Segmental-Based Modeling of Tennis Video Broadcasts

M. Delakis, G. Gravier, P. Gros

IRISA
Campus de Beaulieu, 35042 Rennes Cedex, France
{mdelakis, ggravier, pgros}@irisa.fr

Abstract

Automatic annotation of video documents is a powerful tool for managing large video databases. In this work we aim to describe tennis video broadcasts in a human meaningful way and thus to construct their table of contents. We present a generalization of Hidden Markov Models, the Segment Models, for modeling and extracting high-level information from a video sequence. Segment Models, even though they operate in an enhanced search space, were proved experimentally to perform marginally better compared to Hidden Markov Models.

1. Introduction

The management of the large multimedia databases of nowadays becomes practically intractable without the use of human meaningful semantic indexes describing the content of the documents. The automatic extraction of such a high level information from documents has attracked the interest of the research community in the last few years not only for this practical application, but also as an interesting machine learning problem. In this work, we aim to describe a complete video document via automatically extracted semantic indexes, and thus to costruct its table of contents. We are focused on tennis broadcast videos where the game rules as well as the work of the producer of it result in an structured document.

There are numerous approaches in the relative literature to the extraction of semantic labels of such structured documents, reviewed in a recently published survey [5]. A statistical approach that is usually employed for modeling and information extraction is the Hidden Markov Models (HMMs) [3]. Even though HMMs have been applied with success to speech recognition, the research community quickly realized that this type of modeling makes some rather strong and unrealistic assumptions on the nature of the speech signal [2]. Ostendorf *et al.* proposed there a generalized modeling of a stochastic process, refered to as *Segment Models*, 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 [1] based on HMMs.

The paper is organized as follows. The feature extraction stage is discussed briefly in section 2. Our baseline HMM-based system is presented in section 3. A brief introduction to Segment Models and experimental results are given in section 4. Section 5 concludes this study.

2. Visual and Audio Features

In order to detect hard cuts of the video track and based on the simple yet effective color histogram comparison, we implemented the adaptive threshold selection method of [6]. We also need to detect dissolve transitions between shots as they are frequently used to signal the start or the end of a rediffusion. We used a variation of the approach [7] based on edge features. Having the temporal extend of a dissolve, we formed a new type of shot labeled as shot corresponding to a dissolve transition. Finally, our 6 tennis games of 15 hours total duration were segmented into 12,402 shots with 1,392 of them corresponding to a dissolve.

One can notice easily that the shots of exchanges between the two players are characterized by their homogeneity in color space, where dominates the color of the court. So, in order to detect this type of shots (refered to as "global view"), we can use a color-based distance of each shot to a reference frame, representing an 'ideal' global view. Based on the dominant colors of each key frame, we extracted an initial list of candidates we applied the least median of squares method [4] to get the wanted prototype global view frame K_{ref}. To calculate the visual similarity of each key frame to K_{ref}, we used the simple bin wise distance of the two LUV histograms. As a final result, we attached as visual descriptor to each key frame the observations $O_t = [O_t^{vs}\ O_t^l\ O_t^{diss}]^T$, where O_t^{vs} is the visual similarity, O_t^l is the length of the associated shot and O_t^{diss} indicates if this shot corresponds to a dissolve transition or not. We descritized homogeneously the values of O_t^{vs} and O_t^l into 10 bins each to ease the calculations in sections 3 and 4.

Audio event detection is independently carried out in

a two-step process. First, the soundtrack is segmented into segments with homogeneous content. We utilized the Bayesian information criterion (BIC) using a cepstral representation of the input signal. The principle of the segmentation algorithm is to move two adjacent windows on the soundtrack signal. For each position of the windows, one can compare whether it is best, in terms of BIC, to model the two windows separately with two different Gaussian distributions or with a single one. Then, the detection of ball hits, applause and music is performed based on statistical hypothesis testing. More precisely, we used a two-hypothesis test, where H_0 (resp., H_1) is the hypothesis that the event considered is (resp., is not) present in the segment. Assuming a model for the distribution of y is available under both hypotheses, the decision on the presence of an event is taken by comparing the log-likelihood ratio $R(y) = \ln f(y; H_0) - \ln f(y; H_1)$ to a threshold δ, where $R(y) > \delta$ means that the event is detected in the segment y. In practice, H_0 (resp., H_1) corresponds to a Gaussian mixture model whose parameters were estimated from training data containing the event (resp., non-event) of interest. Finally, taking into consideration the video segmentation, we attached as audio descriptor to each key frame the observations $O_t = [O_t^{bh}\ O_t^{appl}\ O_t^m]^T$, where O_t^{bh} denotes the presence or absence of ball hits, O_t^{appl} of applause, and O_t^m of music in the associated shot. We used half of the games for parameter estimation and threshold evaluation for both visual and audio features and reserved the other half for testing.

3. Hidden Markov Modeling

Having carefully examined our tennis data, we observed the existence of some temporal patterns. For example, in an exchange we usually see a global view with relatively long duration. In parallel, the visual content contains ball hits followed by applauds. So, we can approach the video data (both visual and audio) as a sequence of observations, produced by a random process as it evolves through time. A first possible assumption we can make on this random process is that it can be modeled as a Hidden Markov process [3].

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. We can see them in fig. 1. We have seperated them into four groups corresponding to the four basic types of scenes: first missed serve and exchange, exchange, rediffusion and break. 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.

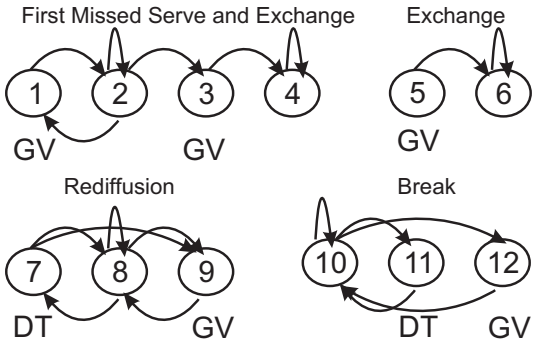

Figure 1: The 12 states of the HMM we used, grouped into four scenes. 'GV' stands for 'global view' and 'DT' for 'dissolve transition'. To make the presentation simpler, arcs interconnecting the four scenes are not shown.

We manually labeled all the shots with the respective state label for computing the ground truth of the video files. Having this ground truth, it was straightforward to estimate the parameters of the model by simply using relative frequencies. Regarding the emission probabilities $b_j(O_t)$ of state j, we supposed independecy between all the components of the observation vector O_t. We used half of the games for the estimation and reserved the other half for testing. The arcs between the states we see in fig. 1 give us the resulted dominant transition probabilities after training. Having estimated the parameters of the model, we can then *decode* an observation sequence to the corresponding most likely hidden state sequence, given by:

$$S^* = \arg\max_{s_1^T} p(O_1^T|s_1^T)p(s_1^T),$$

where s_1^T is the hidden state sequence, O_1^T is the observation sequence and T is the sequence length. The uncovered 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. We achieved 71.67% correct classification rate on the test files, while the perorformance decreases to 66.25% when not using audio features.

4. Segmental-Based Modeling

4.1. 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 arbritary durations. In this way a segment can emmit several observations before the transition into another segment. The 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

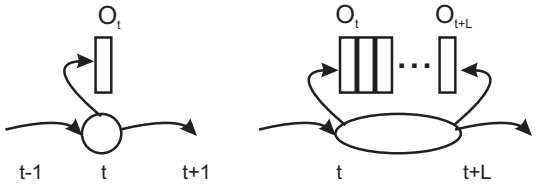

Figure 2: The generation of the observation sequence according to Hidden Markov Models (left) and to Segment Models (right).

emmits one observation symbol and then transits to another state. On the right we see the difference of Segment Models. At a given time instant the stochastic process enters into a segment and remains there according to a probability given by the segment duration modeling. The segment emits a train of observations, instead of a single one, according to a distribution conditioned on the segment label. Then the process transits to a new segment 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 (i.e., an ensemble of shots), as a segment. Indeed, we can observe that the complete sets of observations emmited by each scene of fig. 1 share a lot of common elements. For example, a scene of a break is an ensemble of observations of very small shots (commercials) or of long duration (statistics). Regarding the audio content and at the scene level, the audio observations of an exhange, for example, contain usually some ball hits followed by applause.

The parameters to be estimated for the segment models are the transition probability $p(i|j)$ from segment j to segment i, the duration modeling $p(l|a)$ of segment a, and the segment-level observation probability $b_a(O_1, O_2, \ldots, O_l)$, conditioned on the segment label a (in their general formalism of [2], it was also conditioned on the segment duration l). The first two probabilities were computed via the ground truth (easily adapted from shot-based labels to scene-based ones), as in the case of HMMs. On the contary, Segment Models offer a lot of possibilities of different modeling regarding the observation probabilities, which will be examined in detail later in section 4.2.

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:

$$(L, A)^* = \arg \max_{l_1^N, a_1^N} p(O_1^T|a_1^N)p(l_1^N|a_1^N)p(a_1^N),$$

where T is again the observation sequence length, N is the number of segments found, a_1^N is the segment labels and l_1^N is the segment durations. For avoiding 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, roughly speaking, a computation cost of 70 times higher than that of the HMM-based Viterbi alghorithm. But this computation cost is still of order of few seconds, which is clearly negligible compared to the computational cost of the feature extraction.

4.2. Feature Modeling

As we mentioned above, there are various ways to approach feature modeling in Segment Models. Starting from the simplest case, we will 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 $O_t = [O_t^{vs}\ O_t^l\ O_t^{diss}\ O_t^{bh}\ O_t^{appl}\ O_t^m]^T$ and a is the segment label. We will 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:

$$b_a(O_1 O_2 \ldots O_t) \equiv P(O|\lambda_a) = \sum_{Q} P(O, Q|\lambda_a),$$

where λ_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 probability of the right term can be done easily by the forward pass of the forward-backward procedure [3]. We will call this approach 'AVhmm'. When not using audio observation, we will refer to the 'Vhmm' approach.

As we can now model ensembles 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. We will call this approach 'VhmmA1gram' (which implies the use of HMMs for the video content). Another possibility is to use as features the succession of audio events in the segment, which can be done simply by a 2-gram modeling:

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

where O_t^{a} is a symbol indicating the detection of applause, ball hits or music in the segment (silence or other audio signals were discarded as irrelevant). We also used the special symbols 'start' and 'end' of the observation sequence. We will call this approach 'VhmmA2gram'.

4.3. Experimental Results

For estimating the probability products of the 'AVprod' approach, we used again relative frequencies thanks to the

Table 1: Experimental results for various approaches on test sets regarding percentage of correct classification (%), precision (P), and recall (R) rates.

	%	P	R
HMMs-V	70.72	68.90	80.51
HMMs-AV	74.57	73.69	82.51
AVprod	60.19	6.05	33.56
Vhmm	76.37	70.97	80.82
AVhmm	77.81	72.39	83.69
VhmmA1gram	76.95	72.28	72.47
VhmmA2gram	79.17	75.11	80.13
$p(l_1^N \mid a_1^N) = 1$	72.88	71.07	86.73
$p(a_1^N) = 1$	77.60	70.55	83.39
$p(s_1^T) = 1$	47.98	18.58	71.58

available ground truth. We initialized the topologies of the HMMs according to fig. 1 (i.e., same number of states and the allowed transitions were identical to the dominant transitions of the figure), with the exception of the last scene where we employed a 2-state HMM with an unconstrained topology. We used the standard Baum-Welch algorithm [3] to estimate the parameters. Regarding the bi-grams distributions, we noticed a light improvement in performance when using a simple back-patching scheme. We used half of the games for parameter estimation, and the other half for testing, as in sections 2 and 3. As performance measurements, we considered the percentage of shots assigned with the correct segment label and the precision and recall rates regarding the detection of the segment boundaries.

The average performance on the test sets is shown in table 1. We firstly see the performance of the HMM of section 3 without (HMMs-V) or with (HMMs-AV) audio observations, after transforming the shot labels to segment labels and detecting the points of scene boundaries. We see in the next five rows of table 1 the performance of Segment Models under various observation modeling alternatives. Firstly, it is clear that the observation independence assumption gives very poor results (approach AVprod). On the contrary, it is clear that the performance increases significantly when modeling the observation distributions via an HMM (cases Vhmm and AVhmm). Comparing the performance of Vhmm to that of AVhmm, we see that the audio observations are again usefull for Segment Models (or more precisely, for the HMMs that model the observation sequences). This is not clear when comparing Vhmm with VhmmA1gram, where we used audio observations based on their native segmentation. But the performance does improve when modeling the audio observations via 2-grams models (VhmmA2gram). Overall, it is clear that Segment Models give results of the same performance compared to HMMs, if not better.

In the following rows we see the performance of Segment Models (VhmmA2gram) when removing from the Viterbi decoding the duration modeling ($p(l_1^N|a_1^N) = 1$) or the segment transition probabilities ($p(a_1^N) = 1$). We notice that the performance of the system decreases very little, especially for the case of transitions probabilities. This should be contrasted to the corresponding performance of HMMs (HMMs-AV) when setting the transition probabilities equiprobable ($p(s_1^T) = 1$), where the performance cleary deteriorates as we see in the last row of table 1. Generally, we see clearly that Segment Models rely basically to the observations and not to the content (transition probabilities), as HMMs do.

5. Conclusions

In this study, we proposed an alternative modeling of a video sequence based on Segment Models, where the problem of finding semantic labels is coupled with that of finding their extend in time. The experimental results demonstrated that Segment Models with some straightforward solutions regarding the feature modeling can perform margninally better compared to HMMs. In addition, Segment Models do not require fine tuning of the state labels according to the producer's style, a time consuming and erroneous preliminary step for the HMMs of section 3. Encouraged by their generalized nature, we plan to extend this work to other domains of sport video.

References

[1] E. Kizak, G. Gravier, L. Oisel, and P. Gros. Audiovisual integration for sport broadcast structuring. In *Den to jeroume pros stigmhn*, 2003.

[2] M. Ostendorf, V. Digalakis, and O. Kimball. From HMMs to segment models: A unified view of stochastic modeling for speech recognition. *IEEE Transactions on Speech and Audio Processing*, 4(5):360–378, 1996.

[3] L. Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–285, 1989.

[4] P. J. Rousseeuw and A. M. Leroy. *Robust Regression and Outlier Detection*. John Wiley & Sons, 1987.

[5] C. Snoek and M. Worring. Multimodal video indexing: A review of the state-of-the-art. *Multimedia Tools and Applications*, 25(1):5–35, 2005.

[6] B.T. Truong, C. Dorai, and S. Venkatesh. New enhancements to cut, fade, and dissolve detection processes in video segmentation. In *Proc. ACM on Multimedia*, pages 219–227, 2000.

[7] R. Zabih, J. Miller, and K. Mai. A feature-based algorithm for detecting and classifying scene breaks. In *Proc. ACM on Multimedia*, pages 189–200, 1995.