AUDIOVISUAL INTEGRATION FOR TENNIS BROADCAST STRUCTURING

E. Kijak [1,2], *G. Gravier* [2], *L. Oisel* [1], *P. Gros* [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}@irisa.fr

ABSTRACT

This paper focuses on the integration of multimodal features for sport video structure analysis. The method relies on a statistical model which takes into account both the shot content and the interleaving of shots. This stochastic modelling is performed in the global framework of Hidden Markov Models (HMMs) that can be efficiently applied to merge audio and visual cues. Our approach is validated in the particular domain of tennis videos. The model integrates prior information about tennis content and editing rules. 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. As a result, typical tennis scenes are simultaneously segmented and identified.

1. INTRODUCTION

Video content-based analysis is an active research domain that aims at automatically extracting high-level semantic events from video. The extracted semantic information can be used to produce indexes or table of contents that enable efficient search and browsing of the video content. Indexing tasks usually attempt to identify a predetermined set of events within a video. 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 [1]. The critical step is to automatically group shots into "scenes". Scenes are referred in the literature as story units. They are defined as a coherent group of shots that is meaningful for the end-user. The main problem relies on the definition of the "coherence" of the shots within a scene. We are able to group shots with similar low-level features, but defining a "scene" depends more on a subjective semantic correlation. A way to extract high-level information is then to focus on particular types of videos, like news or sports, and introduce a priori information about content and structure. In sports analysis, a common approach consists in combining low-level features with heuristic rules to infer specific highlights [2, 3, 4]. Statistical models like Hidden Markov Models (HMMs) have also been used for this task [5, 6]. However, most of these approaches use one single media.

In this paper, we address the problem of recovering sport video structure, through the example of tennis which presents a strong structure. 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 [7], and to represent the hierarchical structure of a tennis match. As a result each shot/group of shots is classified within one of the following four categories: *missed first serve*, *rally*, *replay*, and *break*.

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 [8, 9]. Multimodal approaches have been investigated for different areas of content-based analysis, such as scene boundary detection [10], structure analysis of news [11], and genre classification [7]. 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 audiovisual feature vector before the classification. The multimodal features have to be synchronized [11]. 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.

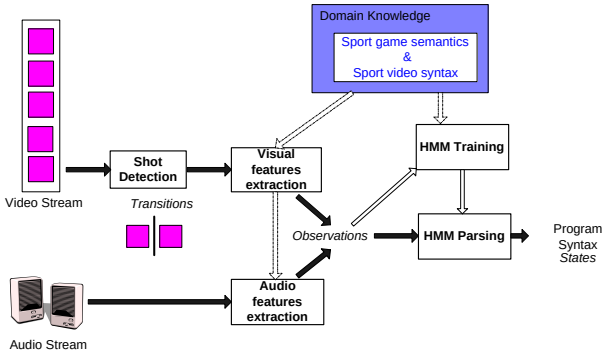

Fig. 1. Structure analysis system overview

However, some dependencies among features from different modalities are not taken into account in this late integration scheme. For example, applying to the detection of rally shots in a tennis video, [12] define independently a visual likelihood of a frame to be a court view, and an audio likelihood of a segment (synchronized on frame sampling rate) to represent a racket hit. The final decision is taken by multiplying these two likelihoods.

But usually theses approaches rely on a successive use of visual and audio classification [13, 14]. For example in [14], 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 1). 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.

This paper is organized as follows. Section 2 provides elements on tennis video syntax. Section 3 gives an overview of the system and describes the visual and audio 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. Considtbh

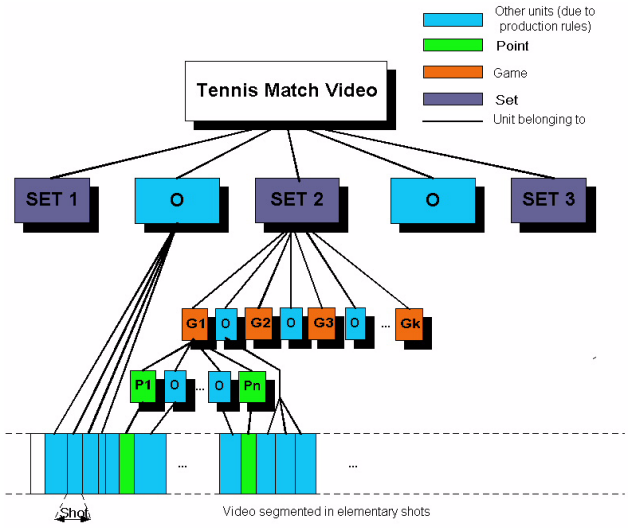

Fig. 2. Structure of Tennis Game

ering a given instant, the point of view giving the most relevant information is selected by the broadcaster. 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. Because of the presence of typical scenes, production rules, and the finite number of views, the tennis video has a predictable temporal syntax.

We identify here four typical patterns in tennis videos that are: *missed first 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. As opposed to 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, 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. One specificity of our system is to identify global views from non global view shots, and then to analyse the temporal interleaving of these shots.

3. SYSTEM OVERVIEW

In this section, we give an overview of the system (Figure 1) and 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 t, we compute the following visual and audio features:

- the duration d_t of the shot
- a visual similarity v_t representing a visual distance between a keyframe extracted from the shot, and a global view model
- a binary audio vector a_t describing which audio classes, among *speech, applause, ball hits, noise* and *music*, are present in the shot

The extraction of these higher-level features is described in sections 3.1 and 3.2. The sequence of shot features results in a sequence of observations, which is then parsed by a HMM process. Final classification of each shot in typical tennis scenes is given by the resulting sequence of states.

Similar approaches are used in [15] and [16]. In [15] dialog scenes are identified using a HMM shot-based classification. Each shot is characterized by three labels which indicate respectively if the shot contains speech or silence or music, if a face is detected or not, and if the scene location has significantly changed or not. In our scheme, preliminary decisions on audio and visual features are less deterministic, since no classifications into "global views" or "short/long" duration are performed before the HMM classification. Uncertainties are deliberately let at feature level to allow the system to take a decision at a higher level and based on audio visual cues simultaneously.

In [16], a multimodal classification method of baseball shots based on the maximum entropy method is proposed. Eight typical scenes are identified. To catch the temporal transition within a shot, each shot is divided into three equal segments. However, inter-shot temporal transitions are not taken into account. In our approach, we perform a simpler shot categorization into global or non-global views, but we aim to analyze temporal interleaving of shots in order to identify higher-level segments. 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: *missed first serve, rally, replay,* and *break*.

3.1. 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 tennis court 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. Commercials are also taken into account. Dominant colors of commercial keyframes are unpredictable. However, due to the cost of air time, the visual characteristic of a piece of commercial is that it usually has more actions, corresponding to more shots and faster motion within each shot.

The first class can thus be characterized by a small camera motion while the lasts 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.

Each keyframe K_i is then described by a dominant color vector F_i and its spatial coherency C_i, and the activity A_i of the corresponding shot. The visual similarity measure 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 first reduce the set of candidate keyframes by discarding those whose highest percentage of dominant color is less than 50%. In the resulting subset of keyframes, global views represent more than 50% of the data. We then 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}. The visual similarity measure $\mathrm{v}(\mathrm{K}_t, \mathrm{K}_{ref})$, denoted by v_t in the following, is then computed between each keyframe K_t and K_{ref}.

3.2. Audio Features

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 soundtrack segmentation is carried out using a Viterbi based system with Gaussian mixtures for each sound class and combination of sound classes. The audio signal is demultiplexed from the video and converted into features representative of the audio content. Classically, the features considered in this work are cepstral coefficients plus log-energy, extracted on 20 ms consecutive windows with a 50% overlap. The 16 cepstral coefficients and the energy are completed by the first and second order derivatives and the absolute log-energy is discarded from the feature vector.

A training step is performed on hand labelled data to estimate the parameters of a 64 component Gaussian mixture model (GMM) for each of the 5 audio classes considered. One problem with soundtrack segmentation is that more than one audio event maybe present simultaneously. For example, it is quite usual to have speech superimposed with ball hits or with applauses. Such events cannot be detected using a Viterbi based segmentation unless a model is given for each possible combination of superimposed class. Since parameter estimation for such models would require a (unavailable) very large amount of data, we assumed that a maximum of two events could be simultaneously present and derived models for every pair (c_1, c_2) of audio classes from the single class models, according to

$$p(y(t)|c_1, c_2) = \frac{1}{2} p(a(t)|c_1) p(a_t|c_2) \ , \qquad (2)$$

where y_t denotes the feature vector corresponding to frame t. The underlying assumption for this model is that either one or the other class is present in a given feature vector $y(t)$, with equal probability.

Finally, segmentation and classification of the soundtrack is done using a Viterbi decoder along with an ergodic HMM where each state represent either a single audio class or a combination of two audio classes. In the first case, the state conditional probabilities are defined by the GMM corresponding to the audio class associated with the considered state while, in the second case, the conditional probability is given by eq. (2). 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.

To assess the quality of the audio segmentation alone, models where trained on the first hour of a tennis video and tested on the second hour. The former corresponds to the introduction plus a part of the first set while the latter includes the end of the first set and the entire second set plus some commercials. The frame correct classification rate obtained is 76.9% while the total frame classification error rate is 46.3% due to insertions. The corresponding confusion matrix is given in Table 1 and shows that *tennis* and *music* are well classified while *noise* is often classified as *tennis*, probably due to the fact that *tennis* is a mix of ball hits and court noises. The last row also shows that the *tennis* class is often inserted.

	noise	tennis	appl.	music	speech	del
noise	18.5	66.6	4.8	0.1	9.9	0.0
tennis	0.2	95.8	0.4	0.1	0.6	2.8
appl.	0.0	23.2	69.1	4.6	0.8	2.3
music	0.2	1.9	1.8	96.0	0.1	0.0
speech	0.3	3.2	1.4	6.8	86.1	2.2
ins	0.0	82.1	14.1	52.3	7.3	–

Tab. 1. Confusion matrix for the audio segmentation system.

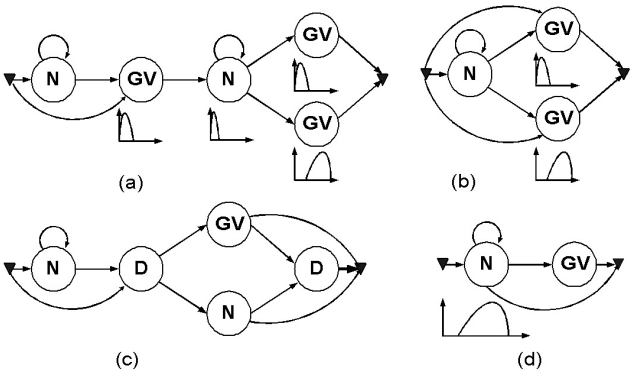

Fig. 3. Hidden Markov Models of the four basic structural units: (a) missed first serve and rally (b) rally (c) replay (d) break. GV stands for Global View, N for Non-global view, and D for Dissolve transition

4. 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 dealing with video segments where no play occur (breaks and replays). Each of these units is modelled by a HMM. These HMMs rely on the temporal relationships between shots (Figure 3), according to the common editing rules explained in section 2.

In a tennis match, only global views are generally of interest and the structural units aim to identify the global views as being a first missed serve or the point that ended a set. Nevertheless, *non global views* convey important information about structure of the video according to production style, and help in the parsing process.

Each HMM state models either a single shot or a dissolve transition between shots. Three observation streams are associated with each state: the visual similarity between the shot keyframe and K_{ref}, the shot duration, and the audio vector which characterizes the presence or absence of the predetermined audio events. More formally, for a shot t, the observation o_t consists of the similarity v_t, the shot duration d_t and the audio description vector a_t.

The probability of an observation o_t conditionally to

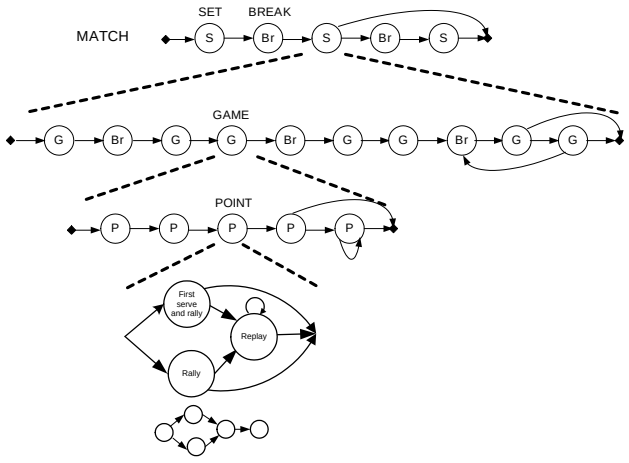

Fig. 4. Content hierarchy of broadcast tennis video

state j is then given by:

$$b_j(o_t) = p(v_t|j) \;\; p(d_t|j) \;\; P[a_t|j] \tag{3}$$

where $p(v_t|j)$ is a Gaussian distribution, $p(d_t|j)$ can be a Gaussian distribution, a mixture of Gaussians or an 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 into the different structural elements are performed simultaneously using a Viterbi algorithm. The most likely sequence state is given by:

$$\hat{s} = \arg\max_{s} \;\; \ln p(s) \;\; + \;\; \sum_{t} \ln b_{s_t}(o_t) \tag{4}$$

To take into account the long-term structure of a tennis game, the four HMMs are connected through a higher level HMM, as illustrated in Figure 4. This higher level represents the tennis syntax and the hierarchical structure of a tennis match, described in terms of points, games and sets. 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. Considering these rules in the model ensures a long-time correlation between shots, and avoids, for example, the apparition of an interleaving of points and breaks.

Transition probabilities between states of the higher level HMM result entirely from prior information about tennis rules, while transition probabilities for the sub-HMMs result from a learning process.

5. EXPERIMENTAL RESULTS

In this section, we describe the experimental results of the audiovisual tennis video segmentation by HMMs. Experimental data are composed of about 2 hours of manually labelled tennis video. About half of the data are used to train the HMM while the remaining half part is reserved for the tests.

	Visual features		Audio features	
Segmentation accuracy	65%		77%	
Classification	precision	recall	precision	recall
First serve	67%	63%	93%	80%
Rallies	97%	54%	95%	84%
Replay	100%	80%	71%	41%
Break	51%	93%	77%	87%

Tab. 2. Classification and segmentation results with visual features only and with manually segmented audio features only for β = 10%

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.

5.1. Using Visual Features Only

For each observation, only the visual similarity that includes dissolve transition detection, and the shot duration are taken into account. That means the decoding process relies on the type of shot (or more specifically its similarity to a global view) and its duration. Recall and precision rates are given in Table 2. Precision is defined as the ratio of the number of shots correctly classified to the total number of shots retrieved. Recall is defined as the ratio of the number of shots correctly classified to the total number of relevant shots.

Rallies detection presents a low recall rate (54%) that means many shots are missed. The similarity measure is not totally robust to strong variations in illumination conditions that could happened during an outdoor match. As a result, the concerned global views could not be identified as rallies. Another 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 analysis of the interleaving of shots based on the shot duration failed. The high precision rate (97%) shows that the similarity measure works well to discriminate non global views since few non global views are classified as rally. This is confirmed by the break detection rates which presents on the contrary a low precision (51%) and a high recall (93%).

Replay detection relies essentially on dissolve transition detection. This explains the high precision rate which corresponds to correctly detected dissolve transitions.

5.2. Using Audio Features Only

For each observation, only the audio vector and the shot duration are taken into account. Dissolve transitions are then

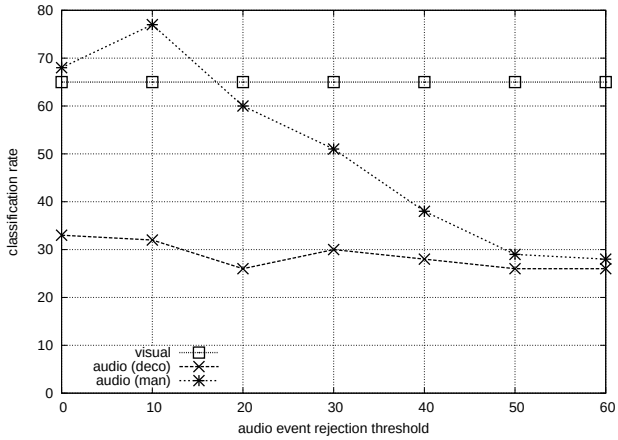

Fig. 5. Classification rate using manually (man) and automatically (deco) segmented audio features, when β varies from 0 to 60 %

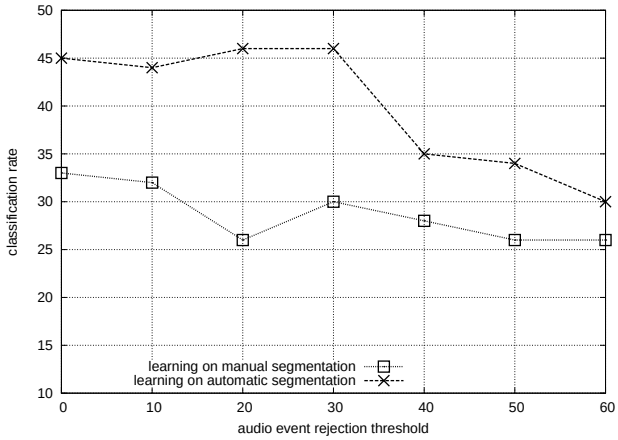

Fig. 6. Classification rate using automatically segmented audio features, when the learning is performed on manually or automatically segmented features

not considered. The audio vectors represent the presence or absence of predetermined audio events in a shot without any measure of the event importance. This means that audio events, such as *applause*, may be detected in a shot, although it appears only a few seconds and is not representative of the shot audio content. We introduce a threshold β which represents a percentage of the shot duration. If the duration of an audio event within a shot is less than β, the audio event is discarded.

We use two types of audio features. The first one results from a manual segmentation and is used to validate the approach. In this case, audio features are generated from the ground truth audio segmentation. The last one results from the automatic segmentation process described in section 3.2.

In a first experiment, audio probabilities are estimated from the manually segmented features. Figure 5 represents the classification rate when β varies from 0 to 60%, and where the audio segmentation is either manual (man) or decoded (deco). The classification rate from visual decoding is given as reference.

The results show that the errors from the automatic audio segmentation spread over the structuration process, and drag down the performance. However using manually segmented audio features (that symbolize a "perfect" audio segmentation), the classification rate is significantly higher for β = 10% (77%) than using visual features only (65%). When β increases, the performance goes down as audio events become less representative of the audio shot content.

The increase in recall rates for rallies and first missed serve (Table 2) 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*. Therefore, there are less confusion between missed first serves and rallies, and less missed detections.

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, and then classified as break, especially if the replay happens just before a break.

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. In fact, the break state works in this case like a commercial detector, and avoids false detections. However it increases the probability to miss a short break without commercials.

Concerning the poor performances given by automatically decoded audio features, there is a mismatch between the manually segmented audio features on which the HMM parameters were estimated, and the automatically generated audio features. In the first one, presence of *ball hits* or *music* are respectively synonymous of rally and break. In the last one, *ball hits* are present in other states than rallies, *music* is present in other states than break, and *noise* are often not detected.

To tackle this problem, another experiment was conducted where audio probabilities were estimated from the automatically segmented audio features, in order to take into account the audio segmentation errors in the training process . Figure 6 shows the improvement of the decoding performance on automatically segmented vectors. However, the results are still less accurate than using visual features only.

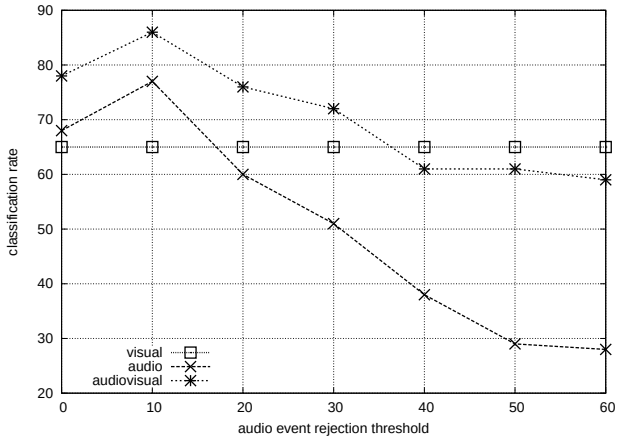

Fig. 7. Classification rate using both manually segmented audio features and visual similarity, when β varies from 0 to 60 %

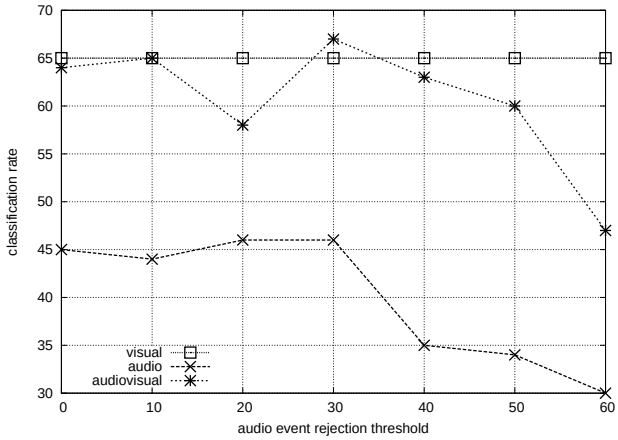

Fig. 8. Classification rate using both automatically segmented audio features and visual similarity, when β varies from 0 to 60 %

5.3. Using Both Audio-Visual Features

Finally, both the visual similarity and the audio vector are taken into account.

Figure 7 represents the classification rate when audio probabilities are estimated from the manually segmented features, and β varies from 0 to 60%. The observation sequence is decoded using manually segmented audio features. The classification rate from visual and audio (manually segmented) decoding are given as references.

Figure 8 represents the classification rate when audio probabilities are estimated from the automatic audio segmentation. In this case, audio performance is enhanced by fusing audio and visual features, but it is not sufficient to improve the performance given by visual features only. Finally, the poor results given by automatic audio segmentation tend to degrade the decoding performance of visual features.

Nevertheless, Figure 7 shows that fusing the audio and visual cues significantly enhanced the performance, when audio features are good. As with audio features only, the best result (86%) is given for β = 10%. For this case, detailed classification rates are given in Table 3.

In particular, for rallies identification, there are no false detections anymore. Comparing with results using visual features only, there are two significant improvements: the recall rate for rallies, and the precision rate for breaks. Introducing audio cues increases the correct detection rate thanks to tennis sound in the first case, and avoids false detection thanks to music detection in the second case.

5.4. In the search of the relevant audio information

From the previous analysis about audio features, *applause*, *ball hits* and *music* seems to be the most relevant events for the structure analysis process. In the following experiment, we use manually segmented audio features and successively discard *speech* and *noise* classes from the audio features a_t (see Figure 9).

	AudioVisual features	
Segmentation accuracy	86%	
Classification	precision	recall
First serve	86%	88%
Rallies	100%	90%
Replay	83%	90%
Break	95%	84%

Tab. 3. Classification and segmentation results using both manually segmented audio features only and visual similarity, for β = 10%

Discarding *speech* provides a slight improvement in the classification rate although discarding *noise* decreases the performance. As a commentator speaks during almost all the broadcast, the *speech* event is present in all states with quite the same probability, and is therefore not informative. On the contrary, *noise* is also present in all states but not with the same importance. For example, the probability to have *noise* in a rally shot is more important than in a break shot.

6. CONCLUSION

In this paper, we presented a system based on HMMs that uses simultaneously visual and audio cues for tennis video structure parsing. The tennis video is simultaneously segmented and classified into typical scenes of higher level than a tennis court view classification.

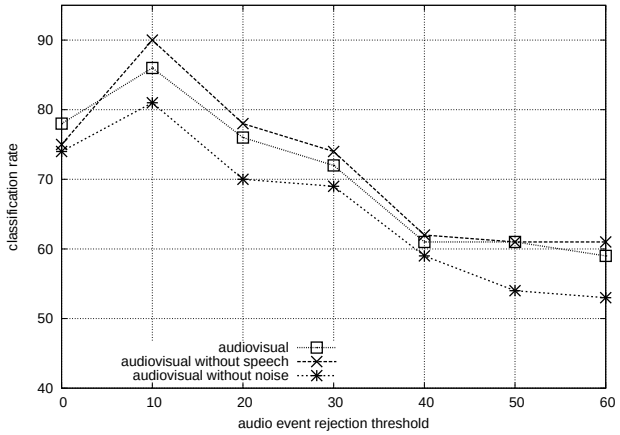

Fig. 9. Classification rate using both manually segmented audio features and visual similarity, when β varies from 0 to 60 %

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. There are no further decision made on these features before the HMM classification, like a global view classification, or major audio event detection. Such late decisions are taken a at a higher level, considering the context, and based on audio visual cues simultaneously.

The results show an encouraging improvement in classification when both audio and visual cues are combined. However, these preliminary results should be validated on a larger database. The automatic audio segmentation has also 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. We are currently working on such approaches where a combined model is defined on low level features.

7. REFERENCES

[1] R. Lienhart. Reliable transition detection in videos: A survey and practitioner's guide. *Intl. Journal of Image and Graphics*, 1(3):469–486, 2001.

[2] P. Xu, L. Xie, S-F. Chang, A. Divakaram, A. Vetro, and H. Sun. Algorithms and system for segmentation and structure analysis in soccer video. In *Proc. of ICME*, 2001.

[3] G. Sudhir, J. C. M. Lee, and A. K. Jain. Automatic Classification of Tennis Video for High-level Content-Based Retrieval. *IEEE Work. on Content-Based Access of Image and Video Databases*, 1998.

[4] W. Zhou, A. Vellaikal, and C.-C. J. Kuo. Rule-based video classification system for basketball video indexing. In *Proc. ACM Intl. Multimedia Conf.*, pp. 213–216, 2000.

[5] P. Chang, M. Han, and Y. Gong. Extract highlights from baseball game video with hidden markov models. In *Proc. of ICIP*, 2002.

[6] L. Xie, S-F. Chang, A. Divakaran, and H. Sun. Structure analysis of soccer video with hidden markov models. In *Proc. of ICASSP*, 2002.

[7] J. Huang, Z. Liu, and Y. Wang, editors. Integration of Multimodal Features for Video Scene Classification based on HMM. In *IEEE Work. on Multimedia Signal Proc.*, 1999.

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

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

[10] H. Jiang, t. Lin, and H. Zhang. Video segmentation with the support of audio segmentation and classification. In *Proc. of ICME*, v 3, pp 1551–1554, 2000.

[11] Z. Liu and Q. Huang. Detecting news reporting using audio/visual information. In *Proc. of ICIP*, v 1, pp 324–328, 1999.

[12] R. Dayhot, A. Kokaram, N. Rea, and H. Denman. Joint audio visual retrieval for tennis broadcasts. In *Proc. of ICASSP*, 2003.

[13] 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*, v 2383, pp 278–288, 2002.

[14] 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 *Proc. of ICASSP*, 2003.

[15] A. A. Alatan, A. N. Akansu, and W. Wolf. Multi-modal dialog scene detection using hidden markov models for content-based multimedia indexing. *Multimedia Tools and Application*, 14(2):137–151, 2001.

[16] W. Hua, M. Han, and Y. Gong. Baseball scene classification using multimedia features. In *Proc. of ICME*, 2002.