O_i | S_i = m, \Omega) = \int N(\mu; \mu_{m,j}, \Sigma_{\mu,m}) \prod_{k=1}^{T} N(O_{k}^{(i,j)} | \mu, \Sigma_{\mu,m}) d\mu, \]  
  (3)

where \( \Omega = \{ \mu_{m,j}, \Sigma_{\mu,m}, \Sigma_{\mu,i,j} | j = 1, 2; m = 1, ..., N_{j} \} \) specifies the two segmental models, \( N_{j} \) is the number of states in channel \( j \) (here \( N_{1} = N_{2} = 4 \)), and \( O_{k}^{(i,j)} \) denotes the observation of frame \( k \) in shot \( i \) and channel \( j \). We refer the reader to [4] for more details.

Given the dual-channel observations of \( T \) shots, \( O = \{ O_{i}^{(j)} | i = 1, ..., T; j = 1, 2 \} \), the joint likelihood is defined as:

\[ p(S, F, O | \Theta) = P(S_{1}^{T}) P(S_{1}^{T}) P(F_{i} | S_{i}^{T}) P(O_{i}^{(j)} | S_{i}^{T}) \prod_{i=1}^{T} P(F_{i} | S_{i}^{T}, S_{i-1}^{T}) \prod_{i=2}^{T} \prod_{j=1}^{2} P(O_{i}^{(j)} | S_{i}^{T}). \]  
(4)

The Expectation Maximization (EM) algorithm is often used for learning a DBN, such as the MCSHMM, which finds the optimal parameters by maximizing the likelihood function (e.g., (4)) through iterative Expectation (the E-step) and Maximization (the M-step). However, the maximum likelihood (ML) learning of the MCSHMM could be problematic due to the complex state space specified by \( S \) and \( F \), and a large set of model parameters, such as \( A \) given in (2).

### 3.2 Learning of the MCSHMM

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 advancedly, the two learning processes could be unified into one framework where the model structure. More advancingly, the two learning processes could be unified into one framework where the model structure and parameters can be optimized simultaneously [13, 2]. In this work, we will use the ideas of entropic prior and parameter extinction proposed in [2] that result in a maximum a posteriori (MAP) estimator. This MAP 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 this work, the MAP estimator focuses on \( A \) defined in (2) that has many possible state transitions due to a large number of possible states of \( F \) (\( N_{1} \times N_{2} = 16 \) here). In other words, we want to simplify \( A \) by keeping only important state transitions, that would effectively reduce the number of useful states in \( F \) and balance the incoming and outgoing transitions between the two layers. Consequently, the MAP-based EM estimator find the optimal parameter by

\[ \Theta^* = \arg \max_{\Theta} P_{c}(\Theta | O), \]  
(5)

where

\[ P_{c}(\Theta | O) \propto P(O | \Theta) P_{c}(\Theta) = \left( \sum_{S,F} p(S,F,O | \Theta) \right) P_{c}(\Theta), \]  
(6)

where \( p(S,F,O | \Theta) \) is given in (4) and \( P_{c}(\Theta) \) is the entropic prior of model \( \Theta \) that, in this work, depends on \( A \) as

\[ P_{c}(\Theta) \propto \exp \left( \sum_{w} \sum_{p} \sum_{q} P_{w} P_{w,q} \log P_{w,q} \right), \]  
(7)

where \( P_{w,q} \) 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,

\[ A^* = \arg \max_{A} \left( \log P_{c}(\Theta | O) + \sum_{w} \sum_{p} \sum_{q} \lambda_{w} (\sum_{p} P_{w,q} - 1) \right), \]  
(8)

where \( \lambda_{w} \) is the Lagrange multiplier to ensure \( \sum_{q} P_{w,q} = 1 \).

Using a similar optimization technique discussed in [2], the MAP estimate of \( \Theta \) can be achieved by setting the derivative of log-posterior given in (8) to zero. Consequently, this step enhances the important states in \( F \) and drives the weakly supported ones towards zero. According to the transitions of small probabilities, the states in \( F \) 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. After \( A \) is optimized, other parameters in \( \Theta \) can also be updated in the M-step.

We resort to the junction tree algorithm in [10] 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. Our EM algorithm implementation was based on the Bayes Net Matlab Toolbox developed by KP Murphy. After learning, the Viterbi algorithm can be used to obtain optimal state sequences for both channels at the first layer, i.e., \( S \), which encodes two semantic structures, i.e., camera views and play types. The second layer state set \( F \) is available that shows the dependency between two semantic structures and it is trimmed to six states that is less than 40% of the complete state set.

### 4. EXPERIMENTAL RESULTS

The algorithm was implemented in Matlab 7.0 and tested on a PC with 3.2G CPU and 1G memory for six 30-min NFL videos. The proposed MCSHMM is compared with our previous methods that include the HMM with Gaussian emission (HMM\(^{(1)}\)) and GMM emission (HMM\(^{(2)}\)) [3], and the embedded GMM-HMM (HMM\(^{(3)}\)) in [5] and the SHMM in [4]. We also implemented two CHMM-based approaches [1], among which CHMM\(^{(1)}\) uses shot-wise features like the HMM\(^{(1)}\) and CHMM\(^{(2)}\) frame-wise features like the SHMM.

We adopted a coarse-to-fine progressive initialization strategy that uses the training result of a simpler model to initialize a more complex model. For example, we used the K-mean result to initialize HMM\(^{(1)}\) whose training result was used to initialize HMM\(^{(2)}\), and the training result of SHMM was used to initialize MCSHMM. Additionally, some prior knowledge can also be used. For example, the first shot is always in the central view. The initialization of \( A \) may require some actual calculation of the frequencies of different state transitions in a couple of games.

\(^{1}\)http://www.cs.ubc.ca/~murphyk/Software/BNT/bnt.html
### Table 1: Shot-based classification results of seven algorithms for six 30-min NFL videos.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Video-1</td>
<td>Play Type</td>
<td>43.59%</td>
<td>60.90%</td>
<td>77.56%</td>
<td>41.03%</td>
<td>80.13%</td>
<td>78.85%</td>
<td>82.05%</td>
</tr>
<tr>
<td>(156 shots)</td>
<td>Camera View</td>
<td>55.77%</td>
<td>72.44%</td>
<td>76.28%</td>
<td>58.33%</td>
<td>79.49%</td>
<td>81.41%</td>
<td>84.62%</td>
</tr>
<tr>
<td>Video-2</td>
<td>Play Type</td>
<td>49.07%</td>
<td>61.34%</td>
<td>70.56%</td>
<td>53.37%</td>
<td>77.91%</td>
<td>76.69%</td>
<td>81.59%</td>
</tr>
<tr>
<td>(163 shots)</td>
<td>Camera View</td>
<td>61.35%</td>
<td>67.48%</td>
<td>79.14%</td>
<td>65.03%</td>
<td>84.05%</td>
<td>80.98%</td>
<td>85.28%</td>
</tr>
<tr>
<td>Video-3</td>
<td>Play Type</td>
<td>41.91%</td>
<td>48.10%</td>
<td>68.08%</td>
<td>52.10%</td>
<td>68.96%</td>
<td>70.68%</td>
<td>75.45%</td>
</tr>
<tr>
<td>(167 shots)</td>
<td>Camera View</td>
<td>53.29%</td>
<td>58.68%</td>
<td>61.48%</td>
<td>58.08%</td>
<td>74.85%</td>
<td>77.25%</td>
<td>81.44%</td>
</tr>
<tr>
<td>Video-4</td>
<td>Play Type</td>
<td>58.33%</td>
<td>64.67%</td>
<td>70.66%</td>
<td>67.86%</td>
<td>77.98%</td>
<td>69.62%</td>
<td>82.74%</td>
</tr>
<tr>
<td>(168 shots)</td>
<td>Camera View</td>
<td>64.88%</td>
<td>70.24%</td>
<td>73.21%</td>
<td>69.05%</td>
<td>79.17%</td>
<td>75.60%</td>
<td>84.52%</td>
</tr>
<tr>
<td>Video-5</td>
<td>Play Type</td>
<td>64.70%</td>
<td>71.11%</td>
<td>70.59%</td>
<td>72.35%</td>
<td>75.88%</td>
<td>76.47%</td>
<td>83.33%</td>
</tr>
<tr>
<td>(170 shots)</td>
<td>Camera View</td>
<td>68.24%</td>
<td>72.35%</td>
<td>75.29%</td>
<td>67.06%</td>
<td>81.18%</td>
<td>71.76%</td>
<td>87.06%</td>
</tr>
<tr>
<td>Video-6</td>
<td>Play Type</td>
<td>56.14%</td>
<td>65.50%</td>
<td>72.51%</td>
<td>68.82%</td>
<td>78.95%</td>
<td>82.94%</td>
<td>84.80%</td>
</tr>
<tr>
<td>(171 shots)</td>
<td>Camera View</td>
<td>62.57%</td>
<td>75.44%</td>
<td>77.78%</td>
<td>76.02%</td>
<td>80.11%</td>
<td>81.18%</td>
<td>88.30%</td>
</tr>
<tr>
<td>Average Accuracy</td>
<td>53.90%</td>
<td>66.78%</td>
<td>71.89%</td>
<td>62.45%</td>
<td>78.21%</td>
<td>75.90%</td>
<td>84.43%</td>
<td></td>
</tr>
</tbody>
</table>

In this work, we are interested in two semantic structures of American football videos, camera views and play types. Our previous four algorithms can only be applied to explore them individually and independently. On the other hand, the proposed MCSHMM can jointly optimize the two tasks with the aim to improve the overall mining performance. Table 1 shows the comparison of the experimental results where the six algorithms are evaluated in terms individual classification accuracy for camera views and play types as well as the overall classification accuracy for all six videos. Specifically, we have following observation and discussion.

- **Observation models:** The major difference between our previous HMM-based algorithms (HMM\(^{(1)}\), HMM\(^{(2)}\), HMM\(^{(3)}\) and SHMM) is the observation model used. The SHMM is able to capture the rich statistics of a variable-length sequence of observations, and it serves as the building block for the MCSHMM.

- **Dynamic models:** The improvement of MCSHMM over SHMM and two CHMMs lies in the new two-layer dynamic mechanism imposed on the hidden states that effectively characterize the mutual interaction between two Markov chains and feeds it back into each channel to boost the performance of inference and learning.

- **Model learning:** Simultaneous structure learning and parameter learning are crucial for the application of MCSHMM. The traditional ML estimator cannot fully take advantage of the MCSHMM, and the entropic prior-based MAP estimator is capable of finding the optimal model structure and parameter jointly.

- **Computational cost:** However, the MCHSMM (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 states at the second-layer and the complexity of the frame-wise observation model.

### 5. CONCLUSION

We presented a new multi-channel segmental hidden Markov model (MCSHMM) for sports video mining. The MCSHMM incorporates the ideas from CHMMs, SHMMs, and HHMMs, and it offers more flexibility, functionality, and capacity than its predecessors. Moreover, we developed an effective MAP-based estimator to optimize the model structure and parameters simultaneously. Although the case of two channels is discussed, this work can be extended to the cases with more channels. We showed the usefulness of MCSHMM for exploring two kinds of semantic structures in American football videos. The proposed MCSHMM could also be applied to other video mining applications where multiple data sources are involved. Future research will focus on how to speed up the model learning and training and how to incorporate more semantic structures in the multi-channel framework.

### Acknowledgements

This work is supported by the National Science Foundation (NSF) under Grant IIS-0347613. The authors would like to thank the reviewers for their comments and suggestions.

### 6. REFERENCES