Abstract

We present a generative model approach to explore intrinsic semantic structures in sport videos, e.g., the camera view in American football games. We will invoke the concept of semantic space to explicitly define the semantic structure in the video in terms of latent states. A dynamic model is used to govern the transition between states, and an observation model is developed to characterize visual features pertaining to different states. Then the problem is formulated as a statistical inference process where we want to infer latent states (i.e., camera views) from observations (i.e., visual features). Two generative models, the hidden Markov model (HMM) and the Segmental HMM (SHMM), are involved in this research. In the HMM, both latent states and visual features are shot-based, and in the SHMM, latent states and visual features are defined for shots and frames respectively. Both models provide promising performance for view-based shot classification, and the SHMM outperforms the HMM by involving a two-layer observation model to accommodate the variability of visual features. This approach is also applicable to other video mining tasks.

1. Introduction

Recently, there is a growing interest in video mining research that is to discover knowledge and semantic structures in the video data stored either in databases, data warehouses, or other online repositories [10, 3]. This trend is mainly driven by the ever increasing need of numerous multimedia and online database applications. Major benefits of video mining range from efficient browsing and summarization of video content to facilitating video access and retrieval in a large database. One goal of video mining is to allow a viewer to access the video data like reading a book that has two main tools for readers, i.e., Contents and Index. The former one provides an overview, while the latter one allows the user to quickly find the information of interest. We believe that both tools could be delivered through video mining. In this work, we focus on sport video mining.

In sport videos, there are a variety of underlying rules that control the game procedure as well as video recording. Usually, cameras recording the broadcast football video are installed around the field to record the game from different view angles. Given an example of a football field in Fig. 1, the camera view of a shot reflects the play location, and the camera view transition between two consecutive shots will indicate the types of plays in the game, e.g., advancing the ball. As a fundamental semantic structure, view-based sport video analysis could provide a common platform for many other sport video mining tasks.

Figure 1. The OSU football field (www.okstate.com).
2. Related works

We will briefly review the recent works related to sport video mining. There are two major kinds of approaches: deterministic rule-based (data/feature driven bottom-up) methods [8, 16, 9] and statistical model-based (concept/knowledge driven top-down) methods [14, 15, 13].

Due to plenty sport rules and a given video recording setup, rule-based methods were proposed to represent semantic events in the sport video. In [8], camera motion related criteria are defined to classify each play shot into one of seven different plays. By evaluating ball trajectories, Yu et al. [16] 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 may not be able to deal with the uncertainty and ambiguity in the real video, and their task-specificity limits their flexibility.

Another trend of sport video mining is using statistical models to discover underlying semantic structures in sport videos, such as play/break or highlights. Descriptive models [17] and generative models [14, 15] are two typical statistical approaches. Compared with descriptive models, generative models involve a dynamic model that is defined on the latent states (usually with specific semantic meaning) and an observation model that characterizes visual features of different states, such as hidden Markov models (HMM). In [14], a soccer video is segmented into plays and breaks by involving a Hierarchical HMM (HHMM). In [15], a 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/mid-level features. However, the latent states of these models are usually not directly or explicitly associated with semantic structures in the video, or their semantic meaning has to be interpreted after model training and data classification.

HMMs are a popular probabilistic framework for modeling temporal series that have structures in time. However, one constraint of traditional HMMs is that the state and observation are usually on the same temporal scale. However, this may not be true for shot-based video analysis where frame-based features are used. In [7], a segmental hidden Markov model (SHMM) is proposed for speech signal processing, which involves a two-layer observation model to accommodate temporal dependencies between observations in the same state. Using the SHMM, the speech signal can be modeled as a sequence of phone level segments which are composed by corresponding frame level observations. In this work, the SHMM will be extended to view-based video mining where shot-based states and frame-based observations are involved in statistical inference.

3. HMMs and Segmental HMM

Our research involves both the traditional HMM and the SHMM, both of which will be introduced briefly in the following. We refer the reader to [11, 7] for more details.

3.1. Hidden Markov Models (HMMs)

In HMM, given the state sequence $S = \{S_i|i = 1, ..., T\}$ and the observation sequence $O = \{O_i|i = 1, ..., T\}$ as shown in Fig. 2, the model can be parameterized by

$$
\Gamma^* = \arg \max_{\Gamma} P(O_{1:T}|\Gamma),
$$

where $\Gamma = \{\pi_k, a_{k,j}, \Omega_k|k, j = 1, ..., K\}$, in which $K$ is the number of possible hidden states, $\pi_k = p(S_1 = k)$ is initial state probabilities vector, $a_{k,j} = p(S_i = j|S_{i-1} = k)$ is the state transition probabilities between states $i$ and $j$, $\Omega_k$ is the parameter set of the emission function for each possible hidden state, and in addition

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

Usually, the emission function $p(O_t|S_t = k)$ is specified as the Gaussian or the mixture of Gaussian functions, e.g.,

$$
p(O_t|S_t = k) = \mathcal{N}(O_t|\mu_k, \Sigma_k),
$$

or

$$
p(O_t|S_t = k) = \sum_{n=1}^{N} (a_{n}N(O_t|\mu_{n,k}, \Sigma_{n,k})).
$$

In HMM, observations are only dependent on current states, the temporal measurement of states and observations should be the same. For example, in video analysis, if we adapt shots as hidden states (semantic units), the corresponding observations should be shot-based too, otherwise the independence assumption could be invalid. In video semantic analysis, we want to utilize the frame-based feature as the observation, so that we can obtain more detailed information that describes the corresponding shot-based state. However, it is hard to establish the mapping relationship between multiple observation and one state in traditional HMM, thus a new approach, the SHMM, is employed in this paper to handle this problem.
3.2. Segmental HMM (SHMM)

As aforementioned, the independence assumption in HMMs could be invalid in some applications, so we are requested to consider the situation where a set of continuous observations are correlated. The segmental hidden Markov model (SHMM) is introduced in [7] in order to handle the speech modeling task. The structure of SHMM is shown in Fig. 3. Instead of generating one observation by each hidden state of the HMM, each hidden state of the SHMM can emit a sequence of observations, which can be called a segment. In SHMM all observations in a given segment are supposed to be independent to observations belonging to other segments. In addition, in each segment, all observation are conditionally independent on the mean of the segment, i.e., given the mean of the segment, all observations within a segment are independent.

Therefore, as shown in Fig. 3, the model can be characterized by

$$\Theta^* = \arg \max_{\Theta} P(O_{1:T} | \Theta),$$

(5)

where $\Theta = \{\pi_k, a_{i,k}, \mu_{\cdot,k}, \Sigma_{\cdot,k}, \Sigma_{\cdot,k} | k, j = 1, ..., K\}$. Different from HMM, we assume that for each shot-based state $S_i$, $i = 1, ..., T$, there are $\tau_i$ corresponding frame-based observations, i.e., $O = \{O_{i}^t | i = 1, ..., T; t = 1, ..., \tau_i\}$. Therefore, we use $\mu$ and $\Sigma_k$ to characterize the observation $O$ following Gaussian distribution, and use $\mu_{\cdot,k}$ and $\Sigma_{\cdot,k}$ to characterize the mean of the segment, i.e., $\mu$, following the Gaussian distribution. Then, given the mean $\mu$ of the segment, we can define

$$p(O_{i}^{1:\tau_i} | S_i = k, \Theta) = \int p(\mu | S_i = k, \Theta) p(O_{i}^{1:\tau_i} | \mu, S_i = k, \Theta) d\mu,$$

(6)

where $O_{i}^{1:\tau_i}$ represents all observations contained in the $i$th state. Then, under the conditionally independent assumption, we can rewrite it as

$$p(O_{i}^{1:\tau_i} | S_i = k, \Theta) = \int p(\mu | S_i = k, \Theta) \prod_{t=1}^{\tau_i} p(O_{i}^t | \mu, S_i = k, \Theta) d\mu.$$

(7)

In this work, $p(\mu | S_i = k, \Theta)$ and $p(O_{i}^t | \mu, S_i = k, \Theta)$ are both characterized by a Gaussian function:

$$p(\mu | S_i = k, \Theta) = \mathcal{N}(\mu_{\cdot,k}, \Sigma_{\cdot,k}) = \frac{1}{(2\pi)^{d/2} |\Sigma_{\cdot,k}|^{1/2}} \exp\left( -\frac{1}{2} (\mu - \mu_{\cdot,k}) \Sigma_{\cdot,k}^{-1} (\mu - \mu_{\cdot,k})' \right),$$

(8)

and

$$p(O_{i}^t | \mu, S_i = k, \Theta) = \mathcal{N}(O_{i}^t | \mu, \Sigma_k) = \frac{1}{(2\pi)^{d/2} |\Sigma_k|^{1/2}} \exp\left( -\frac{1}{2} (O_{i}^t - \mu) \Sigma_k^{-1} (O_{i}^t - \mu)' \right).$$

(9)

As illustrated in [7], in speech recognition, the author regards $\mu$ as the characteristic of the segment, i.e., the phone or model level of speech signals, which are variable due to different speakers or stress conditions. All observations in a segment are employed to characterize given $\mu$, i.e., the variation of a certain speaker due to other factors. It is easy to extend this type of generative procedure to semantic video analysis. For instance, in [4], shots belonging to the same semantic unit can show very different appearance, e.g., different locations in the same camera view, which can be described by different segments, and frame-by-frame observations can characterize the corresponding segment. Therefore, by SHMM, we can analyze the shot-based semantics based on frame-based observations, which can also provide us more detailed information about the semantic existing in a shot.

3.3. Comparison between two models

From the observation analysis above, we find the major difference between HMM and SHMM is that the variation $\Sigma_{\cdot,k}$ of the inter variable $\mu$. If we set $\Sigma_{\cdot,k} = 0$, then (7) can be reduced to

$$p(O_{i}^{1:\tau_i} | S_i = k, \Theta) = \prod_{t=1}^{\tau_i} p(O_{i}^t | \mu_{\cdot,k}, S_i = k, \Theta),$$

(10)

which is the same expression as the HMM with the Gaussian mixture emission which has the mean $\mu_{\cdot,k}$ and covariance $\Sigma_k$, therefore, we can regard the HMM with the the Gaussian mixture emission to be a sub-category of SHMM. We have the knowledge that the objective of regarding the emission function of HMM as Gaussian mixture function rather than Gaussian function is to enable the model have more capabilities to characterize variations existing in the observation. Therefore, compared with the HMM, the SHMM considers the dependency existing among observations belonging to the same segment. This approach provides us with a powerful tool to handle the situation where multiple observations correspond to one state, which describes the condition in video semantic analysis very well, i.e., capturing different characteristics existed in different frames which usually corresponds to the same semantic meaning contained in a shot.
4. Problem formulation

In this section, we will first introduce the concept of semantic space and corresponding low-level features existed in the video data, then we will demonstrate how to establish the mapping relationship between them via two stochastic approaches, i.e., the HMM and the SHMM.

4.1. Semantic space

Figure 5. The 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. 5, where the semantic units and their dynamic relationship are defined to characterize the semantic structure in a video. As stated in [12], a semantic unit should serve as a description of video segment with certain semantics, such as play and break. In a sport video, 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.

Given a football field as shown in Fig. 4, 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 sections, we will show the corresponding visual feature and the dynamic relationship between semantic units, i.e., the camera views, in this semantic space can be represented by different Markov chains.

4.2. View-based visual features

Figure 6. Typical feature distributions in four camera views.

As shown in Fig. 4, in an American football video, viewers can easily identify the camera view of each shot from the yard numbers along the yard line. However, it is hard for a computer to automatically do so due to the high complexity of video content and the occlusion problem during the play. Therefore, we need more robust and accessible visual features to characterize the semantic unit in the semantic space, i.e., different camera views.

In this research, we identify the spatial color distribution and the angle of yard lines to be relevant features. As typical scenes shown in Fig. 6, we can observe that the field color distribution and the yard line angle vary a lot from one view to another. For example, the color of the field, usually occupying the biggest portion of a frame, will change its position and portion in different views, e.g., from left to the right.
4.3. Visual features extraction

![Image](https://example.com/image1.png)

Figure 7. Visual features: a. The sample frame; b. Dominant color extraction; c. Canny detection; d. Yard line detection.

In this work, we use the dominant color to estimate the spatial color distribution by dividing a frame image into multiple regions and computing the ratios between different regions as shown in Fig. 7. Specifically, we use the Robust Dominant Color Region Detection algorithm in [6] to extract the dominant color region, i.e., the play ground. Fig. 7(b) 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 the playing field as shown in Fig. 7(d).

In order to estimate the color distribution, we segment a frame into five regions as shown in Fig. 7,b, each region occupies 1/3 area of the frame. The \( W \) is the total number of pixels in a frame, and \( W_f \) is the total number of pixels in the dominant color region. Moreover, \( W_{left}, W_{right}, W_{top}, W_{bottom}, W_{center} \) are the number of pixels of the dominant color region in the left, right, top, bottom and central part respectively. As the the angles of yard lines, we define \( \theta_i \) as the angle of the \( i \)th detected line and \( L \) is the number of detected lines. Then, based on the detected playing field and yard lines, we extract 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: \( D_{ave} = \sum_{i=1}^{L} \theta_i / L \);
- Angle difference between the first and last frames in a shot: \( D_{shot} = \theta_{last} - \theta_{first} \).

4.4. HMM-based video representation

In the HMM, states represent camera views following a dynamic relationship shown in Fig. 8. Observations are shot-based visual features which are computed by averaging the frame-wised features in a shot, and the state emission function are the GMM for each state. Given a series of shot-based observations \( \{O_t| t = 1, ..., T\} \), the HMM is parameterized as \( \Gamma = \{\pi_k, \alpha_k, \alpha_n, \mu_n, \Sigma_n| k = 1, ..., 4; n = 1, ..., N\} \):

1. \( S = k, k \in \{1, 2, 3, 4\} \): the hidden states represent the left, central, right, and end-zone camera views.
2. \( \pi_k = p(S_1 = k) \) and \( k \in S \): initial state probabilities vector, which represents the initial probabilities of four cameras;
3. \( \alpha_{kj} = p(S_t = j|S_{t-1} = k)| t = 1, \ldots, T; k, j \in S \) is the state (camera) transition probabilities between states (cameras) \( i \) and \( j \);
4. \( p(O_t| S_t = k) = \sum_{n=1}^{N} (\alpha_{nk} N(O_t, \mu_n, \Sigma_n)) \): is the \( N \)-order GMM emission function of hidden states.

Then we can use the Expectation Maximization (EM) algorithm [1] to obtain the maximum likelihood-based parameter estimate of this HMM. After the EM training, we can use the Viterbi decoding algorithm to estimate the optimal state sequence \( S^*_{1:T} \),

\[
S^*_{1:T} = \arg \max_{S_{1:T}} P(S_{1:T}|O_{1:T}, \Gamma^*), \quad (11)
\]

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

In HMM-based method, due to the mismatch between shot-level states and frame-level feature, we have to adapt the average feature in a shot to serve as the observation corresponding to the shot-based hidden state, which ignores both the statistical dependency and varieties among frames.

Therefore we will employ SHMM instead. In video analysis, we can regard the segment as the shot and observations in a segment as frames in that shot. Therefore we need to consider two kinds of independencies: first, segments are independent given the state; second, all observations in a segment are independent given the mean of the segment.

By these assumptions, we are able to establish a statistical relationship between multiple observations and a single semantic unit.

As introduced in [7], we assume the $p(o|s, \Gamma)$ and $p(O_i^t|\mu, s_i, \Gamma)$ are both characterized by Gaussian distributions. Following (8) and (9), we can rewrite (7) as

$$p(O_i^1|\tau_i|S_k = i, \Theta) = \int N(\mu|\mu_{i,k}, \Sigma_{i,k}) \prod_{t=1}^{\tau_i} N(O_i^t|\mu, \Sigma_k) d\mu. \quad (12)$$

This density function describes the probability distribution of a semantic unit which generates multiple observations, i.e., visual features distribution for a specific camera view within a shot with length $\tau_i$. Then following the description in [7], let

$$\mu_i = \sum_{t=1}^{\tau_i} O_i^t, \quad \Sigma^{-1}_{i,k} = \Sigma^{-1}_{\mu,k} + \tau_i \Sigma^{-1}_k, \quad \mu_{i,k} = (\Sigma_{i,k}(\Sigma^{-1}_{\mu,k}\mu_{\mu,k} + \Sigma^{-1}_k \mu_i))', \quad (13-15)$$

then using the log probability we can obtain

$$\log p(O_i^1|\tau_i|S_k = k, \Theta) = \log(\int R_k^T \mu_{i,k} R_k + 1 \Sigma_{i,k} R_k^T \mu_{i,k} + \frac{1}{2} \Sigma_{i,k}^{\tau_i} (O_i')' - \frac{1}{2} \mu_{i,k}^{\tau_i} (\mu_{i,k}')) + \frac{\tau_i}{2} \Sigma_{i,k}^{\tau_i} (O_i')' - \mu_{i,k}^{\tau_i} (\mu_{i,k}'). \quad (16)$$

where

$$R_k = \frac{1}{(2\pi)^\frac{N}{2} |\Sigma_k|^{-\frac{N}{2}}} \times R_k^T \mu_{i,k} = \frac{1}{(2\pi)^\frac{N}{2} |\Sigma_{\mu,k}|^{-\frac{N}{2}}},$$

and

$$R_i = \frac{1}{(2\pi)^\frac{N}{2} |\Sigma_{i,k}|^{-\frac{N}{2}}}.$$

Then we can employ the Backward-Forward algorithm to calculate the total probability summed over all possible paths. As the size of each segment, i.e., the length of each shot, is known to us, so we can represent the forward and backward processes as (17) - (20), rather than considering all the possible length of segment introduced in [7].

$$\alpha_k(i) = p(O_1, ..., O_i, S_i = k|\Theta), \quad (17)$$

$$\beta_k(i) = p(O_{t+1}, ..., O_T|S_i = k, \Theta), \quad (18)$$

$$\gamma_k(i) = \frac{\alpha_k(i)\beta_k(i)}{\sum_{k=1}^{K} \alpha_k(i)\beta_k(i)}. \quad (19)$$

$$\xi_{kj}(i) = \frac{\alpha_k(i) a_{kj} p(O_{i+1}|S_{i+1} = j, \Theta) \beta_j(i+1)}{\beta_k(i)}. \quad (20)$$

Then the SHMM parameters are estimated as follows:

$$\hat{\mu}_{\mu,k} = \sum_{t=1}^{T} \gamma_k(i) \frac{\mu_i}{\gamma_k(i)} / \sum_{t=1}^{T} \gamma_k(i), \quad (21)$$

$$\hat{\Sigma}_{i,k} = \sum_{t=1}^{T} \gamma_k(i) \frac{(\tau_i \mu_{\mu,k} - \mu_i)^2}{\gamma_k(i)}, \quad (22)$$

$$\hat{\Sigma}_k = \sum_{t=1}^{T} \gamma_k(i)(\sum_{t=1}^{T} \gamma_k(i) - 1) \gamma_k(i)(\tau_i - 1), \quad (23)$$

$$\hat{a}_{kj} = \frac{\xi_{kj}(i)}{\gamma_k(i)}, \quad \hat{\pi}_k = \gamma_k(1). \quad (24-25)$$

Comparing with the traditional HMM, we only add one additional parameter $\Sigma_k$ to each state in the SHMM. After the EM training, we can use the Viterbi algorithm to estimate the optimal state sequence $S_{1:T}$ as before.

Algorithm 1: EM for SHMM

Input: Iteration number $W$, the initial SHMM $\Pi_0$, the video $O = \{O_i^t|i = 1, ..., T; t = 1, ..., \tau_i\}$

Output: New Parameter set $\{\Pi_k|k = 1, 2, 3, 4\}$

for $w = 1$ to $W$ do

for $k = 1$ to $K$ do

E-step:

$p(O_i^1|\tau_i|S_i = k, \Theta) \Leftarrow (16)$;

$\alpha_k(i), \beta_k(i), \gamma_k(i), \xi_{kj}(i) \Leftarrow (17-20)$;

M-step:

$\hat{\mu}_{\mu,k}, \hat{\Sigma}_{i,k}, \hat{\Sigma}_k, \hat{a}_{kj}, \hat{\pi}_k \Leftarrow (21-25)$;

end

end
5. Experimental results

In this section, we will first demonstrate how to select salient visual features that provide discriminative observation models for the two generative models, i.e., HMM and SHMM; then we test the HMM and SHMM-based methods on two real American football videos in terms of their effectiveness for camera view-based video analysis.

5.1. Feature selection and evaluation

As aforementioned in the previous section, six candidate visual features are extracted to compose model observations. In HMM, the observation should be shot-based, but in SHMM, we need to employ the frame-based observation to serve as the observation. Therefore, selected features are a little bit different in each model due to the variation of the observation measurement.

Following the feature evaluation method mentioned in [4], we choose four salient visual features to construct the observation vector in HMM, i.e., $O_t$, including $D_{ave}$, $R_f$, $D_{shot}$, and $R_{lr}$. Because the feature $D_{shot}$ is designed for the shot-based approach, in SHMM, we employ the feature $R_d$, i.e., the ratio difference of dominant color between the left/right and center parts, instead of $D_{shot}$.

As shown in Fig. 9, we compare the $D_{ave}$, i.e., average angle of all yard lines, distribution in each camera view between shot-based and frame-based observations. We find that frame-wised features not only provide us the evidence about the statistical characteristics of the observation but also offer us richer and smoother information about the observation concerning different semantic meaning compared with shot-based features. Therefore we can establish a more powerful mapping relationship, e.g. SHMM, as the generative model between observation and latent states to discover the underlying semantic structure existing in the video data.

5.2. View-based shot classification

Our experiment was conducted on two 30-minute American NFL football videos. Videos A and B contain 156 shots and 171 shots, respectively, where each shot was manually classified into one of the four views for algorithm evaluation. Four methods were tested, i.e., K-mean, HMM with Gaussian emission (HMM-G), HMM with GMM emission (HMM-GMM) and SHMM, among which the first three methods use shot-based features and the SMM-based method involves frame-based features. In Table 1, we show the shot classification results of four different approaches.

<table>
<thead>
<tr>
<th>Videos</th>
<th>K-mean</th>
<th>HMM (Gaussian)</th>
<th>HMM (GMM)</th>
<th>SHMM</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>41.03%</td>
<td>55.77%</td>
<td>72.44%</td>
<td>79.49%</td>
</tr>
<tr>
<td></td>
<td>(64/156)</td>
<td>(87/156)</td>
<td>(113/156)</td>
<td>(124/156)</td>
</tr>
<tr>
<td>B</td>
<td>41.52%</td>
<td>62.57%</td>
<td>75.44%</td>
<td>80.11%</td>
</tr>
<tr>
<td></td>
<td>(71/171)</td>
<td>(107/171)</td>
<td>(129/171)</td>
<td>(137/171)</td>
</tr>
</tbody>
</table>

Table 1. The view-based shot classification accuracy.

For the HMM-based method, we use the K-mean clustering for model initialization, and $N = 3$ was found to be a reasonable choice for HMM-GMM. 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 $\pi_3 = \pi_4 = 0$. We also initialize the state transition matrix $A$ by estimating the frequency of actual state transitions and averaging over a few games. After K-mean initialization, we use the EM algorithm for HMM training with two different emissions: the Gaussian (HMM-G) and the GMM (HMM-GMM). 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 a single Gaussian.
For the SHMM-based method, we initialize model parameters $\Theta$ by the training result of HMM-G, which is mainly because the HMM-G can provide us a coarse estimation about the statistical property in each shot, therefore we can use them to represent the initialization of the SHMM. We initialize $\Sigma_k$ by multiplying $\Sigma_{\mu,k}$ with a small scalar, which is mainly because comparing with $\Sigma_{\mu,k}$, the $\Sigma_k$ only specify a small variation of observations. It is shown the SMM-based outperforms other methods in Table. 1. The boost in the performance is mainly because its capability to capture the diversity of the statistical property existing in each shot. In other words, the SHMM provides a two-layer observation model that can take advantage of the abundant frame-based observations and accommodate statistical variability of visual features.

6. Conclusions and discussions

We have presented a generative model based approach to explore the semantic structures in the sport video, i.e., the camera view in the American football videos. We have specified a semantic space for structured sport video representation where a dynamic model and an observation model are involved. The former one is defined on the latent states (i.e., camera views), and the latter one reflects the visual differences between across views. Then camera view based video analysis is formulated as a statistical inference problem. Three generative models are tested, i.e., HMM with Gaussian emission (HMM-G), HMM with GMM emission (HMM-GMM), and Segmental HMM (SHMM). The first two methods use shot-wise features, and the SHMM involves frame-wise ones. It is shown the SHMM outperforms the two HMM methods because its two-layer observation model is able to take advantage of rich frame-wise features. It is expected that the SHMM can be applied to other video mining tasks, e.g., play analysis [5].

In summary, we have addressed three interrelated issues for sport video mining. The first one is how to define a semantic space for structured video presentation where the semantic units (e.g., events of interest) directly correspond to the latent states. The second one is how to specify a dynamic model and an observation model by which we can infer the late states from the visual features. The third one is how to invoke powerful machine learning tools to fully take advantages of the two models for effective statistical inference. We treat this work as a case study to show that the importance of these three issues in video mining research.

Acknowledgements

The authors would like to thank the anonymous referees whose comments help us improve this paper. This work is supported by the National Science Foundation (NSF) under Grant IIS-0347613 (CAREER).

References