HMM BASED STRUCTURING OF TENNIS VIDEOS USING VISUAL AND AUDIO CUES

E. Kijak [1,2], G. Gravier [2], P. Gros [2], L. Oisel [1] and F. Bimbot [2]

[1] Thomson multimedia R&D, France
1, av. de Belle-Fontaine
35510 Cesson-Sevigne, France
{ewa.kijak, lionel.oisel}@thomson.net

[2] IRISA
Campus Universitaire de Beaulieu
35042 Rennes Cedex, France
{ggravier, pgros, bimbot}@irisa.fr

ABSTRACT

This paper focuses on the use of Hidden Markov Models (HMMs) for structure analysis of videos, and demonstrates how they can be efficiently applied to merge audio and visual cues. Our approach is validated in the particular domain of tennis videos. The basic temporal unit is the video shot. Visual features are used to characterize the type of shot view. Audio features describe the audio events within a video shot. The video structure parsing relies on the analysis of the temporal interleaving of video shots, with respect to prior information about tennis content and editing rules. As a result, typical tennis scenes are identified. In addition, each shot is assigned to a level in the hierarchy described in terms of point, game and set.

1. INTRODUCTION

Video structure parsing consists in extracting logical story units from the considered video. It's a mandatory step to efficiently organize and retrieve video contents. The structure to be estimated relies on the nature of the video. In this paper, we address the problem of recovering sport video structure that can be useful in a further indexing task. Sport video analysis is motivated by the broadcasters needs of a detailed annotation of video contents. This annotation is usually used to select relevant excerpts in order to easily build summaries or magazines. Up to now, this logging task is performed manually.

Broadcasted video data consist of several multimodal information streams. Numerous approaches presented in the literature have discussed how to extract high-level semantic events from video. To achieve such a goal, algorithms have to be dedicated to one particular type of videos such as sports or news. In this context, most of the approaches use the individual visual or audio cues. Only a few attempts dealing with multimodal analysis of audio-visual information have appeared recently, such as [1] on sports, or [2] on news. Theses approaches rely on a successive use of visual and audio features. In a first step, visual features are used to detect plays in the original flow. In a second step, a detection of audience or commentator excitement is used to select the most exciting plays [3]. The reverse scheme has also been proposed in [4]: cheering sounds are first extracted to locate an important event. Visual analysis is then performed in cheering sounds regions to recognize the event.

In this paper, we propose a tennis scene classification and segmentation method that automatically labels each segment of a tennis video with one of the following four types: *first missed serve*, *rally*, *replay*, and *break*. Hidden Markov Models (HMMs) are used to merge audio-visual informations, and to represent the hierarchical structure of a tennis match. Domain knowledge is implicitly taken into account through the learning process associated with HMMs. This implicit approach is more suitable than heuristic rules oriented schemes because the learning process can derive more complete knowledge. In addition, previous works that use HMMs for video parsing or segmentation suggest good potential, even if only visual features are considered in sport domain [5, 6].

Our solution differs from existing methods in the following main points: (1) it combines visual/audio cues simultaneously; (2) different levels of classifications are generated over the entire tennis video at different levels; and (3), the segmented content are semantically meaningful to users.

This paper is organized as follows. Section 2 gives an overview of the system and briefly describes the exploited audio and visual features. Section 3 introduces the structure analysis mechanism. Experimental results are presented and discussed in section 4. Section 5 provides the concluding remarks.

2. SYSTEM OVERVIEW

In this section, we briefly describe the extraction of audio and visual cues. First, the video stream is automatically segmented into shots by detecting cuts and dissolve transitions. For each shot, we extract from the image stream, one

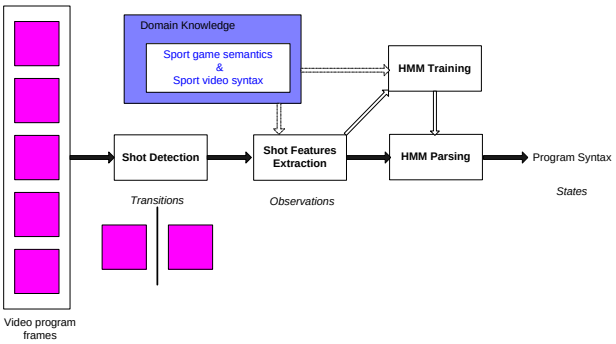

Fig. 1. Structure analysis system overview

keyframe along with its image features (color and motion intensity), and from the audio stream, sound classes such as ball hits, applause, or speech. The segmented video results in an observation sequence which is parsed by a HMM process (Figure 1). As a result, each shot is assigned to a level in the hierarchy described in terms of point, game and set, and is labelled with one of these four types: *first missed serve*, *rally*, *replay*, and *break*.

Such a tennis scene classification and segmentation is possible because tennis video is characterized by a typical production style, known as the tennis video syntax.

2.1. Tennis Syntax

Sport video production is characterized by the use of a limited number of cameras at almost fixed positions. Considering a given instant, the point of view giving the most relevant information is selected by the producer, and broadcast. Therefore sports are composed of a restricted number of typical scenes producing a repetitive pattern. For example, during a rally in a tennis video, the content provided by the camera filming the whole court is selected (we name it global view in the following). After the end of the rally, the player who has just carried out an action of interest is captured with a close-up. As close-up views never appear during a rally but right after or before it, global views are generally significant of a rally. Another example consists in replays that are notified to the viewers by inserting special transitions.

We identify here four typical pattern in tennis videos that are: *first missed serve*, *rally*, *replay*, and *break*. The *break* class includes the scenes unrelated to games, such as commercials. A break is characterized by an important succession of such scenes. It appears when players change ends, generally every two games. We also take advantage of the well-defined and strong structure of tennis broadcast. On the contrary of time-constrained sports that have a relatively loose structure, tennis is a score-constrained sport that can be broken down into sets, games and points (Figure 2).

The different types of views present in a tennis video can be divided into four principal classes: global, medium,

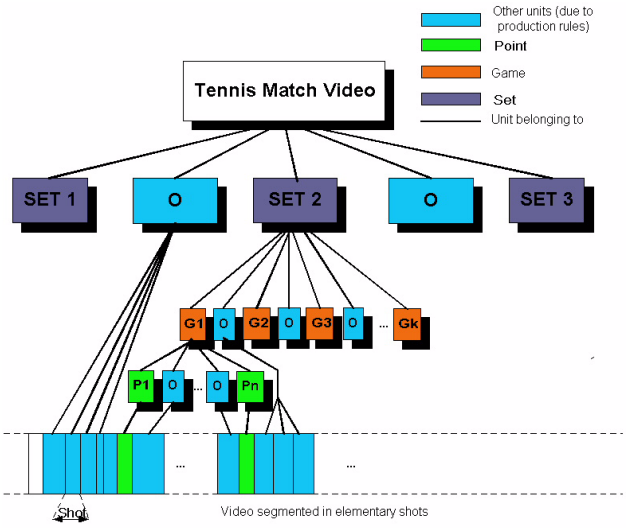

Fig. 2. Structure of Tennis Game

close-up, and audience. In a tennis video production, global views contain much of the pertinent information. The remaining information relies on the presence or the absence of non global views but is independent of the type of these views.

2.2. Visual Features

We used visual features to identify the global views within all the extracted keyframes. The process can be divided into two steps. First, a keyframe K_{ref} representative of a global view is selected without making any assumption about the playing area color. Once K_{ref} has been found, each keyframe is characterized by a similarity distance to K_{ref}. In the following, we motivate the choice of the visual features we use, before briefly describing the identification process.

Global views are characterized by a rather homogeneous color content (the colors of the court and its surrounding), although medium and audience views are characterized by scattered color content. In addition, a global view must capture at each time the main part of the court, whereas in close-up views, the camera is generally tracking the player. The first class can thus be characterized by a small camera motion while the second implies important camera translations. Based on these observations, we choose two features to identify global views: (1) a vector of dominant colors *F* and its spatial coherency *C*, and (2) activity *A* that reflects camera motion during a shot.

The visual similarity measure v(K_1, K_2) between two keyframes K_1 and K_2 is defined as a weighted function of the spatial coherency, the distance function between the dominant color vectors, and the activity:

$$v(K_1, K_2) = \\ w_1 \, |C_1 - C_2| + w_1 \, d(F_1, F_2) + w_3 \, |A_1 - A_2| \quad (1)$$

where w_1, w_2, and w_3 are the weights.

To find a keyframe K_{ref} representative of a global view, we consider dominant colors ratios. In a reduced subset of keyframes whose highest percentage of dominant color is more than 50%, we apply the least median square method, to minimize the median distance of all the remaining keyframes to a randomly selected keyframe. This leads to obtain K_{ref} (see details in [7]). The visual similarity measure v(K_t, K_{ref}), denoted by v_t in the following, is then computed between each keyframe K_t and K_{ref}.

2.3. Audio Features

Generally speaking, there are two approaches to combine audio and visual features for structure analysis: one is to combine audio and visual features into a single audiovisual feature vector before the classification. The other consists in classifying separately according to each modality before integrating the classification results.

In this work, an intermediate strategy is used which consists in extracting separately "high level" audio and visual cues, the classification being made using the audio and visual cues simultaneously. As mentioned previously, the video stream is segmented in a sequence of shots. Since shot boundaries are more suitable for structure analysis than boundaries extracted from the soundtrack, the shot is considered as the base entity and features describing the audio content for each shot are used to provide additional information. For each shot, a binary vector a_t describing which audio classes, among *speech*, *applause*, *ball hits*, *noise* and *music*, are present in the shot is extracted from an automatic segmentation of the audio stream. The automatic soundtrack segmentation is carried out using a Viterbi based system with Gaussian mixtures for each sound class and combination of sound classes. The soundtrack is segmented independently and the vectors a_t are created from this independent segmentation.

3. STRUCTURE ANALYSIS

Prior information is integrated by deriving syntactical basic elements from the tennis video syntax. We define four basic structural units: two of them are related to game phases (first missed serves and rallies), the two others denote non-game phases (breaks and replays). Each of these units is modelled by a HMM. These HMMs rely on the temporal relationships between shots.

Each HMM state models either a single shot or a dissolve transition between shots. The observation consists of three elements: the visual similarity v_t between the shot keyframe and K_{ref}, the shot duration d_t, and the audio vector a_t which characterizes the presence or absence of the predetermined audio events. More formally, the probability of an observation o_t conditionally to state j is then given by

$$b_j(o_t) = p(v_t|j)\ \ p(d_t|j)\ \ P[a_t|j] \qquad (2)$$

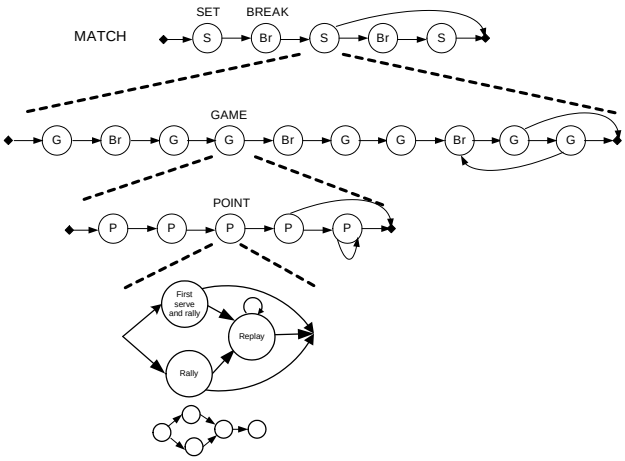

Fig. 3. Content hierarchy of broadcast tennis video

where $p(v_t|j)$ and $p(d_t|j)$ are given by a smoothed histogram, and $P[a_t|j]$ is the product over each sound class k of the discrete probability $P[a_t(k)|j]$.

Segmentation and classification of the whole observed sequence $o = o_1 o_2 ... o_T$ into the different structural elements are performed simultaneously using a Viterbi algorithm. The most likely state sequence $s = s_1 ... s_T$ is given by

$$\hat{s} \quad = \quad \arg\max_{s} \ \ \ln p(s) \ \ + \ \ \sum_{t} \ln b_{s_t}(o_t) \qquad (3)$$

To take into account the long-term structure of a tennis game, the four HMMs are connected to a higher-level HMM which represents the tennis syntax as illustrated in Figure 3. The transition probabilities between states of the higher-level HMM result entirely from a priori information, while the transition probabilities of the sub-HMMs are estimated.

4. EXPERIMENTAL RESULTS

Experimental data are composed of about 2 hours of manually labelled tennis video. About half of the data is used to train the HMM while the remaining half is reserved for the tests.

Recall and precision rates are given in Table 1 for three experiments: visual features only, audio features only and combined audiovisual approach. Shot duration is considered for all experiments. Experiments using audio features are sub-divided into two parts: using audio features manually annotated (denote by "man." in Table 1) to validate the model, and using audio features extracting from automatic segmentation (denote by "segm." in Table 1). The audio probabilities are estimated from the manually annotated audio features.

	Visual features		Audio features				Audio-visual			
			man.		segm.		man.		segm.	
Segmentation accuracy	71%		68%		44%		78%		66%	
Classification	precision	recall	precision	recall	precision	recall	precision	recall	precision	recall
First serve	67%	63%	92%	66%	66%	5%	87%	75%	71%	41%
Rallies	97%	54%	95%	63%	62%	28%	96%	83%	82%	65%
Replay	100%	80%	68%	38%	75%	15%	81%	88%	86%	82%
Break	51%	93%	91%	81%	30%	90%	94%	81%	68%	92%

Table 1. Classification and segmentation results with visual features only, audio features only and both audio-visual features

We observed that one single stream cannot provide enough information to identify different scenes. Rallies high precision rate (97%) and breaks recall rate (93%) show that low level visual features are efficient to identify global views from others. However, these features are not sufficient to distinguish a rally from a simple view of the court or from a first serve, as indicate by rallies low recall rate (54%). This means also that many global views are missed using only visual features. Replays detection relies essentially on dissolve transitions identification. The high precision rate corresponds to correctly detected dissolve transitions.

Classification using reliable audio features increases precision rates for first missed serve. Comparable results are obtained for rallies detection. This suggests that audio features are effective to describe rally scenes. Indeed, a rally is essentially characterized by the presence of *ball hits* and *applause* which happen at the end of the exchange, although a missed first serve is only characterized by the presence of *ball hits*. Therefore, there are less confusion between missed first serves and rallies. Breaks precision rate increases also. Since break is the only state characterized by the presence of *music*, the break state works in this case like a commercial detector, and avoids false detections. On the contrary, replays are not characterized by a representative audio content, and almost all replays are missed.

The approach using only audio features automatically segmented gives poor performance. Recall rate for audio segmentation as described in 2.3 is 76.9%, and precision is 46.3%. *Ball hits* and *music* are well classified, whereas *noise* is often classified as *ball hits*. The errors from the automatic audio segmentation spread over the structuration process and drag the performance down. Another source of errors is the mismatch between the manually segmented audio features on which the HMM parameters were estimated, and the automatically generated audio features. An improvement was obtained by estimating the HMM parameters using automatically generated audio features, but it did not outclass the video only system.

Nevertheless, integrating multiple cues from different media improves performance. In particular, *ball hits* detection improves classification rates for rallies and first serves. And *music* detection prevents from false detection of breaks.

5. CONCLUSION

In this paper, we have presented an audio-visual model for tennis video structure parsing. Tennis structure is modelled by an HMM that integrates a priori information about tennis content and editing rules. HMMs provide an efficient framework to merge audio and visual cues, and to perform simultaneously classification and segmentation into scenes. Experiments show that audio information improve the performance of video classification, when features are quite reliable. Indeed errors from feature extraction spread over the entire analysis process. It is therefore necessary to enhance audio segmentation and classification.

To extend our tennis video structure analysis, we also intend to include a more sophisticated audio-video fusion model by refining audio and video content analysis.

6. REFERENCES

[1] W. Hua, M. Han, and Y. Gong, "Baseball scene classification using multimedia features," in *Proc. of Int'l Conf. on Multimedia and Expo*, 2002.

[2] W. Qi, L. Gu, H. Jiang, X-R. Chen, and H-J. Zhang, "Integrating visual, audio and text analysis for news video," in *Proc. of Int'l Conf. on Image Processing*, 2000.

[3] B. Li and I. Sezan, "Event detection and summarization in american football broadcast video," in *SPIE Storage and Retrieval for Media Databases*, v. 4676, pp. 202–213, 2002.

[4] K. Kim, J. Choi, N. Kim, and P. Kim, "Extracting semantic information from basketball video based on audio-visual features," in *Proc. of Int'l Conf. on Image and Video Retrieval*, vol. 2383, pp. 278–288, 2002.

[5] J. Assfalg, M. Bertini, A. Del Bimbo, W. Nunziati, and P. Pala, "Soccer highlights detection and recognition using hmms," in *Proc. of Int'l Conf. on Multimedia and Expo*, 2002.

[6] P. Chang, M. Han, and Y. Gong, "Extract highlights from baseball game video with hidden markov models," in *Proc. of Int'l Conf. on Image Processing*, 2002.

[7] E. Kijak, L. Oisel, and P. Gros, "Temporal structure analysis of broadcast tennis video using hidden markov models," in *SPIE Storage and Retrieval for Media Databases*, v. 5021, pp. 289–299, 2003.