AUTOMATIC VIDEO STRUCTURING BASED ON HMMS AND AUDIO VISUAL INTEGRATION

P. Gros[(1)]*, E. Kijak*[(2)] *and G. Gravier*[(1)]

[(1)] IRISA – CNRS [(2)] IRISA – Université de Rennes 1
Campus Universitaire de Beaulieu
35042 Rennes Cedex, France
{Patrick.Gros, Ewa.Kijak, Guillaume.Gravier}@irisa.fr

ABSTRACT

This paper focuses on the use of Hidden Markov Models (HMMs) for structure analysis of sport videos. The video structure parsing relies on the analysis of the temporal interleaving of video shots, with respect to a priori information about video content and editing rules. The basic temporal unit is the video shot and both audio and visual features are used to characterize its type of view. Our approach is validated in the particular domain of tennis videos. 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 content analysis is an active research domain that aims at automatically extracting high-level semantic events from video. This semantic information can be used to produce indexes or tables-of-contents that enable efficient search and browsing of video content. Low-level visual features, largely used for indexing generic video contents, are not sufficient to provide a meaningful information to an end-user. To achieve such a goal, algorithms have to be dedicated to one particular type of videos.

One domain-specific application is the detection and recognition of highlights in sport videos. Sport video analysis is motivated by the growing amount of archived sport video material, and by the broadcasters needs of a detailed annotation of video contents to select relevant excerpts to be edited for summaries or magazines. Up to now, this logging task is performed manually by librarians.

Most of the existing works in the domain of sports video analysis are related to specific events detection. A common approach in event detection consists in combining the extraction of low-level features with heuristic rules to infer predetermined highlights [1, 2]. Recent approaches use Hidden Markov Models (HMMs) for the event classification in soccer [3] and baseball [4]. Nevertheless, these works attempt to detect specific events, but the reconstruction of higher-level temporal structure is not addressed.

Inside the category of sport videos, a distinction should be made between time-constrained sports such as soccer, and score-constrained sports such as tennis or baseball. Time-constrained sports have a relatively loose structure. The game can be decomposed into equal periods. During a period, the content flow is quite unpredictable. As a result, structure analysis of a soccer video is restricted to ”play”/”out-of-play” segmentation [5].

In score-constrained sports, the content presents a strong hierarchical structure. For example, a tennis match can be broken down into sets, games and points. A previous work on tennis and baseball [6] studies the detection of basic units, such as serve in tennis or pitch in baseball, however, the well-defined structure of these sports is not taken into account to recover the whole hierarchical structure.

In this paper, we address the problem of recovering sport video structure, through the example of tennis which presents a strong structure. Video structure parsing consists in extracting logical story units from the considered video. The structure to be estimated relies on the nature of the video. For instance, a news video can be considered as a sequence of units which starts with an image frame presenting an anchor person followed by a variety of news and commercials [7].

Producing a table of contents implies to perform a temporal segmentation of the video into shots. Such a task has been widely studied and generally relies on the detection of discontinuities into low-level visual features such as color or motion [8]. The critical step is to automatically group shots into ”scenes”, or “story units” that are defined as a coherent group of shots that is meaningful for the end-user.

Recent efforts have been made on fusing information provided by different streams. It seems reasonable to think that integrating several media improve the performance of the analysis. This is confirmed by some existing works reported in [9, 10]. Multimodal approaches have been inves-

tigated for different areas of content-based analysis, such as scene boundary detection [11], structure analysis of news [12], and genre classification [13]. However, fusing multimodal features is not a trivial task. We can highlight two problems among many others.

- a synchronization and time scale problem: sampling rate to compute and analyse low-level features is not the same for the different medias;
- a decision problem: what should be the final decision when the different medias provide opposite information ?

Multimodal fusion can be performed at two levels: feature and decision levels. At the feature level, low-level audio and visual features are combined into a single audio-visual feature vector before the classification. The multimodal features have to be synchronized [12]. This early integration strategy is computationally costly due to the size of typical feature spaces. At the decision level, a common approach consists in classifying separately according to each modality before integrating the classification results. However, some dependencies among features from different modalities are not taken into account in this late integration scheme.

But usually theses approaches rely on a successive use of visual and audio classification [14, 15]. For example in [15], visual features are first used to identify the court views of a tennis video. Then ball hits, silence, applause, and speech are detected in these shots. The analysis of the sound transition pattern finally allows to refine the model, and identify specific events like scores, reserves, aces, serves and returns.

In this work, an intermediate strategy is used which consists in extracting separately shot-based "high level" audio and visual cues. The classification is then made using the audio and visual cues simultaneously (Figure 2). In other words, we choose a transitional level between decision and feature levels. Before analyzing shots from raw image intensity and audio data, some preliminary decisions can be made using the features of the data (e.g. representation of audio features in terms of classes like music, noise, silence, speech, and applause). In this way, after making some basic decisions, the feature space size is reduced and each modality can be combined more easily.

Our aim is to exploit multimodal information and temporal relations between shots in order to identify the global structure. The proposed method simultaneously performs a scene classification and segmentation using HMMs. HMMs provide an efficient way to integrate features from different media [13], and to represent the hierarchical structure of a tennis match. At the first level, several consecutive shots of a tennis video are classified within one of the following four predefined tennis units : *missed first serve*, *rally*, *replay*, and *break*. At the higher level, the classified segments are grouped and assigned to their corresponding label in the structure hierarchy, described in terms of point, game, and set (see Figure 1).

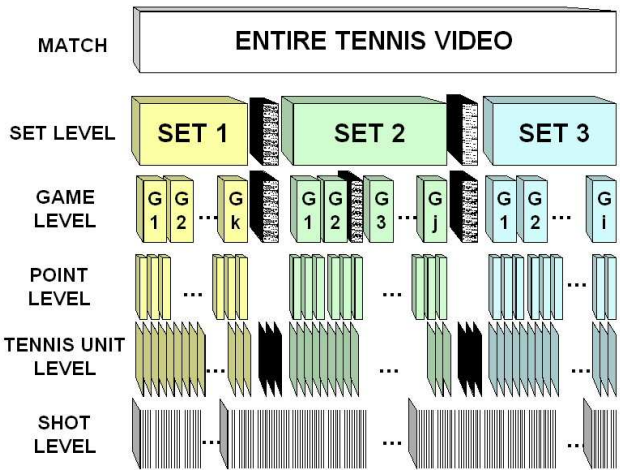

Fig. 1. Structure of Tennis Game

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

2. TENNIS SYNTAX

Sport video production is characterized by the use of a limited number of cameras at almost fixed positions. The different types of views present in a tennis video can be divided into four principal classes: global, medium, 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.

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 (global view). 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. Because of the presence of typical scenes and the finite number of views, the tennis video has a predictable temporal syntax.

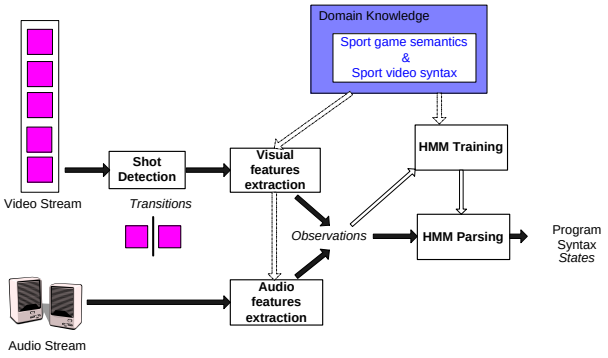

Fig. 2. Structure analysis system overview

We identify here four typical pattern in tennis videos, called tennis units, that are: *missed first serve, rally, replay,* and *break*. A break is characterized by an important succession of scenes unrelated to games, such as commercials or close-ups. It appears when players change ends, generally every two games. We also take advantage of the well-defined and strong structure of tennis broadcast to break it down into sets, games and points.

3. SYSTEM OVERVIEW

In this section, we give an overview of the system (Figure 2) and briefly describe the extraction of visual features. First, the video stream is automatically segmented into shots by detecting cuts and dissolve transitions [16]. For each shot, shot features are computed and one keyframe is extracted from the beginning of the shot along with its image features. The segmented video results in an observation sequence which is parsed by a HMM process.

3.1. Visual Features

The features we currently use are shot length, a visual similarity based on dominant colors desciptor, and relative player position.

Shot length *l*: the shot length is given by the shot segmentation process. It is the number of frames within the shot.

Visual similarity *v*: 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 automatically without making any assumption about the playing area color. Our approach tries to avoid the use of predefined field color as the game field color can largely vary from one video to another. Once K_{ref} has been found, each keyframe K_t is characterized by a similarity distance to K_{ref}. The visual similarity measure v(K_t, K_{ref}) is defined as a weighted function of the spatial coherency, the distance function between the dominant color vectors, and the activity (average camera motion during a shot):

$$v(K_t, K_{ref}) = \quad (1)$$
$$w_1 \, |C_t - C_{ref}| + w_1 \, d(F_t, F_{ref}) + w_3 \, |A_t - A_{ref}|$$

where w_1, w_2, and w_3 are the weights.

Player position *d*: the player position is given by the gravity center of a raw segmentation of the player (also named blob). As it is a time-consuming process, this feature extraction is only performed on potential global view keyframes. It uses domain-knowledge on the tennis court model and associated potential player positions. Only the player on the bottom of the court is considered to ensure a more reliable detection. The blob corresponding to this player is given by an image filtering and segmentation process based on dominant colors. Tennis court lines are also detected by a Hough-transform and the half-court line, i.e. the line separating the left and right halves of the court, is identified. The distance ***d*** between the detected blob and the half-court line is computed. If the extraction process fails, this feature is not taken into account for the considered shot.

3.2. Audio Features

As mentioned previously, the video stream is segmented into a sequence of shots. Since shot boundaries are more suitable for a structure analysis based on production rules 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 events, among *speech, applause, ball hits, noise* and *music*, are present in the shot is extracted from an automatic segmentation of the audio stream.

The soundtrack is first segmented into spectrally homogeneous chunks. For each chunk, tests are performed independently for each of the audio events considered in order to determine which events are present. More details can be found in [17]. Using this approach, the frame correct classification rate obtained is 77.83% while the total frame classification error rate is 34.41% due to insertions. The confusion matrix shows that *ball hits, speech* and *applause* are well classified while *noise* is often misclassified as *ball hits*, probably due to the fact that *ball hits* is a mix of ball hits and court noises. *ball hits* class is often inserted, and *music* class is often deleted.

Finally, the shot audio vectors a_t are created by looking out the audio events that occur within the shot boundary according to the audio segmentation.

4. STRUCTURE ANALYSIS

We integrate a priori information by deriving syntactical basic elements from the tennis video syntax. We define

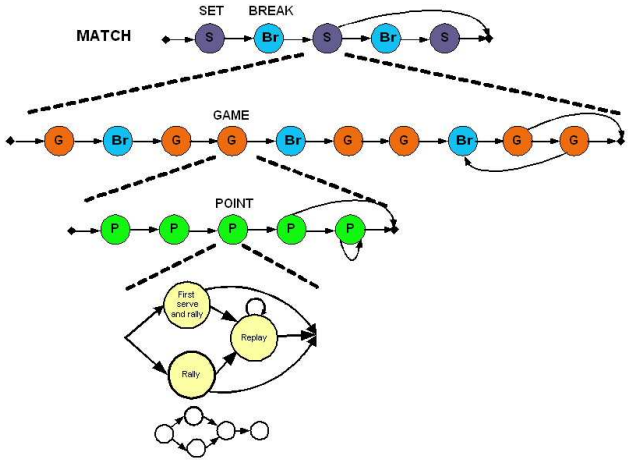

Fig. 3. Content hierarchy of broadcast tennis video

four basic structural units: two of them are related to game phases (missed first serves and rallies), the two others deal with video segments where no play occurs (breaks and replays). Each of these units is modelled by a HMM. The HMMs rely on the temporal relationships between shots.

Each state of the HMMs models either a single shot or a dissolve transition between shots. For a shot t, the observation o_t consists of the similarity measure v_t, the shot duration l_t, the audio description vector a_t, and the relative player position d_t, if it exists. The relative player position to the half-court line is used as a cue to determine if the server has changed or not. The probability of an observation o_t to be in state j at t is then given by:

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

where the probability distributions $p(v_t|j), p(d_t|j)$ and $P[a_t|j]$ are estimated by a learning step. $p(v_t|j)$ and $p(d_t|j)$ are modelled by smoothed histograms, and $P[a_t|j]$ is the product over each sound class k of the discrete probability $P[a_t(k)|j]$. $p(s_t|j)$ is the probability that the server changed given by:

$$p(s_t|j) = \begin{cases} \frac{||d_t|-|d_p||}{Norm} & \text{if state } j \text{ corresponds to a change of server} \\ 1 - \frac{||d_t|-|d_p||}{Norm} & \text{if state } j \text{ doesn't correspond to a change of server} \\ 0,5 & \text{if } d_t \text{ has not been extracted whatever the state} \end{cases}$$

where $Norm$ is a normalization factor that is taken as the half width of the court baseline, and d_p is the previous existing distance extracted. In addition, for each state representing a global view, the player position at the left or right side of the half-court line is considered.

Segmentation and classification of the whole observed sequence into the different structural elements are performed simultaneously using a Viterbi algorithm.

$$\hat{s} = \arg\max_s \;\; \ln p(s) \;\; + \;\; \sum_t \ln b_{s_t}(o_t) \qquad (3)$$

To take into account the long-term structure of a tennis game, the four HMMs are connected to a hierarchical HMM, as represented in Figure 3. This higher level reflects the structure of a tennis game in terms of sets, games, and points. The search space can be described as a huge network where the best state transition path has to be found. The search is performed at different levels. Transition probabilities between states of the hierarchical HMM result entirely from a priori information, while transition probabilities for the sub-HMMs result from a learning process.

Several comments about the hierarchical HMM are in order. The point is the basic scoring unit. It corresponds to a winner rally, that is to say almost all rallies except first missed serves. A break happen at the end of at least ten consecutive points. Boundaries detection between games, and consequently game composition in terms of points, relie essentially on the player position detection, which indicate if the server has changed.

5. EXPERIMENTAL RESULTS

In this section, we describe the experimental results of the audiovisual tennis video segmentation by HMMs. Experimental data are composed of 8 videos, representing about 5 hours of manually labelled tennis video. The videos are distributed among 3 different tournaments, implying different production styles and playing fields. 3 sequences are used to train the HMM while the remaining part is reserved for the tests. One tournament is completely exluded from the training set.

Several experiments are conducted using visual features only, audio features only and the combined audiovisual approach. The segmentation results are compared with the manually annotated ground truth. Classification rates of typical tennis scenes are given in Table 1.

Considering visual features only, the main source of mismatch is when a rally is identified as a missed first serve. In this case the similarity measure is well computed but the player detection or the analysis of the interleaving of shots failed. Replay detection relies essentially on dissolve transition detection. Our dissolve detection algorithm gives a lot of false detections, that leads to a small precision rate (48%). We check that correcting the temporal segmentation improve the replay detection rates up to 100%.

Using only audio features, the precision and recall rates for rallies and first missed serve suggests that audio features are effective to describe rally scenes. Indeed, a rally is essentially characterized by the presence of *ball hits sounds* and *applause* which happen at the end of the exchange,

although a missed first serve is only characterized by the presence of *ball hits*. On the contrary, replays are not characterized by a representative audio content, and almost all replays are missed. The correct detections are more due to the characteristic shot durations of dissolve transitions that are very short. For the same reasons, replay shots can also be confused with commercials that are non-global views of short duration. Break is the only state characterized by the presence of *music*. That means *music* is a relevant event for break detection and particularly for commercials.

Fusing the audio and visual cues enhanced the performance, when audio features are good. Comparing with results using visual features only, there are two significant improvements: the recall and precision rates for rallies, and missed first serve. Introducing audio cues increases the correct detection rate thanks to ball hit sounds and applause.

Recovering the global structure is more interesting to reach a higher-level in the structure analysis (see Table 2). The point boundaries detection is highly correlated with the correct detection of typical tennis scenes. However, the change of server detection is of high relevance for the structure parsing process. Without any information about the end of a game, the Viterbi process falls in local minima when searching the best state transition path, because of the equiprobable transitions between games. The game boundaries detection is then very sensitive to the player extraction, and requires this process to be robust. All misplaced game boundaries are due to errors or ambiguities in player position. Another way to deal with this problem should be to analyze the score displays on superimposed captions.

	Hierarchical segmentation accuracy
Point boundaries	93%
Game boundaries	72%

Table 2. Hierarchical structure identification

6. CONCLUSION

In this paper, we presented a system based on HMMs that uses simultaneously visual and audio cues for tennis video structure parsing. Tennis structure is modelled by an HMM that integrates a priori information about tennis content and editing rules. The tennis video is simultaneously classified and segmented into typical scene, and each segment is matched to one level of abstraction in the hierarchy that describes the structure in term of point, game, and set. The structure parsing allows the building of a "table-of-contents" for efficient browsing and summarization.

The multimodal integration strategy proposed is intermediate between a coarse low-level fusion and a late decision fusion. The audio features describe which classes, among *speech, applause, ball hits, noise* and *music*, are present in the shot. The video features correspond to visual similarity between the shot keyframe and a global view model, and the shot duration.

The results have been validated on a large and various database. They show an encouraging improvement in classification when both audio and visual cues are combined. However, the automatic audio segmentation has to be improved since the errors from the classification spread over the further structuration process. Another solution to avoid this problem is to extend this fusion scheme so that no partial decisions (such as presence or absence of an audio class) is taken before the fusion stage.

7. REFERENCES

[1] G. Sudhir, J. C. M. Lee, and A. K. Jain, "Automatic classification of tennis video for high-level content-based retrieval," in *Proc. Of IEEE Workshop on Content-Based Access of Image and Video Databases*, Bombay, January 1998.

[2] D. D. Saur, Y-P. Tan, S. R. Kulkarni, and P. J. Ramadge, "Automated analysis and annotation of basketball video," in *IS&T/SPIE Storage and Retrieval for Still Image and Video Databases IV*, February 1997, vol. SPIE-3022, pp. 176–187.

[3] J. Assfalg, M. Bertini, A. Del Bimbo, W. Nunziati, and P. Pala, "Soccer highlights detection and recognition using hmms," in *IEEE International Conference on Multimedia and Expo (ICME02)*, August 2002.

[4] P. Chang, M. Han, and Y. Gong, "Extract highlights from baseball game video with Hidden Markov Models," in *Proc. of IEEE International Conference on Image Processing (ICIP02)*, Rochester, NY, USA, September 2002.

[5] L. Xie, S-F. Chang, A. Divakaran, and H. Sun, "Structure analysis of soccer video with Hidden Markov Models," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP02)*, Orlando, FL, USA, May 2002.

[6] D. Zhong and S-F. Chang, "Structure analysis of sports video using domain models," in *IEEE International Conference on Multimedia and Expo (ICME01)*, Tokyo,Japan, August 2001.

[7] H. Zhang, S-Y. Tan, S.W. Smoliar, and G. Yihong, "Automatic parsing and indexing of news video," *Multimedia Systems*, vol. 2, no. 6, pp. 256–266, 1995.

	Visual features		Audio features		AudioVisual features	
Segmentation accuracy	77%		65%		86%	
Classification	precision	recall	precision	recall	precision	recall
First serve	67%	74%	61%	65%	86%	88%
Rallies	92%	63%	91%	55%	94%	86%
Replay	71%	82%	49%	47%	57%	66%
Break	89%	85%	92%	81%	92%	87%

Table 1. Classification and segmentation results with visual features only, manually segmented audio features only, and both manually segmented audio and visual features, in a reduced subset of videos belonging to the same tournament

[8] R. Lienhart, "Reliable transition detection in videos: A survey and practitioner's guide," *International Journal of Image and Graphics*, vol. 1, no. 3, pp. 469–486, 2001.

[9] C.G.M. Snoek and M. Worring, "Multimodal video indexing: A review of the state-of-the-art," *Multimedia Tools and Applications*, 2003, to appear.

[10] Y. Wang, Z. Liu, and J.-C. Huang, "Multimedia content analysis using both audio and visual cues," *IEEE Signal Processing Magazine*, pp. 12–36, November 2000.

[11] H. Jiang, t. Lin, and H. Zhang, "Video segmentation with the support of audio segmentation and classification," in *IEEE International Conference on Multimedia and Expo (I)(ICME00)*, August 2000, vol. 3, pp. 1551–1554.

[12] Z. Liu and Q. Huang, "Detecting news reporting using audio/visual information," in *Proc. of IEEE International Conference on Image Processing (ICIP99)*, October 1999, vol. 1, pp. 324–328.

[13] J. Huang, Z. Liu, and Y. Wang, "Integration of multimodal features for video scene classification based on hmm," in *Proc. Of IEEE Workshop on Multimedia Signal Processing*, Copenhagen, Denmark, September 1999, pp. 53–58.

[14] 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*, London, UK, July 2002, vol. 2383, pp. 278–288, Springer, Lecture Notes in Computer Science.

[15] M. Xu, L-Y. Duan, C-S. Xu, and Q. Tian, "A fusion scheme of visual and auditory modalities for event detection in sports video," in *IEEE Int. Conference on Acoustics, Speech, and Signal Processing (ICASSP03)*, Hong Kong, April 2003.

[16] H.J Zhang, A. Kankanhalli, and S.W. Smoliar, "Automatic partitioning of full-motion video," *Multimedia Systems*, vol. 1, no. 1, pp. 10–28, 1993.

[17] M. Betser, G. Gravier, R. Gribonval, and F. Bimbot, "Extraction of information from video sound tracks - can we dectect simultaneous events ?," in *Third International Workshop on Content-Based Multimedia Indexing (CBMI03)*, September 2003, pp. 71–77.