MULTIMODAL SEGMENTAL-BASED MODELING OF TENNIS VIDEO BROADCASTS

M. Delakis\textsuperscript{1}, G. Gravier\textsuperscript{2}, P. Gros\textsuperscript{2}

\textsuperscript{1}IRISA/University of Rennes 1, \textsuperscript{2}IRISA/CNRS
Campus de Beaulieu, 35042 Rennes Cedex, France

ABSTRACT

Efficient multimodal fusion is a key feature of future video indexing systems. Hidden Markov Models provide a powerful framework for video structure analysis but they require all video modalities to be strictly synchronous. Taking as a case study tennis broadcasts analysis, we introduce into video indexing Segment Models, a generalization of Hidden Markov Models, where the fusion of different modalities can be performed with relaxed synchrony constraints. Segment Models were experimentally proved to perform marginally better compared to Hidden Markov Models.

1. INTRODUCTION

Automatic annotation of video documents is a powerful tool for managing large video databases. In the last few years, modern computer vision techniques were employed for extracting semantic indexes based on the low-level features of a video. As video documents are inherently multimodal, it was quickly realized that an efficient indexing technique should take into consideration all the possible modalities (like images, audio, etc.). There are numerous approaches to multimodal fusion in the relative literature, reviewed in a recently published survey [1].

A statistical approach that is usually employed for modeling and information extraction is the Hidden Markov Models (HMMs) [2, 1]. The drawback of HMMs is that they require all the modalities of a video document to be completely synchronous before their fusion. Due to this constraint, a reference modality is usually chosen and then its segmentation is used to collect information from the other ones. This deficiency, however, of the non-native segmentation of the other modalities could be solved in another framework referred to as Segment Models. They were introduced in speech recognition by Ostendorf et al. [3] as a generalization of HMMs where different modeling assumptions can be easily incorporated. The purpose of this study is to introduce this promising framework into video indexing, providing extensions to previous work [4] based on HMMs. Our main application focuses on tennis broadcasts where game rules as well as production rules result in a structured document. Our aim is to recover this structure and then to construct the table of contents of the video by segmenting it in human meaningful scenes.

The paper is organized as follows. The feature extraction stage is briefly discussed in section 2. HMMs and Segment Models are reviewed in section 3. Multimodal integration under these models is discussed in section 4. Parameter estimation details and experimental results are given in section 5. Finally, section 6 concludes this study.

2. VISUAL AND AUDIO FEATURES

Both the video and audio tracks are characterized by large homogeneous segments. For the video track, these segments are the shots. For the sound tracks, we consider segments whose audio content is homogeneous with respect to sound classes such as ball hits or applause. In this section we discuss the extraction of a unique visual or audio descriptor from these segments. These descriptors (or observations in the HMM terminology) will serve as input features to the modeling stages of the following sections.

2.1. Visual Features

In order to detect hard cuts of the video track we implemented the adaptive threshold selection method of [5]. Starts and ends of replays are usually signaled by a smoothed progressive transitions between two shots, known as dissolve transitions, which were detected via the twin comparison algorithm [6]. Having the temporal extend of a dissolve, we formed a new type of shot labeled as “dissolve shot”.

We detected shots of exchanges between the two players (referred to as “global views”) using a simple color histogram-based distance between the middle frame of the given shot and a reference frame representing an “ideal” global view. This reference frame, different for each game, was found via an automated procedure as described in [4]. As a final result, we attached as visual descriptor to each key frame the vector $O_t = [O_{t}^{vs} \quad O_{t}^{l} \quad O_{t}^{diss}]^T$, where $O_{t}^{vs}$ is the visual similarity, $O_{t}^{l}$ is the length of the associated shot and $O_{t}^{diss}$ indicates a dissolve shot or not and $T$ denotes matrix
transposition. We quantized homogeneously the values of $O_t^{x}$ and $O_t^{j}$ into 10 bins each.

2.2. Audio Features

In order to characterize the content, we track the presence of the following key sound classes: music, applause, and ball hits. Tracking such events is carried out in a two step process as described in [7]. First, the soundtrack is segmented into homogeneous segments using a Bayesian information criterion. It is important to note that this segmentation is carried out independently of the shot segmentation. The presence or absence of sound classes is detected using statistical hypothesis testing with Gaussian mixture models.

3. MODELING OF THE VISUAL CONTENT

Our aim is to decode the tennis game according to some preidentified scenes, namely first missed serve and exchange, exchange, replay and break. The succession of these scenes is modeled by an ergodic HMM. In the first part of this section, we discuss how to model a scene also using an HMM (the resulting model also being an HMM), while in the second one we extend this approach to use segment models, where a segment corresponds to a scene.

3.1. Hidden Markov Models

One can easily observe that tennis videos exhibit strong temporal patterns. For example, a replay can be identified as a sequence of dissolves and non-dissolve shots. So, we can approach the video data as a sequence of observations, produced by a random process as it evolves through time.

After a careful examination of our video sequences, we have distinguished 12 different states for modeling the Markovian process, each of them having its special physical meaning, as illustrated in Fig. 1. We have separated them into four scenes corresponding to our four basic types of scenes mentioned above. The first scene can be modeled as follows: a first missed serve with a shot of global view (state 1), some shots of non-global view follow (state 2), a shot of global view of the normal exchange (state 3), and finally, some shots of non-global view after the exchange (state 4). There is also the possibility to transit from state 2 back to state 1 in cases of repeated missed serves. The states for the remaining scenes can be explained in a similar manner.

Assuming the parameters of the model are known, we can then decode an observation sequence to the corresponding most likely hidden state sequence, given by:

$$S^* = \arg \max_{s^*_1} \prod_{t=1}^{T} p(O_t^x|s^*_t)p(s^*_t)$$

where $s^*_t$ is the hidden state sequence, $O_t^x$ is the observation sequence and $T$ is the sequence length. The state sequence $S^*$ gives us the wanted human meaningful class labels of each video shot. This optimization problem is solved efficiently and fast using the Viterbi algorithm.

3.2. Segment Models

In this new type of modeling, the notion of the segment generalizes the notion of the state of HMMs in that it allows its extension to arbitrary durations. In this way a state can generate several observations before the transition into another state. This situation is depicted in Fig. 2. On the left, we see what happens conceptually in the case of HMMs: at a given time instant the process is in a given state and generates one observation symbol and then transits to another state. On the right, we see how a sequence is generated according to Segment Models. At a given time instant the stochastic process enters into a state and remains there according to a probability given by the segment duration model. A sequence of observations is generated, instead of a single one, according to a distribution conditioned on the segment label. Then the process transits to a new state with a transition probability, as in HMMs, and so on until the complete sequence of observations is generated.

In our tennis video case, we can think of a scene as a segment. Indeed, we can observe that the complete sets of observations of the scenes of Fig. 1 share a lot of common elements. For example, a scene of a break is an ensemble of shots of very short (commercials) or long (statistics) duration. In addition, we expect that all the break scenes will be of long absolute duration while the scenes of replays should be of short absolute duration.

The parameters to be estimated for Segment Models are the transition probability $p(i|j)$ from state $j$ to state $i$, the duration model $p(l|i)$ and the segment-level observation probability $b_s(O_1, O_2, \ldots, O_t)$, conditioned on the segment la-
bel a (in their general formalism of [3], it was also conditioned on the segment duration \( l \)). Details are given later in section 4. During our Viterbi search, we have now to find not only the most likely segment labels, but also the most likely segmentation or, in other words, the most likely duration of each segment. This new enhanced maximization problem can be formulated as:

\[
(L, A)^* = \arg\max_{l^N} p(O^T|l^N, a^N)p(l^N|a^N)p(a^N)
\]

where \( T \) is again the observation sequence length, \( N \) is the number of segments, \( a^N \) the segment labels and \( l^N \) the segment durations. This problem is solved via a straightforward extension of the Viterbi algorithm for HMMs with explicit state duration, described in [2]. To avoid unnecessary computation we restricted our search for possible segmentations into a window of 70 time steps (or shots), as it is difficult to have scenes containing more than 70 shots. This gives a computation cost of roughly 70 times higher than that of the HMM-based Viterbi algorithm, but it is still negligible compared to the cost of the feature extraction.

4. MULTIMODAL INTEGRATION

In section 3 the observation vector was limited to a single visual vector for sake of simplicity. The audio content however is an important source of information that should be taken into consideration in our modeling. For example, states 1, 3, and 5 of Fig. 1 are visually very similar as they correspond to the same global view type of shot. What can essentially differentiate the first state from the other two is the absence (state 1) or the presence (states 3 and 5) of applause after the exchange has finished.

In the HMM framework each state is strictly related to one and only observation symbol \( O_t \). As a consequence, HMMs allow very little flexibility regarding the fusion of multiple modalities: they should be artificially aligned and synchronized. A common approach is to choose a reference modality (the video track, in our case) and to concatenate to the observation vector, observations for the other modalities. In this manner, we collect information from the other sources not based on their native segmentation but in a direct way via the segmentation of the reference modality. The enhanced observation vector for the HMM is

\[
O_t = [O_t^{vs}, O_t^{vdis}, O_t^{bh}, O_t^{appl}, O_t^{mu}]^T
\]

where \( O_t^{vs}, O_t^{vdis} \) were defined in section 2, \( O_t^{bh} \) denotes the presence or absence of ball hits, \( O_t^{appl} \) of applause, and \( O_t^{mu} \) of music in the shot. We supposed again independence between all the components of the observation vector.

There are various ways to approach feature modeling in Segment Models. Generally, we can group these approaches based on the way they integrate the audio content: we can use it in the form of shot-based descriptors, as in HMMs, or with the form of scene-based features.

Starting from shot-based descriptors, the simplest case is to make the assumption of the independence of the observations:

\[
b_a(O_1 O_2 \ldots O_t) = \prod_{k=1}^t P(O_k|a),
\]

where \( a \) is the segment label. We refer to this approach as ‘AVprod’ from now on. We can relax the independence assumption by using an HMM to model the sequence of observations of a segment:

\[
b_a(O_1 O_2 \ldots O_t) \equiv P(O|\lambda_a) = \sum_Q P(O, Q|\lambda_a),
\]

(1)

where \( \lambda_a \) represents the HMM charged to model the observations of segment \( a \) and \( Q \) is a hidden state sequence of it. The calculation of the right term can be done easily by the forward pass of the forward-backward procedure [2]. We will call this approach ‘AVhmm’. When not using audio observations, we will refer to the ‘Vhmm’ approach.

As we can now model sets of observations at the scene level, we can describe the audio content using its native audio-based segmentation. So, instead of collecting a number of descriptors for each shot, we can use features like ‘presence of applause in the scene’, etc. The visual features are still modeled via HMMs as in eq. (1). We will call this approach ‘VhmmA1gram’. Another possibility is to use as features the succession of audio events in the segment, which can be done simply by a bigram modeling:

\[
b_a(O_1^a O_2^a \ldots O_t^a) = \prod_{k=2}^t P(O_k^a|O_{k-1}^a, a),
\]

where \( O_k^a \) is a symbol indicating the detection of applause, ball hits or music in the segment. We will call this approach ‘VhmmA2gram’.

5. PARAMETER ESTIMATION AND EXPERIMENTAL RESULTS

For all the models, parameters are estimated from manual shot and segment labels. The transition probabilities are es-
the performance increases significantly when modeling the sequential evolution of the observations of a segment. Indeed, AVprod gives strong evidence that we should model the temporal evolution of the data under various observation modeling alternatives. Firstly, we can integrate the video and audio content in an asynchronous way while achieving marginally better performance, as we see comparing VhmmA2gram to HMMs-AV and AVhmm.

6. CONCLUSIONS

We proposed an alternative modeling of a video sequence based on Segment Models, which can offer some flexibility regarding the fusion of multiple modalities compared to HMMs. The experimental results demonstrated that the asynchronous fusion of visual and audio observations under the Segment Models can give the same level of performance, if not better. We plan to extend this framework to other domains of sport video, as an alternative to HMMs.

7. REFERENCES


<table>
<thead>
<tr>
<th>Table 1. Experimental results for various approaches on test sets regarding percentage of correct classification (C), precision (P), and recall (R) rates.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
</tr>
<tr>
<td>----------------</td>
</tr>
<tr>
<td>HMMs-V</td>
</tr>
<tr>
<td>HMMs-AV</td>
</tr>
<tr>
<td>AVprod</td>
</tr>
<tr>
<td>Vhmm</td>
</tr>
<tr>
<td>VhmmA1gram</td>
</tr>
<tr>
<td>VhmmA2gram</td>
</tr>
<tr>
<td></td>
</tr>
</tbody>
</table>