Structuration multimodale d'une vidéo de tennis par modèles de Markov cachés

Ewa KIJAK[1,3], Guillaume GRAVIER[2], Lionel OISEL[1], Patrick GROS[3]

[1]THOMSON multimédia R&D France
1 avenue de Belle Fontaine, CS 17616, 35576 Cesson-Sévigné cedex, France

[2]IRISA - projet METISS [3]IRISA - projet TEXMEX
Campus Universitaire de Beaulieu, 35042 Rennes cedex, France
ewa.kijak@thomson.net, ggravier@irisa.fr
lionel.oisel@thomson.net, pgros@irisa.fr

Résumé – Cet article présente une méthode de structuration d'une vidéo utilisant des indices sonores et visuels. Cette méthode repose sur un modèle statistique de l'entrelacement temporel des plans de la vidéo. Le cadre général de la modélisation est celui des modèles de Markov cachés. L'approche est validée dans le cadre de vidéos de tennis télédiffusées. Les indices visuels sont utilisés pour caractériser le type des plans. Les indices audio décrivent les événements sonores apparaissant durant un plan. La structure de la vidéo est représentée par un modèle de Markov caché, intégrant les informations a priori sur le contenu de la vidéo, ainsi que sur les règles d'édition. En résultat du décodage, des éléments structuraux caractéristiques du tennis sont identifiés : premier service raté, échange, rediffusion ou temps mort. De plus, chaque plan de la vidéo est assigné à un niveau de hiérarchie décrit en terme de point, jeu et set. Cette classification et segmentation simultanées peuvent être utilisées pour la création de résumés vidéo ou pour permettre une navigation non linéaire dans le document vidéo.

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

1 Introduction

La structuration de la vidéo consiste à extraire les unités logiques qui composent une vidéo donnée. La structure qui doit être estimée dépend cependant de la nature de celle-ci. Dans cet article, nous nous intéressons à la structure des retransmissions sportives. L'analyse des vidéos de sports est un domaine de recherche motivé par le besoin des diffuseurs de vidéos d'une annotation détaillée de leur contenu. Cette annotation est généralement utilisée pour sélectionner les extraits susceptibles d'être diffusés dans des magazines ou afin de construire des résumés. Jusqu'à présent cette tâche est réalisée manuellement.

Les données télédiffusées possèdent plusieurs types de flux d'informations : l'image, le son, et parfois le texte. Pour extraire des informations de haut-niveau sémantique d'une vidéo, les algorithmes doivent être dédiés à un type particulier de vidéos, tel que le sport ou les journaux télévisés. Dans ce contexte, la plupart des approches utilisent individuellement les indices audio ou vidéo [1]. Seulement quelques travaux récents prennent en compte ces deux sources d'informations [2, 3]. Ces approches reposent sur l'utilisation successive des caractéristiques audio et vidéo. Dans un premier temps, les caractéristiques visuelles sont utilisées pour détecter les phases de jeu. Dans un deuxième temps, la mesure de l'excitation du commentateur ou du public permet de sélectionner les plans les plus intéressants [4]. Le schéma inverse est également proposé dans [5] : les exclamations de l'audience sont d'abord détectées pour localiser les événements importants. L'analyse de l'image est ensuite utilisée dans les régions précédemment sélectionnées afin d'identifier un événement particulier.

Dans cet article, nous présentons une méthode de classification et de segmentation automatique d'une vidéo de tennis qui combine simultanément les indices audio et vidéo. De façon générale, il existe deux approches pour combiner ces caractéristiques. L'une est de les combiner au sein d'un unique vecteur de caractéristiques audiovisuelles avant la classification. L'autre consiste à faire deux classifications indépendantes selon chaque modalité, puis de fusionner leur résultats. Dans ce travail, nous avons utilisé une stratégie intermédiaire qui consiste à extraire séparément des indices audio et vidéo de "haut niveau", puis à réaliser la classification en utilisant simultanément ces indices.

D'autre part, le tennis possède une structure temporelle bien déterminée, puisqu'il se décompose en sets, jeux et points, contrairement au football par exemple qui ne se décompose qu'en périodes de jeu contenant des phases de jeu ou de non-jeu. Nous exploitons la forte structure du tennis afin : (1) d'identifier chaque segment parmi l'une des quatre catégories suivantes : *premier service raté, échange, rediffusion* et *temps morts*, (2) d'identifier chacun de ces segments par son index à différents niveaux de la hiérarchie décrite en terme de point, jeu et set. Les modèles de Markov cachés (ou HMMs) sont utilisés

pour fusionner les informations audio-visuelles, et pour représenter la structure hiérarchique d'un match de tennis.

Nous définissons en section 2 la notion de syntaxe d'un match de tennis, qui a motivé notre approche sur l'analyse de la structure. En section 3 nous présentons le système global ainsi que les caractéristiques audiovisuelles exploitées. L'utilisation des HMMs est détaillée en section 4. Enfin, la section 5 présente les résultats expérimentaux.

2 Structure d'une vidéo de tennis

Règles de production : Dans le cadre de la retransmission télévisée, les évènements sportifs sont généralement soumis à des règles de réalisation spécifiques. Ces règles résultent :

- du nombre fini de caméras nécessaires pour retransmettre un évènement sportif
- de leur position souvent fixe et caractéristique de l'évènement filmé
- de leur utilisation : à un instant donné, le point de vue fournissant l'information la plus pertinente est sélectionné par le réalisateur.

A cause des règles de réalisation, les diffusions télévisuelles sportives sont composées de scènes caractéristiques produisant des motifs répétitifs, désignés par le terme de *syntaxe*. Pour le tennis, par exemple, durant un échange, le point de vue sélectionné est celui capturant la vue globale du terrain, tandis qu'entre les échanges, des plans rapprochés sur les joueurs ou sur le public seront préférés.

Connaissances a priori : Ce sont essentiellement les règles du tennis. Les informations exploitées sont diverses : nombre et position des joueurs, modèle du terrain, déroulement et structure du jeu.

Structuration de la vidéo : Nous utilisons les règles de production et les règles du tennis pour retrouver la structure d'un match dans la vidéo. Les règles de production sont interprétées en identifiant les différents plans, représentés par leur point de vue (vue globale du terrain, gros plan, publicité...), et en analysant leur entrelacement temporel pour identifier les scènes caractéristiques du tennis. Nous avons identifié quatres scènes caractéristiques composant une vidéo de tennis :

- *les échanges* : il s'agit de vues globales au terme desquelles un point est marqué.
- *les premiers services ratés* : ils sont caractérisés par une vue globale d'une courte durée, mais ne concluent pas un point.
- *les temps morts* : de durée significative, ils apparaissent lorsque les joueurs changent de côté (tous les deux jeux en général).
- *les rediffusions* : elles montrent la dernière action menée suivant un autre point de vue, ou au ralenti. Elles sont notifiées au téléspectateur par l'insertion de transitions spéciales.

L'identification de ces quatre scènes caractéristiques est une étape intermédiaire pour la segmentation en points, jeux et sets. Un point est en effet défini comme une séquence de scènes caractéristiques, un jeu comme une séquence de points et un set comme une séquence de jeux.

3 Présentation générale du système

Tout d'abord, une segmentation vidéo est réalisée en détectant les transitions abruptes et graduelles entre les plans vidéos. Pour chaque plan vidéo obtenu, sont extraits : une image-clé représentative du contenu du plan, les attributs vidéos (couleur et mouvement), et les attributs audio. Les plans vidéos étant plus pertinents pour une analyse de la structure que des plans audio, l'unité temporelle de base est le plan vidéo. Ainsi les caractéristiques audio sont extraites de façon à décrire le contenu sonore d'un plan vidéo. La séquence des plans vidéos caractérisés par leurs attributs audio et vidéo forme la séquence d'observation qui est décodée par un processus de HMMs (Figure 1).

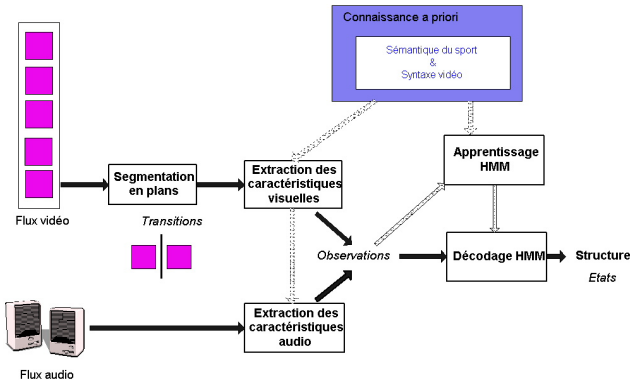

FIG. 1 – Système d'analyse de la structure temporelle

Pour un plan t et son image-clé associée K_t, les attributs utilisés sont :

La longueur du plan l_t : elle est définie comme le nombre d'images contenues dans le plan. Elle est directement obtenue à partir du processus de segmentation automatique en plans.

La similarité visuelle v_t : Lors d'un match de tennis, ce sont les vues globales qui capturent les échanges entre les joueurs. La similarité visuelle est utilisée pour identifier les vues globales parmi toutes les images-clé. D'abord, une image-clé K_{ref} représentative d'une vue globale est sélectionnée, sans aucune hypothèse sur la couleur du terrain. Une fois K_{ref} sélectionnée, pour chaque image-clé K_t de la séquence, la similarité visuelle v_t entre l'image-clé K_t et K_{ref} est calculée. v_t est définie comme une fonction pondérée du vecteur de couleurs dominantes de l'image-clé, de sa cohérence spatiale et de l'activité du plan défini comme la moyenne sur le plan des mouvements de la caméra [6].

La position des joueurs d_t : il s'agit de la position relative des joueurs par rapport au centre du terrain, représentée par la distance entre le centre de gravité du joueur et la ligne centrale (qui sépare les moitiés gauche et droites du terrain). Les joueurs sont détectés par une segmentation grossière (sous forme de blobs), réalisée par des filtrages sur les couleurs dominantes. Les lignes du terrain sont détectées par une transformation de Hough. Seul le joueur situé en bas du terrain n'est en fait considéré, car sa détection est plus robuste. Si le processus d'extraction de la position du joueur échoue, cet attribut n'est pas pris en compte pour le plan considéré. La position relative du joueur par rapport à la ligne centrale du terrain est utilisée comme un indice afin de déterminer si le serveur a changé entre

deux jeux consécutifs.

Un vecteur audio binaire a_t : il décrit quelles classes, parmi *parole, applaudissements, bruit de balle, bruit* et *musique*, sont présentes dans le plan vidéo. Ce vecteur binaire est extrait à partir d'une segmentation automatique de la bande sonore. Cette segmentation est réalisée par un système basé sur une modélisation par HMMs, utilisant des mélanges de Gaussiennes pour chaque classe de son, et pouvant combiner les classes.

4 Analyse de la structure

Nous avons défini quatre éléments structuraux de base qui correspondent aux quatre scènes caractéristiques décrites dans la section 2. Chacun de ces éléments est modélisé par un HMM, qui caractérise les relations temporelles entre les différents plans.

Chaque état d'un HMM modélise soit un plan, soit une transition progressive entre deux plans. Quatre symboles d'observation sont associés à chaque état. Formellement, pour un plan t, l'observation o_t consiste en la durée du plan l_t, la similarité v_t entre l'image-clé du plan et K_{ref}, la position du joueur d_t si elle existe, et le vecteur de description audio a_t qui caractérise la présence ou l'absence d'événements sonores prédéterminés. La probabilité de l'observation o_t d'être dans l'état j à l'instant t est alors donnée par :

$$b_j(o_t) = p(l_t|j)\ \ p(v_t|j)\ \ p(s_t|j)\ \ P[a_t|j] \qquad (1)$$

où $p(l_t|j)$ et $p(v_t|j)$ sont des distributions Gaussiennes, $p(s_t|j)$ est une fonction de d_t représentant la probabilité que le serveur ait changé par rapport au jeu précédent, et $P[a_t|j]$ est le produit sur l'ensemble des classes sonores k de la probabilité discrète $P[a_t[k]|j]$.

Afin de prendre en compte l'ensemble de la structure d'un match de tennis en terme de point, jeu et set, les quatres HMMs sont connectés au sein d'un HMM de plus haut niveau, représenté dans la Figure 2.

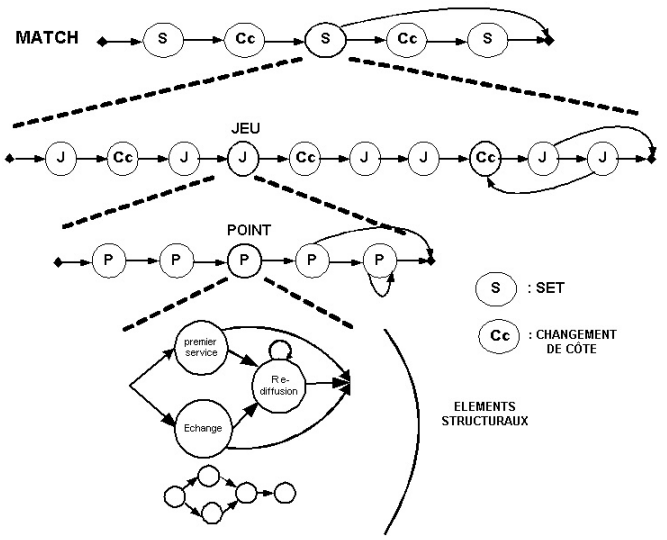

FIG. 2 – HMM représentant la structure hiérarchique d'un match de tennis

La segmentation et la classification de la séquence d'observation en éléments structuraux sont réalisées simultanément par un algorithme de Viterbi [7]. L'algorithme de Viterbi permet de calculer la séquence d'états Q réalisant le plus probablement la séquence d'observation O :

$$Q = \arg\max_q \ \ln p(q) \ \ + \ \sum_t \ln b_{q_t}(o_t) \qquad (2)$$

Les probabilités de transitions entre les états du HMM de plus haut niveau résultent entièrement des connaissances a priori, tandis que les probabilités de transition entre états, et les probabilité des symboles d'observations au sein des sous-HMMs qui le compose, sont estimées par apprentissage.

5 Résultats expérimentaux

Les données expérimentales sont composées d'environ 2-heures de vidéos de tennis au format MPEG-2 manuellement annotées. La moitié des données est utilisée pour entraîner le HMM, tandis que l'autre moitié est réservée pour les tests. Les taux de rappels et de précision sont donnés pour trois expériences (Table 1) : la première utilise uniquement la similarité visuelle et la position des joueurs, la deuxième uniquement les caractéristiques sonores et enfin la troisième combine les données audiovisuelles. Les durées des plans sont prises en compte dans toutes les expérimentations. En ce qui concerne les attributs audio, les résultats sont donnés pour les attributs issus de la vérité terrain (colonne "man." dans la Table 1) et pour ceux issus de la segmentation automatique (colonne "segm.").

Les informations visuelles seules permettent une bonne classification des scènes caractéristiques. La détection du joueur renforce la similarité visuelle, dans la mesure où le processus doit échouer sur les plans ne représentants pas des vues globales. La position du joueur à gauche ou à droite de la ligne centrale fournit un indice supplémentaire à l'entrelacement temporel, permettant de différencier les premiers services ratés des échanges.

Concernant les résultats de la classification utilisant uniquement des vecteurs audio fiables, la précision pour les premiers services ratés (90%) et les échanges (93%) montrent que les caractéristiques audio permettent de caractériser les plans d'échanges. En effet, un échange se caractérise par la présence de *bruits de balle* et d'*applaudissements*, tandis que les premiers services ratés sont seulement caractérisés par la présence de *bruits de balle*. De la même façon, les temps morts sont caractérisés par la présence ou l'absence de *musique* dans un plan. Les résultats sont équivalents à ceux obtenus avec la vidéo seule. En revanche, les rediffusions qui reposent essentiellement sur la détection des transitions progressives, ne sont pas caractérisées par un contenu audio représentatif, et par suite sont souvent non identifiées (rappel 26%).

Les résultats de la classification utilisant les vecteurs audio automatiquement segmentés sont en revanche plutôt mauvais. Le taux de rappel obtenu par la segmentation automatique de la bande sonore est de 76.9%, et la précision est de 46.3%. Les *bruits de balle* et la *musique* sont bien classifiés, tandis que le *bruit* est souvent étiqueté comme *bruit de balle*, probablement parce que la classe *bruit de balle* est caractérisée par un mélange de frappes de balles et de courts silences. Les erreurs issues de la segmentation automatiques se répercutent au niveau de la structuration et dégradent les performances.

TAB. 1 – Résultats de la classification et de la segmentation en éléments structuraux, avec les caractéristiques visuelles uniquement, les caractéristiques audio uniquement et les caractéristiques audiovisuelles

	Caractéristiques Visuelles		Caractéristiques Audio				Audio-visuelles			
			man.		segm.		man.		segm.	
Précision de la segmentation	84%		70%		48%		88%		80%	
Classification	precision	rappel	precision	rappel	precision	rappel	precision	rappel	precision	rappel
Premiers services	80%	83%	90%	68%	50%	19%	92%	88%	84%	88%
Échanges	88%	80%	93%	50%	18%	6%	87%	90%	83%	66%
Rediffusions	98%	88%	53%	26%	87%	34%	87%	96%	75%	78%
Temps morts	93%	87%	92%	84%	87%	79%	95%	84%	95%	84%

Cependant, l'intégration d'indices provenant de différents médias augmente les performance de la segmentation et de la classification. En particulier, la détection des bruits de balle améliore le taux de classification des échanges. De plus, les informations issues de l'audio permettent de corriger les erreurs qui peuvent se produire dans le processus d'extraction de la position du joueur, lorsque par exemple le joueur n'est pas détecté.

L'identification de la structure complète en terme de points et de jeu permet d'atteindre un niveau encore supérieur dans l'analyse de la structure (Table 2). La détection des frontières des points est hautement corrélée à la détection des scènes caractéristiques. Cependant, en l'absence d'information sur le changement de serveur au cours du jeu, l'algorithme de Viterbi ne peut déterminer de façon fiable les limites des jeux. La détection du joueur est donc un indice déterminant pour l'analyse de la structure hiérarchique du match de tennis. Toutes les frontières mal ajustées sont dues à des erreurs ou des ambiguïtés dans la position du joueur.

TAB. 2 – Identification de la structure hiérarchique

	Hierarchical segmentation accuracy
Point boundaries	83%
Game boundaries	75%

6 Conclusion

Nous présentons un modèle audio-visuel pour l'analyse de la structure temporelle d'une vidéo de tennis. La structure est modélisée par un HMM qui intègre l'information a priori sur le tennis et les règles d'édition. Les HMMs fournissent de plus un cadre probabiliste efficace pour l'intégration de données multimodales. La segmentation automatique des caractéristiques audio doit être améliorée. Une autre approche est explorée qui consiste à segmenter a priori la bande sonore en segments homogènes, puis à détecter de manière indépendante la présence ou non de chacun des événements sonores.

Nous étudions actuellement l'identification des scores incrustés dans les vidéos afin de les intégrer au système, soit comme une alternative à l'extraction des joueurs et du terrain, soit afin de compléter le processus de structuration. L'approche doit également être validée sur un nombre plus important de vidéos. En perspective, le modèle doit être adapté à d'autres séquences sportives telles que le baseball.

Références

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

[2] W. Hua, M. Han, and Y. Gong. Baseball scene classification using multimedia features. In *Proc. of IEEE Int'l Conf. on Multimedia and Expo (ICME)*, 2002.

[3] Q. Huang, Z. Liu, A. Rosenberg, D. Gibbon, and B. Shahraray. Automated generation of news content hierarchy by integrating audio, video, and text information. In *Proc. of IEEE Int'l Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, 1999.

[4] B. Li and I. Sezan. Event detection and summarization in american football broadcast video. In *SPIE Storage and Retrieval for Media Databases*, 2002.

[5] K.Kim, J. Choi, N. Kim, and P. Kim. Extracting semantic information from basketball video based on audio-visual feature. In *Proc. of Int'l Conf. On Image and Video Retrieval*, 2002.

[6] E. Kijak, L. Oisel, and P. Gros. Temporal structure analysis of broadcast tennis video using hidden markov models. In *SPIE Storage and Retrieval for Media Databases*, January 2003.

[7] L.R. Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2) :257–286, 1989.