Multimed Tools Appl (2006) 30: 289–311
DOI 10.1007/s11042-006-0031-5

Audiovisual integration for tennis broadcast structuring

Ewa Kijak · Guillaume Gravier · Lionel Oisel · Patrick Gros

Published online: 5 August 2006

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. Two sets of audio features are used in this study: the first one is extracted from a manual segmentation of the soundtrack and is more reliable. The second one is provided by an automatic segmentation and classification process. As a result of the overall HMM process, typical tennis scenes are simultaneously segmented and identified. The experiments illustrate the improvement of HMM-based fusion over indexing using only the best single media, when both media are of similar quality.

Keywords Video structure analysis · Macro-segmentation · Cross-modality · Hidden Markov models

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

E. Kijak (✉)
Université de Rennes I, Campus Universitaire de Beaulieu, 35042 Rennes Cedex, France
e-mail: ewa.kijak@lip6.fr

G. Gravier
CNRS, Campus Universitaire de Beaulieu, 35042 Rennes Cedex, France

L. Oisel
Thomson multimedia R&D, 1 av. Belle Fontaine, 35510 Cesson-Sévigné, France

P. Gros
IRISA, Campus Universitaire de Beaulieu, 35042 Rennes Cedex, France

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 [11]. 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 [15, 18, 22]. Statistical models like Hidden Markov Models (HMMs) have also been used for this task [3, 17]. 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 [14, 16]. Multimodal approaches have been investigated for different areas of content-based analysis, such as scene boundary detection [8], structure analysis of news [12], 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 analyze 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 [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. For example, applying to the detection of rally shots in a tennis video, [4] 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 [10, 19]. For example in [19], 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 3). 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, ball hits, 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. Considering 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.

The different types of views present in a tennis video can be divided into four principal classes: global, medium, close-up, and audience (see figure 1). 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 analyze the temporal interleaving of these shots.

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

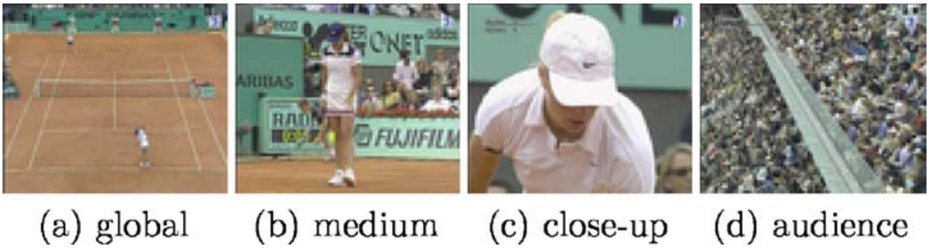


Fig. 1 Four main types of view in tennis video

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).

3 System overview

In this section, we give an overview of the system (figure 3) and describe the extraction of audio and visual cues. First, the video stream is automatically segmented into shots by detecting cuts and dissolve transitions. Straight cut detection is performed by detecting rapid changes of the standard bin-wise difference between luminance histograms of DC pictures. Gradual shot transitions detection is performed by the twin-comparison method [20], using a dual threshold that accumulates significant differences to detect gradual transitions. The content of a detected shot is represented by a single keyframe (as only one keyframe is sufficient to illustrate the whole content of a view).

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 the keyframe extracted from the shot, and a global view model

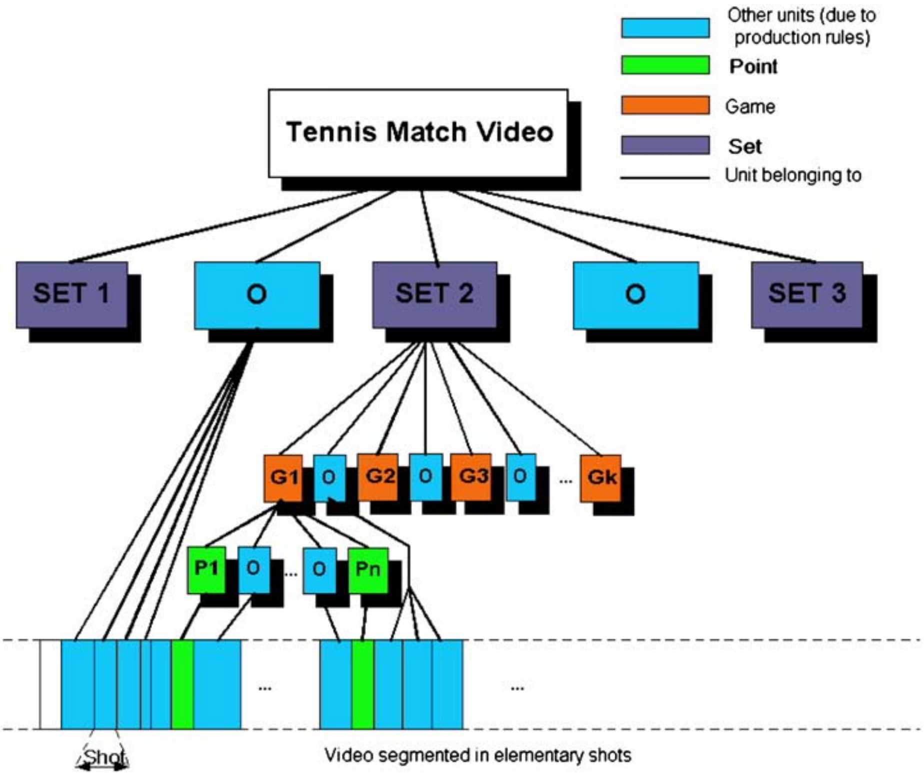


Fig. 2 Structure of tennis game

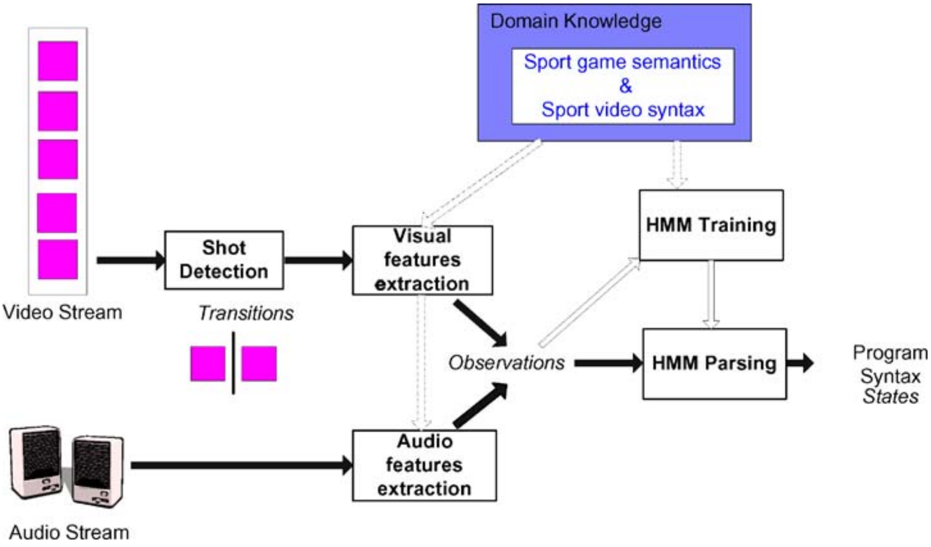


Fig. 3 Structure analysis system overview

- a binary audio vector a_t describing which audio classes, among *speech*, *applause*, *ball hits*, *silence* 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 [1] and [6]. In [1] 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 [6], 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 labelled with one of these four types: *missed first serve*, *rally*, *replay*, and *break*.

3.1 Visual features

We use visual features to identify the global views within all the extracted keyframes. Identifying the different types of view is a necessary step in any sports content analysis. Many works deal with shot classification in the context of sport videos [5, 9, 15, 17, 21]. In most of the methods that use color information, the game field color has to be first evaluated, because it can largely vary from one video to another. Our approach tries to avoid the use of predefined field color to be able to automatically take into account a large type of videos.

The process can be divided into two steps. First, a keyframe K_{ref} representative of a global view is automatically 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 first motivate the choice of the visual features we extract, and then we define the visual similarity measure we use, before describing the K_{ref} selection 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, commercials are usually characterized by more actions, corresponding to more shots and faster motion within each shot. Global views can thus be characterized by a dominant color content and a small camera motion, while the other views imply unpredictable color content and important camera translations. Based on these observations, we choose three features to identify global views:

- a vector F of N dominant colors,
- its spatial coherency C,
- the activity A that reflects the average camera motion during a shot.

Rather than color histogram, we use a global descriptor of dominant colors that is more compact. In addition, dominant colors vectors capture the most significant color information of a frame and are less noise sensitive.

3.1.1 Dominant color vector

Let F be a vector of N dominant colors and p_i the percentage of each color $c_{i,1\leq i\leq N}$ with respect to the whole associated frame. The colors of the original images are quantized into N values using a k-means clustering algorithm. Neighboring dominant colors are merged when their distance are less than a predefined threshold T_d. The goal is to ensure that the N dominant colors are perceptually different. According to MPEG-7, the similarity between two dominant colors features F_1 and F_2 can be then measured by the following simplified quadratic distance function $d(F_1, F_2)$:

$$d^2(F_1,F_2) = \sum_{i=1}^{N_1} p_{1i}^2 + \sum_{j=1}^{N_2} p_{2j}^2 - \sum_{i=1}^{N_1}\sum_{j=1}^{N_2} 2a_{1i,2j}p_{1i}p_{2j} \tag{1}$$

where $a_{k,l}$ is the similarity coefficient between two colors c_k and c_l,

$$a_{k,l} = \begin{cases} 1 - \dfrac{d_{k,l}}{d_{\max}} & \text{if } d_{k,l} \leq T_d \\ 0 & \text{if } d_{k,l} > T_d \end{cases} \tag{2}$$

T_d is the maximum distance for two similar colors, $d_{\max} = \alpha\, T_d$, and $d_{k,l}$ is the Euclidean distance between two colors c_k and c_l.

To take into account the spatial configuration of similar color pixels, a confidence measure C is associated to each dominant color feature. A pixel of color c_k is considered to be coherent if all the pixels in its neighborhood have the same color. As a result, the confidence measure C for the dominant color feature F is defined as:

$$C = \sum_{k=1}^{N} \frac{\text{number of coherent pixels } c_k}{\text{total number of pixels}} \tag{3}$$

In our implementation, we use 4 dominant colors $(N = 4)$ to characterize the most significant color information in the game field (see figure 4).

3.1.2 Activity

The activity A is defined as the average camera motion during a shot and is based on MPEG motion vectors. These vectors are used to estimate the camera dominant motion within the sequence by deriving a parametric model of the 2D dominant apparent motion represented by each motion field. A simple three-parameter model covers the most common categories of global dominant motion (translation, zoom and no motion):

$$\begin{cases} u = & kx + t_x \\ v = & ky + t_y \ . \end{cases} \tag{4}$$

In these equations, (u, v) represent the horizontal and vertical components of the motion vector of the pixel located at (x, y) in the image; k, t_x and t_y are the motion model parameters. As the MPEG motion vectors don't necessarily represent the "real/physical" motion, an outlier rejection step is used to discard vectors that don't match the estimated motion. The rejection process is performed by a robust linear regression.

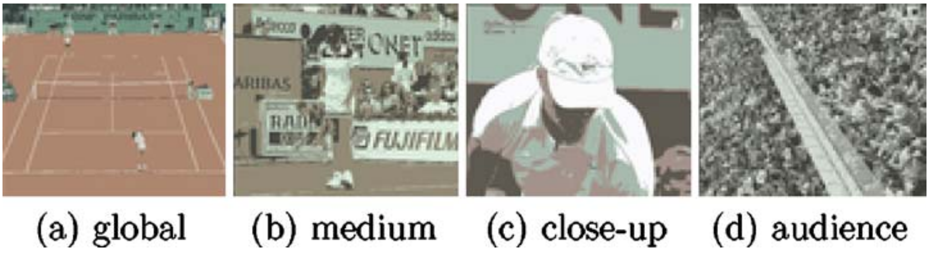
Let Ω be the set of inliers vectors. The activity A is defined by:

$$A^2 = \sum_{(u,v)\in\Omega} \frac{u^2 + v^2}{|\Omega|} \tag{5}$$

3.1.3 Visual similarity

Each keyframe K_t is then described by a dominant color vector F_t, its spatial coherency C_t, and the activity A_t 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:

$$v(K_1, K_2) = w_1 \ |C_1 - C_2| + w_1 \ d(F_1, F_2) + w_3 \ |A_1 - A_2| \tag{6}$$



(a) global (b) medium (c) close-up (d) audience

Fig. 4 Dominant colors extraction

where w_1, w_2, and w_3 are weighting coefficients, set as follows: $w_1 = 0,2,\ w_2 = 0,5$, and $w_3 = 0,3$.

The visual similarity is used to characterize the shot content. This measure is computed between each keyframe K_t and a keyframe K_{ref} representative of a global view, and denoted by v_t in the following:

$$v_t = v(K_t, K_{ref}) \tag{7}$$

As a result, a keyframe K_t is more similar to a global view as v_t is smaller. On the contrary, a close-up keyframe K_t will present a high visual distance v_t.

The given weighting coefficients were tuned to separate as good as possible the visual similarity distributions of the global view class and the non-global view class. However, these coefficients have not a significant influence on the following.

On the next section, the K_{ref} selection process is described.

3.1.4 Selection of a global view keyframe K_{ref}

Our goal is here to identify a global view from all extracted keyframes without making any assumption about the playing area color. Our method can thus get rid of the different types of tennis court (carpet, clay, hard or grass). Analyzing several hours of tennis video reveals that in a video, global views keyframes represent only 20 to 30% of all extracted keyframes including commercials.

To find a keyframe K_{ref} representative of a global view, we consider dominant colors ratios. As it was previously noted, color contents in medium and audience views are more scattered than in global views. Considering that a global view is mainly composed of the playing area, we assume that the percentage of the main dominant color is greater than 50%. We reduce the set of candidate keyframes by discarding keyframes whose highest percentage of dominant color is less than 50%. In the resulting subset of images, global views represent more than 50% of the data (most of medium and audience views have been discarded).

The main problem thus remains the distinction of global views. In other words, the problem is now reduced to an identification of inlier datapoints i.e., global views, in the presence of data outliers. We apply the least median square method (LMS), to minimize the median distance of all the remaining keyframes to a randomly selected keyframe. The number p of samples is chosen in a way that the probability P of finding a representative global view keyframe is greater than 99%. The expression for p is given by [13]:

$$p = \frac{\log(1 - P)}{\log(1 - (1 - \tau)^q)} \tag{8}$$

where τ is the fraction of outlier data, and q the number of features in each sample. Here $q = 1$, as each sample K_t has one associated feature v_t.

The set of candidate keyframes is reduced again by keeping keyframes whose distance is lower than the median distance previously found. The LMS is re-iterated on this new subset to select a better reference keyframe K_{ref}. This last step ensure the selected keyframe K_{ref} to be independent from the random process. In other words, different random pulling lead to the same K_{ref}. Figure 14 shows, for different videos, the dominant color repartition of keyframes in YCbCr space, according to their type of view, and the associated selected K_{ref}.

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*, *silence* and *music*, are present in the shot is extracted from an automatic segmentation of the audio stream (figure 5).

To achieve this goal, 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 16 cepstral coefficients extracted on 20 ms consecutive windows with a 50% overlap. 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. In this step, delta coefficients are considered in addition to cepstral coefficients. The segmentation and classification steps are recalled below. More details can be found in [2].

3.2.1 Soundtrack segmentation

Segmenting the soundtrack into spectrally homogeneous segments is carried out with the Bayesian information criterion (BIC) using a cepstral representation of the input signal. The BIC is defined as the log-likelihood of a segment y given a model, penalized by the model complexity and the segment length. Formally, for a segment of length T with an associated model Λ, the Bayesian information criterion is defined as

$$\mathcal{I}(\Lambda) = \ln f(y;\, \Lambda) - \gamma \frac{\#(\Lambda)}{2} \ln T,$$

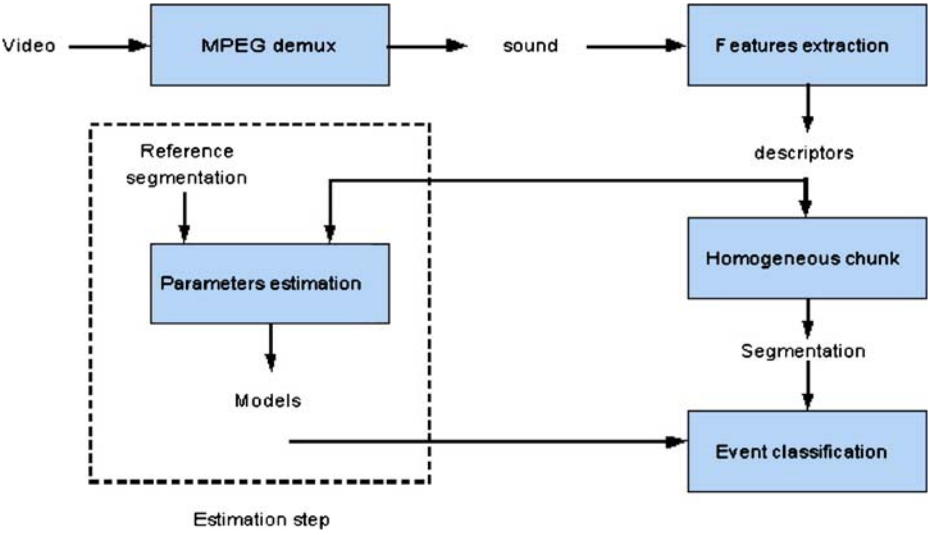


Fig. 5 Architecture of the audio segmentation system

where $f(y;\,\Lambda)$ is the likelihood of a segment y given a model Λ, $\#(\Lambda)$ is the number of free parameters in the model, and γ a tunable parameter theoretically equal to one (but not in practice).

The principle of the segmentation algorithm is to move two adjacent windows on the soundtrack signal. For each position of the two windows, one can compare whether it is best, in terms of BIC, to model the two windows separately with two different Gaussian distributions, say Λ_1 and Λ_2, or with a single one, say Λ_0. If $\mathcal{I}(\Lambda_0)-\mathcal{I}(\Lambda_1)-\mathcal{I}(\Lambda_2)$ is negative, then the two-segment model is more appropriate and a segment boundary is detected at the frontier of the two windows. In this study, 3 s windows are used.

3.2.2 Audio classification

A training step is performed on manually labelled data to estimate the parameters of a 64 component Gaussian mixture model (GMM) for each of the five audio events considered.

In the audio classification step, the presence or absence of each label has to be detected in every previously extracted segments. This detection problem can be classically solved using 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 segment y.

In practice, $f(y;\,H_0)=f(y;\,M)$ where M is a Gaussian mixture model whose parameters were estimated from training data containing the event of interest (whether alone or superimposed with other events). Similarly, $f(y;\,H_1)$ is approximated using a “non-event” model $\overline{M}$ whose parameters were estimated on training data where the event considered is not represented. The decision threshold δ was determined experimentally on training data.

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 corresponding confusion matrix is given in Table 1 and shows that *ball hits*, *speech* and *applause* are well classified while *silence* is often misclassified as *ball hits*, probably due to the fact that *ball hits* is a mix of ball hits and court silences. The last row also shows that *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. The resulting vector has then five components, set to 1 if the corresponding audio event has been detected in the shot, and to 0 otherwise.

Table 1 Confusion matrix for the audio segmentation system

	Music	Speech	Ball hits	Appl.	Silence	Del.
Music	16.49	14.35	17.13	13.33	1.59	37.11
Speech	0.11	87.15	2.21	0.67	1.63	8.23
Ball hits	0	5.67	77.16	2.54	4.90	9.73
Appl.	0.02	0.30	2.33	88.47	0.58	8.29
Silence	0.93	12.08	30.60	15.52	40.87	0.00
Ins.	3.3	2.96	65.93	14.11	0.00	–

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 deal 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 6), according to the common editing rules explained in Section 2 that are:

- a rallye is displayed by a global view;
- a missed first serve is a global view of short duration followed by close-up views of short duration too (as the players do not have to change their positions) and followed by an other global view corresponding to the second serve;
- in a broadcast video, the producers notify the viewers that a replay is being displayed by inserting special transitions;
- a break is characterized by an important succession of close-up, public views and advertisements. This set of consecutive shots has a particular long duration. It appears when players change ends, generally every two games.

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 shot duration, the visual similarity between the shot keyframe and K_{ref}, and the audio vector which characterizes the presence or absence of the predetermined audio events. More formally, for a shot t, the

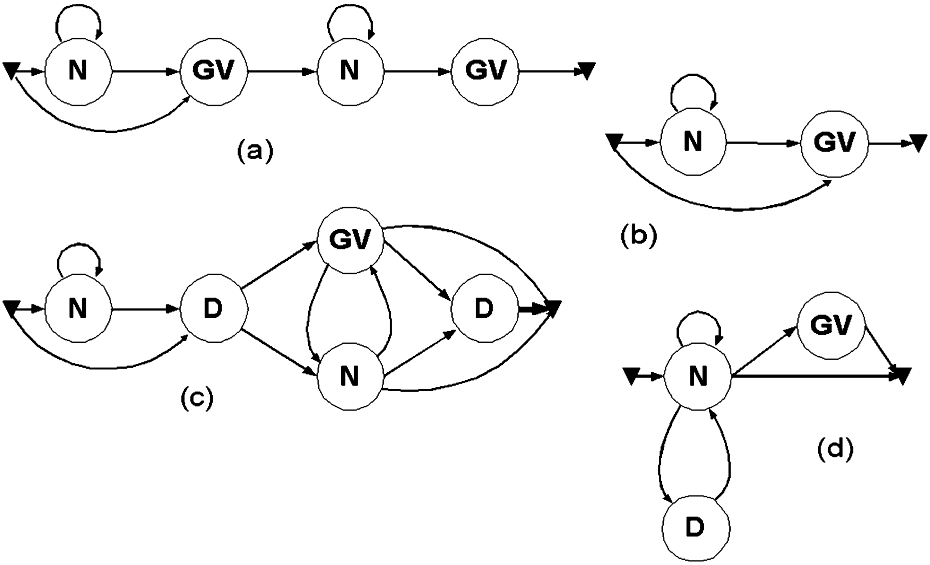


Fig. 6 Hidden Markov Models of the four basic structural units: (a) missed first serve and rally, (b) rally, (c) replay, (d) break. *GV* stands for Global View, *N* for Non-global view, and *D* for Dissolve transition

observation o_t consists of the shot duration d_t, the similarity v_t, and the audio description vector a_t.

The probability of an observation o_t conditionally to state j is then given by:

$$b_j(o_t) = p(d_t|j)^{\alpha} \ \ p(v_t|j)^{\alpha} \ \ P[a_t|j]^{\beta} \tag{9}$$

with $\alpha, \beta \in]0, 1[$, $2\alpha + \beta = 1$, and 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 $P[a_t(k)|j]$. Examples of the resulting probability distributions are given in figure 8 and Table 2 for missed first serve, rally and break states. As $P[a_t|j]$ is a product of discrete probabilities, its value is generally smaller than $p(v_t|j)$ and $p(d_t|j)$. The weighting coefficients are used to re-enforce the audio influence and are set to : $\alpha = 0.375$ and $\beta = 0.25$ (as $p(v_t|j)$,$p(d_t|j)$ and $P[a_t|j] \in [O, 1]$).

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:

$$\widehat{s} = \arg\max_{s} \ \ \ln p(s) \ \ + \ \ \sum_{t} \ln b_{s_t}(o_t) \tag{10}$$

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 7. This higher level represents the tennis syntax and the hierarchical structure of a tennis match, described in terms of points, games and sets. 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.

Several comments about the higher level 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. 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. It prevents also two breaks from being too much consecutive. The detection of breaks provides high level information about the video structure like "two games happen between two breaks." However, boundaries between games, and consequently game composition in terms of points, are not available at this stage. The observation streams used and the transition probabilities are not sufficient to determine if a game ended or not. To provide a segmentation in point, game and set, extra-information about the game status (like which player is the server) is needed. Such a feature is specific to tennis, and this issue is not addressed in this paper.

Table 2 Probability of each audio event conditionally to a given state, estimated respectively from manually and automatically segmented audio features

	Man. audio vectors					Segm. audio vectors				
States	Appl.	Speech	Tennis	Music	Silence	Appl.	Speech	Tennis	Music	Silence
First serve	0.06	0.9	0.84	0	0.74	0.09	0.8	0.55	0.02	0.62
Rallies	0.78	0.95	0.98	0	0.52	0.71	0.88	0.82	0.02	0.58
Break	0.21	0.82	0	0.52	0.3	0.32	0.82	0.15	0.37	0.35
Other	0.42	0.74	0.06	0	0.5	0.46	0.73	0.33	0.02	0.5

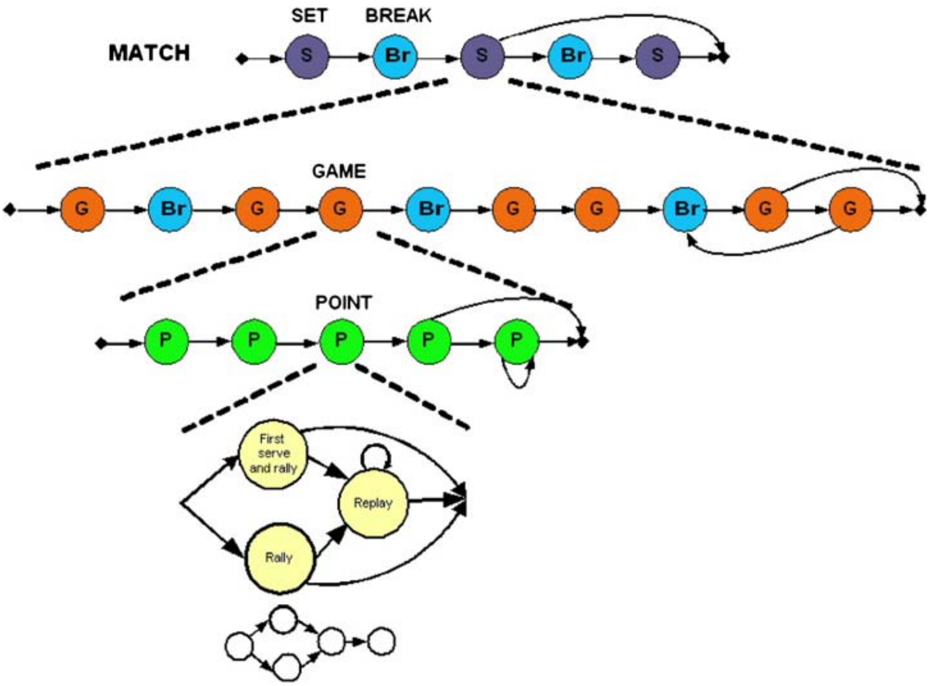


Fig. 7 Content hierarchy of broadcast tennis video

5 Experimental results

In this section, we describe the experimental results of the audiovisual tennis video segmentation by HMMs. Experimental data are composed of eight videos, representing about 5 h of manually labelled tennis video. The videos are distributed among three different tournaments, implying different production styles and playing fields. Three 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.

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. Precision and recall rates, as well as F-measure,[1] are given in Table 3 for the video of the testing set that gives the worst results. Precision is defined as the ratio of the number of shots correctly

[1] defined by: $F = \frac{2 \cdot \text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}$

Table 3 Precision (*P*), recall (*R*) and *F*-measure (*F*) for segmentation and classification with visual features only, manually segmented audio features only, and both manually segmented audio and visual features

	Visual features			Audio features			Audio visual features		
Classification	*P*	*R*	*F*	*P*	*R*	*F*	*P*	*R*	*F*
First serve	0.61	0.74	0.67	0.72	0.70	0.71	0.85	0.93	0.89
Rallies	0.57	0.48	0.52	0.80	0.64	0.71	0.89	0.80	0.84
Replay	0.48	0.70	0.57	0.36	0.35	0.35	0.56	0.65	0.60
Break	0.90	0.65	0.75	0.88	0.64	0.74	0.91	0.66	0.76

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 low recall and precision rate compared to missed first serve. In fact the similarity measure works well to discriminate global views from non-global views. Regarding only the shot classification into global views, the correct classification rate is 95%. 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 analysis of the interleaving of shots based on the shot duration failed.

Replay detection relies essentially on dissolve transition detection. Detecting dissolve transitions is a hard task, critical for replay 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%.

5.2 Using audio features only

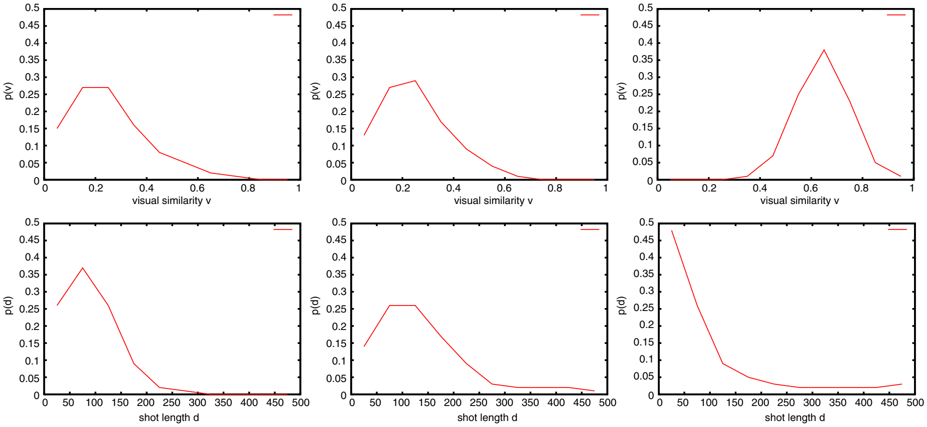
For each observation, only the audio vector and the shot duration are taken into account. Dissolve transitions are then not considered. The audio vectors represent the presence or absence of predetermined audio events in a shot without any measure of the event importance. This means that audio events, 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 (figure 8).

In a first experiment, audio probabilities are estimated from the manually segmented features. Figure 9 represents the average classification rate over the testing set 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. When β increases, the performance goes down as audio events become less representative of the audio shot content.

The increase in precision and recall rates for rallies and first missed serve (Table 3) 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



(a) missed first serve **(b) rally** **(c) break**

Fig. 8 Estimated probability distributions for (*from left to right*): missed first serve, rally, and break states. *Top to bottom*: visual similarity and shot length

of *ball hits* (see Table 2). 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

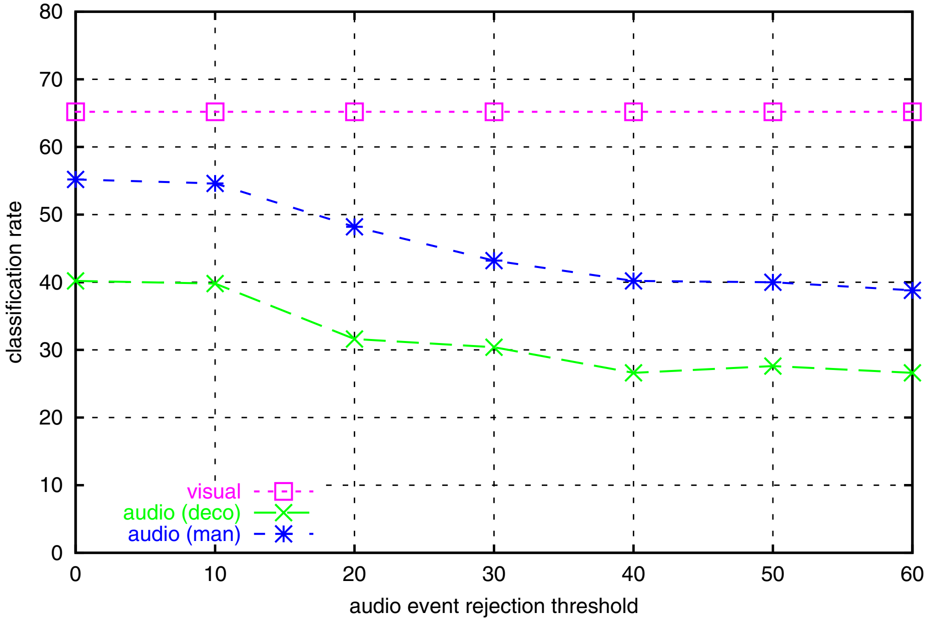


Fig. 9 Classification rate using manually (man) and automatically (deco) segmented audio features, when β varies from 0 to 60%

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, due to insertions of *ball hits* class (Table 1), and *music* is present in other states than break, as illustrated in Table 2.

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 10 shows the improvement of the decoding performance on automatically segmented vectors. However, the results are still less accurate than using visual features only.

5.3 Using both audio–visual features

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

Figure 11 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.

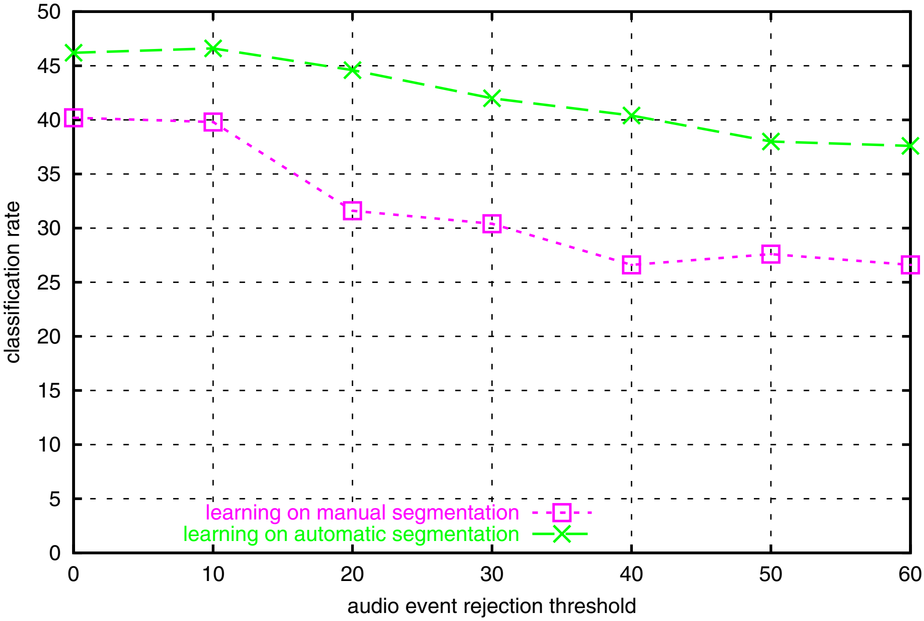


Fig. 10 Classification rate using automatically segmented audio features, when the learning is performed on manually or automatically segmented features

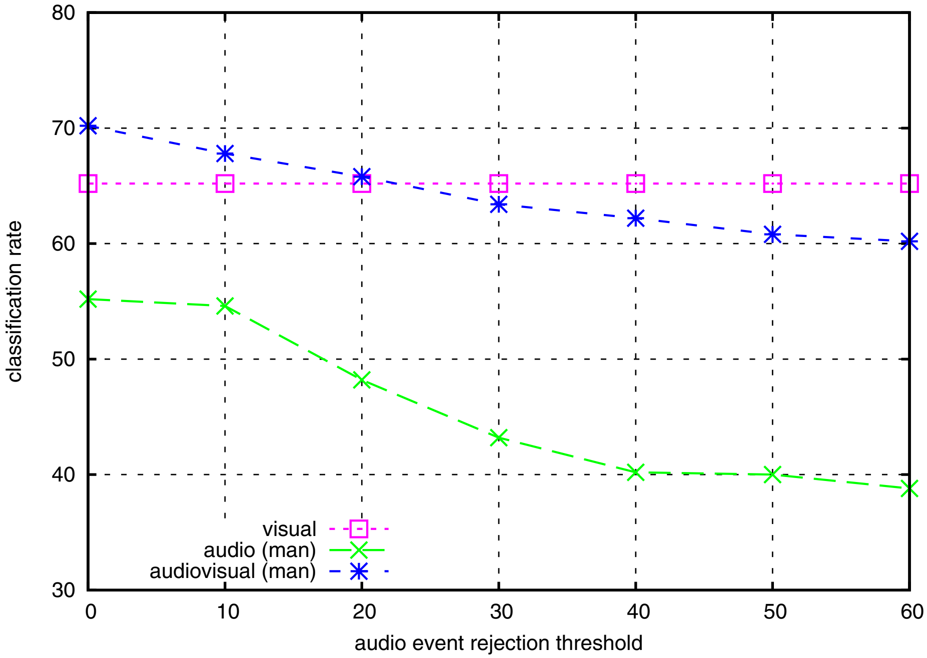


Fig. 11 Classification rate using both manually segmented audio features and visual similarity, when β varies from 0 to 60%

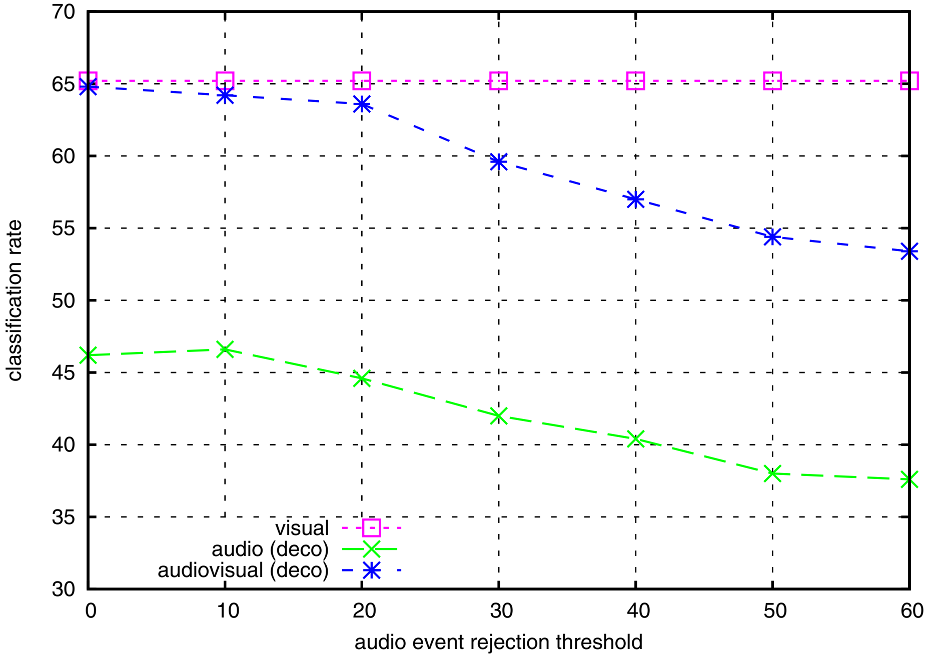


Fig. 12 Classification rate using both automatically segmented audio features and visual similarity, when β varies from 0 to 60%

Table 4 Precision (*P*), recall (*R*) and *F*-measure (*F*) for segmentation and classification 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

	Visual features			Audio features			Audio visual features		
Segmentation accuracy	0.65			0.65			0.86		
Classification	*P*	*R*	*F*	*P*	*R*	*F*	*P*	*R*	*F*
First serve	0.67	0.63	0.65	0.61	0.65	0.63	0.86	0.88	0.87
Rallies	0.78	0.47	0.58	0.91	0.55	0.68	0.94	0.86	0.90
Replay	0.63	0.81	0.71	0.49	0.47	0.48	0.57	0.66	0.61
Break	0.93	0.87	0.90	0.92	0.81	0.86	0.92	0.87	0.89

Figure 12 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 11 shows that fusing the audio and visual cues enhanced the performance, when audio features are good. Detailed classification rates are given in Table 3.

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.

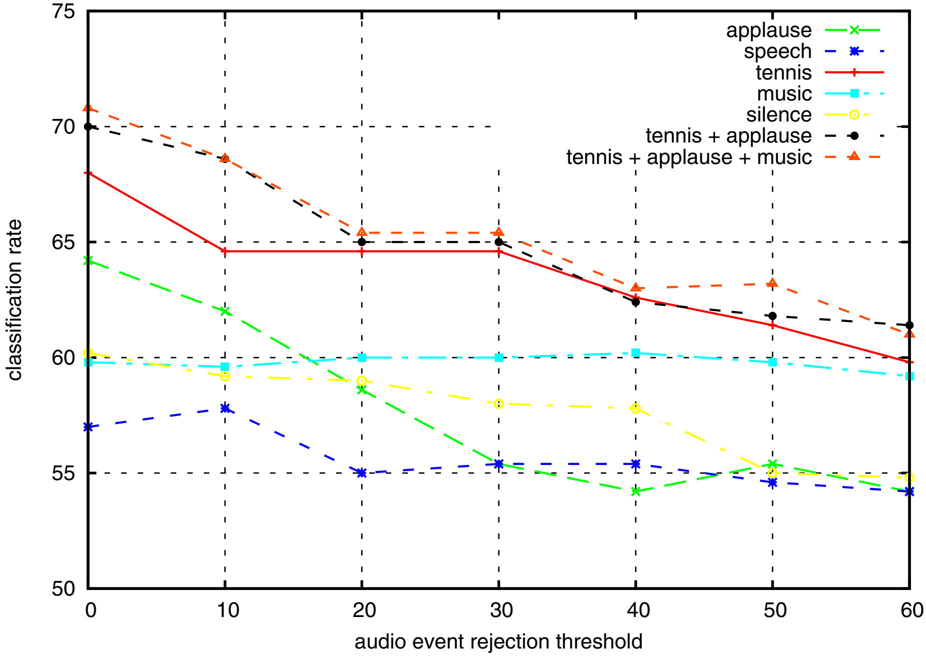
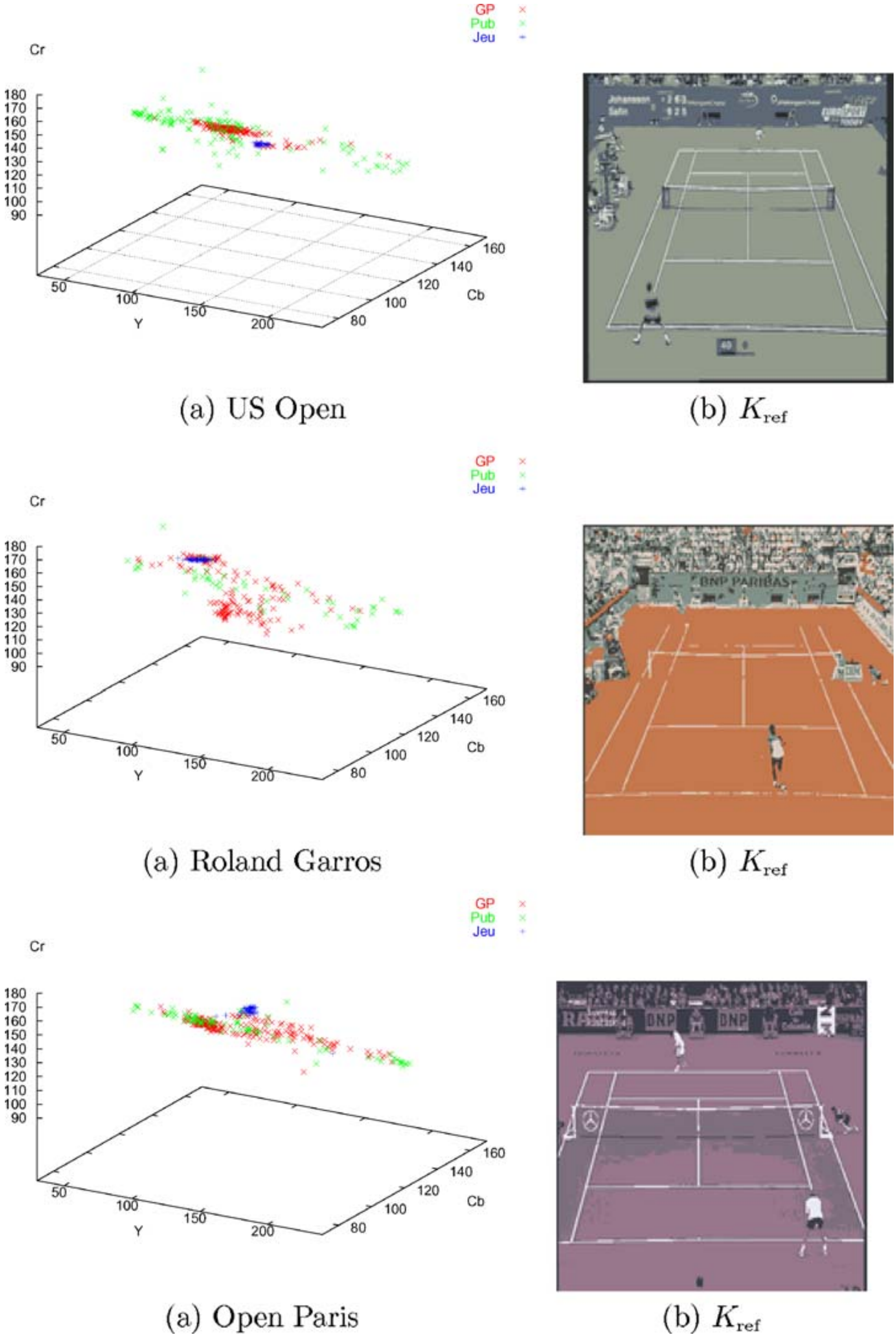


Fig. 13 Classification rate using both manually segmented audio features and visual similarity, when β varies from 0 to 60%, considering only some audio classes



(a) US Open

(b) K_{ref}

(a) Roland Garros

(b) K_{ref}

(a) Open Paris

(b) K_{ref}

Fig. 14 (a) Dominant color repartition of keyframes according to their type of view, for different videos. (b) selected K_{ref} for each video

It should be noticed here that the results are presented for bad conditions: a small learning set, and videos in the testing set belonging to tournaments excluded from the learning set. The performance and the audio influence significantly increase when learning and testing sets are reduced to the same tournament (Table 4).

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 consider only one class for the audio features a_t. Figure 13 confirms that *applause*, *ball hits* and *music* are the most relevant events. 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. Finally, the best performance is obtained by discarding *speech* and *silence* from audio vectors (figure 14).

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.

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*, *silence* 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 at a higher level, considering the context, and based on audio–visual cues simultaneously.

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. We are currently working on such approaches where a combined model is defined on low level features. We are also testing this approach on another sport, baseball, which is well-structured like tennis.

Another approach could be to integrate higher-level features, like score or players detection for the visual part, and keywords from automatic speech transcription for the audio part. Such more specific features would give very relevant information, but could tend to make the system even more application dependent.

References

1. Alatan AA, Akansu AN, Wolf W (2001) Multi-modal dialog scene detection using Hidden Markov Models for content-based multimedia indexing. Multimed Tools Appl 14(2):137–151

2. Betser M, Gravier G, Gribonval R, Bimbot F (2003, September) Extraction of information from video sound tracks—can we dectect simultaneous events? In: Third International Workshop on Content-Based Multimedia Indexing (CBMI03), pp 71–77
3. Chang P, Han M, Gong Y (2002, September) Extract highlights from baseball game video with Hidden Markov Models. In: Proc. of IEEE International Conference on Image Processing (ICIP02), Rochester, NY, USA
4. Dayhot R, Kokaram A, Rea N, Denman H (2003, April) Joint audio visual retrieval for tennis broadcasts. In: IEEE Int. Conference on Acoustics, Speech, and Signal Processing (ICASSP03), Hong Kong
5. Duan L-Y, Xu M, Tian Q (2003, January) Semantic shot classification in sports video. IS&T/SPIE storage and retrieval for media databases, SPIE-5021, pp 300–313
6. Hua W, Han M, Gong Y (2002, August) Baseball scene classification using multimedia features. In: IEEE International Conference on Multimedia and Expo (ICME02)
7. Huang J, Liu Z, Wang Y (1999, September) Integration of multimodal features for video scene classification based on HMM. In: Proc. of IEEE Workshop on Multimedia Signal Processing, Copenhagen, Denmark, pp 53–58
8. Jiang H, Lin T, Zhang H (2000, August) Video segmentation with the support of audio segmentation and classification. In: IEEE International Conference on Multimedia and Expo (I)(ICME00), Vol. 3, pp 1551–1554
9. Kawashima T, Tateyama K, Iijima T, Aoki T (1998, October) Indexing of baseball telecast for content-based video retrieval. In: IEEE International Conference on Image Processing (ICIP98), Vol. 1, pp 871–875
10. Kim K, Choi J, Kim N, Kim P (2002, July) Extracting semantic information from basketball video based on audio-visual features. In: Proc. of Int'l Conf. on Image and Video Retrieval, Vol. 2383, London, UK, Springer, Lecture Notes in Computer Science, pp 278–288
11. Lienhart R (2001) Reliable transition detection in videos: a survey and practitioner's guide. International Journal of Image and Graphics 1(3):469–486
12. Liu Z, Huang Q (1999, October) Detecting news reporting using audio/visual information. In: Proc. of IEEE International Conference on Image Processing (ICIP99), Vol. 1, pp 324–328
13. Rousseeuw PJ, Leroy AM (1987) Robust regression and outlier detection. Wiley and Sons, New York
14. Snoek CGM, Worring M (2003) Multimodal video indexing: a review of the state-of-the-art. Multimed Tools Appl (to appear)
15. Sudhir G, Lee JCM, Jain AK (1998, January) 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
16. Wang Y, Liu Z, Huang J-C (2000, November) Multimedia content analysis using both audio and visual cues. IEEE Signal Process Mag 12–36
17. Xie L, Chang S-F, Divakaran A, Sun H (2002, May) Structure analysis of soccer video with Hidden Markov Models. In: IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP02), Orlando, FL, USA
18. Xu P, Xie L, Chang S-F, Divakaram A, Vetro A, Sun H (2001, August) Algorithms and system for segmentation and structure analysis in soccer video. In: IEEE International Conference on Multimedia and Expo (ICME01), pp 928–931
19. Xu M, Duan L-Y, Xu C-S, Tian Q (2003, April) 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
20. Zhang HJ, Kankanhalli A, Smoliar SW (1993) Automatic partitioning of full-motion video. Multimedia Syst 1(1):10–28
21. Zhong D, Chang S-F (2001, August) Structure analysis of sports video using domain models. In: IEEE International Conference on Multimedia and Expo (ICME01), Tokyo, Japan
22. Zhou W, Vellaikal A, Kuo C-CJ (2000, November) Rule-based video classification system for basketball video indexing. In: Proc. ACM International Multimedia Conference, Los Angeles, California, pp 213–216

Ewa Kijak has been an assistant professor at Pierre et Marie Curie University (UPMC or Paris 6), working in the Computer Science Laboratory (LIP6) of the university since 2004. From 2000 to 2003, she worked at Thomson Corporate Research, in collaboration with the IRISA research unit, to achieve a Ph.D. in signal processing. Her research interests are video indexing and multimedia information fusion.

Guillaume Gravier is a research fellow at Centre National pour la Recherche Scientifique (CNRS), working at the Institut de Recherche en Informatique et Systèmes Aléatoires (IRISA). His research are in the fields of speech recognition, speaker recognition, audio indexing and multimedia information fusion. Guillaume Gravier also worked on speech synthesis at ELAN Informatique from 1996 to 1997 and on audiovisual speech recognition at IBM Research from 2001 to 2002.

Lionel Oisel received the Ph.D degree in electrical engineering from the University of Rennes 1, France in 1998. His research interests were in the field of 3D reconstruction and motion estimation. He joined the Thomson research group, Rennes France in 2000 and is currently a research staff member. He is investigating problems in multimedia indexing, object tracking and visual servoing.

Patrick Gros has been involved in research in the field of Computer Vision since 1990. After having finished his studies in Engineering Science at “École Polytechnique” and “École Nationale Supérieure de Techniques Avancées” in Paris, he joined the Fundamental Computer Science and Artificial Intelligence Laboratory (LIFIA) in 1990, to achieve a Ph.D. in computer vision. Since July 1993 and the defense of this thesis, he has had a research position at CNRS, still in LIFIA, which became GRAVIR since then. From November 1995 until October 1996, he was visiting research scientist at the Robotics Institute of Carnegie Mellon University in Pittsburgh, PA, USA working on a project of automatic landmark recognition for vehicles in urban environment. In July 1999, he moved from Grenoble to Rennes where he joint the IRISA research unit. In 2002, he founded a new research group devoted to multimedia document analysis and management, with a special emphasis on the problems raised by the management of very large volumes of documents.