Segmentation temporelle de vidéos numériques fondée sur l'utilisation de mosaïques 1D

A. Manoury et H. Nicolas

IRISA/INRIA, projet TEMICS, Campus de Beaulieu, 35042 Rennes Cedex

anne.manoury@irisa.fr, henri.nicolas@irisa.fr

Résumé – *Cet article présente une nouvelle méthode permettant de segmenter temporellement une séquence vidéo. Trois niveaux de décomposition sont introduits: les plans-séquences, les scènes et les hyper-scènes. Chaque plan est caractérisé à l'aide de deux images mosaïques 1-D obtenues à partir des vecteurs de mouvement du flux compressé MPEG. Deux types de distances, l'une globale et l'autre spatiale, sont définies de manière à évaluer la distance entre les plans et/ou les hyper-scènes. Un critère de décision basé sur ces distances permet alors la création des hyper-scènes. Les résultats expérimentaux ont permis d'obtenir la création d'hyper-scènes cohérentes en terme de contenu.*

Abstract *– This paper presents a new method which allow a temporal segmentation of a video sequence. Three decomposition levels are introduced: scene shots, scenes and hyper-scenes. Each shot is characterized using two 1-D mosaic images obtained using the MPEG motion vectors. Two kinds of distances (global and spatial) are defined to evaluate the shot and/or hyper-scenes distances. A decision criterion, based on these distances, is therefore used to create new hyper-scenes. Experimental results show that coherent hyper-scenes can be obtained using the proposed technique.*

1. Introduction

Les terminaux numériques de réception et d'enregistrement de vidéos ont atteint une capacité de stockage telle que, pour en faciliter l'utilisation, il est nécessaire de mettre en oeuvre de nouveaux outils. En particulier, un soin important doit être porté au techniques d'indexation, facilitant la navigation dans la base de données des vidéos ainsi que dans les vidéos elles-mêmes. L'utilisateur doit pouvoir choisir de regarder tout ou une partie de la vidéo, mais aussi n'en visionner qu'un résumé ou la parcourir selon des thématiques fixées. Pour se faire, le contenu de la vidéo doit être étayé par une structure temporelle.

C'est dans cette optique que se situe la méthode de construction d'une structure temporelle vidéo proposée dans cet article. Cette méthode a pour objectif le regroupement des images de la vidéo en *plans*, eux même formant des *scènes* puis *des hyper-scènes*. Les scènes sont définies comme le regroupement de plans consécutifs ayant un contenu proche au sens d'un certain critère. Les hyper-scènes contiennent les scènes de contenus communs mais éloignées temporellement.

La mise en œuvre algorithmique de l'étape de construction de la structure d'une vidéo donnée implique la définition d'un critère de similarité que l'on appliquera à des composantes caractéristiques des plans. En pratique, on utilisera ici comme caractéristique des plans les *mosaïques1-D* obtenues directement à partir du flux MPEG2.

Dans le paragraphe 2 de cet article nous définirons les objectifs et les propriétés de la structure proposée. Nous présenterons ensuite la notion de *mosaïque 1-D* comme représentante d'un plan séquence. Le paragraphe 4 décrit la méthode de classification proposée. Finalement, le paragraphe 5 donne quelques résultats expérimentaux et le paragraphe 6 conclu sur quelques perspectives.

2. Structure d'une vidéo

Une navigation aisée à l'intérieur d'un document vidéo nécessite que son contenu soit étayé par une structure non-linéairement dépendante du temps. Il s'agit notamment de permettre à un utilisateur de ne regarder qu'un thème sélectionné redondant dans la vidéo (il doit par exemple pouvoir choisir de ne visionner que des scènes d'extérieur dans un documentaire).

La segmentation temporelle de la vidéo est une étape nécessaire dans la définition d'une telle structure. Elle est usuellement obtenue en détectant les changement de plans dans le flux vidéo. Plusieurs auteurs se sont intéressés à ce problèmes et des états de l'art sont donnés dans [1,2] .

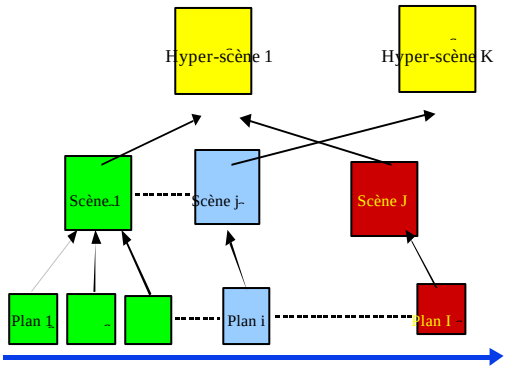

FIG. 1 : Schéma de structuration d'une vidéo en plans, scènes et hyper-scènes.

La méthode que nous proposons permet d'aller plus loin dans la structuration de la vidéo en introduisant une structure hiérarchique à trois niveaux : les plans, les scènes et les hyper-scènes. La figure 1 présente cette structure. Les plans sont les plus petits éléments de cette structure, ils sont composés d'un ensemble d'images qui ont été filmées de façon continu. Les scènes sont composées d'ensembles de plans consécutifs considérés comme similaires au sens d'un certain critère. Enfin, les hyper-scènes sont composés d'un ensemble de scènes ou de plans similaires mais non consécutifs. Nous considérerons dans la suite de l'étude que les plus petits éléments d'analyse sont les plans et que la segmentation de la vidéo en plan est disponible a priori. Le paragraphe suivant présente la notion de mosaïque 1-D utilisée pour caractériser les plans.

3. Représentation des plans à l'aide de mosaïques 1-D

Un plan séquence contient un ensemble d'images très redondant en terme de contenu, il n'est donc a priori pas utile d'utiliser l'ensemble de ces données images pour caractériser un plan. De plus, l'ensemble des images représente un volume de données important, et, donc généralement une complexité opératoire plus grande. Une solution pour représenter un plan consisterai à choisir comme image clé une ou éventuellement plusieurs images de ce plan. Cela pose alors le problème du choix de ces images. De plus, une image clé ne représente généralement pas la totalité du contenu du plan, dont une partie du contenu n'est alors pas considérée.

La caractérisation du contenu d'un plan via une représentation de type mosaïque apparaît alors comme intéressante dans la mesure où les mosaïques sont, par définition, des images construites de façon à contenir l'ensemble de l'information disponible sur l'arrière-plan de la séquence. Cela signifie alors que les objets en mouvement en avant-plan du fond ne sont pas pris en compte, à moins de construire une mosaïque particulière pour chacun d'entre eux. Dans le contexte de cet article, nous nous limiterons à une mosaïque représentant le fond fixe des plans, ce qui suppose que la classification en scènes et hyper-scènes obtenue se base exclusivement sur les caractéristiques de ces arrière-plans. La construction de ces mosaïques suppose la suppression des objets en mouvement ainsi que la compensation des mouvements induits par la caméra. L'utilisation de mosaïques pour l'indexation et la représentation vidéo est présentée dans [3,4].

La phase de classification revient alors ici à regrouper des mosaïques similaires au sens d'une certaine distance. Dans certain cas (si les plans comparés sont similaires), ce regroupement peut conduire à la création d'une nouvelle mosaïque par compensation des paramètres de mouvement, de la même façon que la mosaïque initiale a été crée progressivement par traitement des images successives contenus dans la séquence.

Afin de réduire la quantité de données à traiter, Dupuy *et al.* [5] définissent et utilisent des objets mosaïques 1-D correspondants à des projections sur un axe donné (typiquement selon les axes horizontaux et verticaux) d'une mosaïque 2-D représentative du fond d'un plan. En pratique, il est possible de créer de tels mosaïques avec une précision suffisante pour les objectifs de structuration de la vidéo en scènes en utilisant directement les paramètres de mouvement contenu dans le flux MPEG.

Nous utilisons alors ici pour caractériser chaque plan deux mosaïques 1-D, correspondant à des projections horizontales et verticales et que nous appellerons *mosaïque 1+1D*. La figure 2 présente un exemple de mosaïque1+1D et la mosaïque 2-D correspondante.

FIG. 2 : mosaïque1+1-D et 2-D d'un plan séquence de la vidéo « Aquaculture ».

4. Classification des plans vidéos

L'objectif consiste à définir un algorithme de structuration générique, c'est à dire capable de traiter des vidéos de types très différents, tel que, par exemple, des documentaires, des journaux télévisés, des émissions sportives ou des films. De ce fait, aucun information sur la vidéo ou sur les mosaïques n'est considéré comme a priori disponible, ce qui suppose l'utilisation d'algorithmes de classification non supervisée. Parmi ces algorithmes, notre choix s'est arrêté sur une méthode de classification hiérarchique ascendante. Cette méthode permet de classer les éléments suivant des partitions de moins en moins fines jusqu'au critère d'arrêt [6].

Dans la méthode proposée ici, l'algorithme est initialisé en affectant un plan à chaque classe hyper-scènes. Les classes ainsi définies seront alors regroupées les unes aux autres selon le critère de décision explicité ci-dessous. L'arrêt de l'algorithme, déterminant le nombre d'hyper-scènes composant la structure de la vidéo aura lieu quand la mesure de similarité aura atteint une valeur critique.

4.1 Mesures de similarités

La principale difficulté de la méthode de segmentation proposée réside dans la définition de la notion de « thème». Le sens de ce terme n'a pas pour tous la même interprétation et donne lieu sur des exemples concrets à des classifications très subjectives variant d'une personne à une autre. Le résultat obtenu par une approche automatique sera donc inévitablement d'autant plus subjective que le nombre d'hyper-scènes obtenu est réduit.

La phase de regroupement des plans et hyper-scènes se base fondamentalement sur une notion de distance. Nous définissons ici les deux catégories de distances suivantes :

- les distances globales
- les distances spatiales

Les **distances globales** sont proches de certaines distances utilisées en segmentation de vidéo comme les comparaisons d'histogrammes[1,8], elles mesurent le degré de ressemblance entre des paramètres dépendants des couleurs présentes dans les mosaïques.

La distance globale que nous proposons est une mesure de l'homogénéité des mosaïques. L'algorithme permettant de la calculer est le suivant :

- Les vecteurs mosaïques sont découpés en un même nombre de segments.
- Des caractéristiques statistiques (la moyenne a été utilisée dans les expérimentations présentées ici) sont calculées pour chaque segment.
- A chaque segment de la mosaïque 1+1D du premier plan, on fait correspondre un unique segment dans la mosaïque 1+1D du second plan.
- La norme L2 est utilisée pour le calcul de la distance entre deux segments.
- La distance entre les mosaïques 1+1D est alors celle qui minimise la somme des distances entre segments, calculées pour toutes les partitions possibles.

Dans la pratique, le nombre de segments doit être choisi très petit (de l'ordre de 4) pour d'une part respecter la volonté d'avoir une mesure caractérisant le comportement « global » de la mosaïque et d'autre part pour ne pas produire un coup de calcul excessif (on a N! partitions à calculer, où N est le nombre de segment dans chaque mosaïque 1+1D).

Les distances spatiales représentent l'ensemble des distances utilisées conventionnellement (corrélation, norme1, norme2, Mahalonobis, Battacharya). Nous les appliquons aux mosaïques après compensation du mouvement estimé entre elles. Le modèle de mouvement utilisé correspond uniquement à des mouvements de translation et zoom. Un algorithme de type « recherche exhaustive » est utilisé pour l'identification de ces paramètres. Une mosaïque 1+1D caractéristique de l'hyper-scène est alors construite lorsque la distance obtenue est suffisamment faible. Cette « hyper-mosaïque » servira par la suite de représentant de l'hyper-scènes.

4.2 Critère de décision

Le critère de décision permettant de regrouper entre eux des plans et/ou des hyper-scènes est basé sur les deux distances définies ci-dessus. Pour cela, un **critère de décision par seuillages successifs** est défini de la manière suivante : la distance globale, beaucoup moins coûteuse en temps de calcul, est d'abord calculée. Le processus de décision de regroupement des plans et/ou des hyper-scènes suit l'algorithme suivant :

- si la valeur du critère se situe en deçà d'un *seuil de regroupement*, on considère que les plans appartiennent à la même classe.
- si cette valeur est supérieur à un *seuil de rejet* les plans sont définitivement considérés comme faisant partie de classes différentes.
- entre ces deux seuils, la décision est prise selon la valeur de la distance spatiale et sa position par rapport à un seuil.

Cette approche revient à considérer que la distance globale permet les regroupements ou les rejets les plus évidents, ce qui permet de restreindre l'utilisation de la distance spatiale, plus coûteuse en terme de calcul.

4.3 Stratégie d'agrégation

La distance globale permet d'éviter des regroupement d'hyper-scènes de contenus trop éloignés. Afin de conserver cet aspect lors du calcul de la distance entre deux hyper-scènes, nous avons défini la distance globale entre deux hyper-scènes de la manière suivante : la distance entre deux Hyper-scènes est la plus grande distance séparant un élément (ou plan séquence) de la première hyper-scène d'un élément de la seconde. Lorsque la distance spatiale est inférieure au seuil de décision, on peut construire une hyper-mosaïque caractéristique de l'hyper-scène. Le calcul des distances entre hyper-scènes sera alors le calcul des distances entre ces nouvelles mosaïques. Si on ne peut pas construire cette hyper-mosaïque, on la définie comme l'ensemble des hyper-mosaïques représentantes de chacune des hyper-scènes. Le calcul des distances spatiales se fait alors deux à deux sur chaque mosaïque.

5. Résultats expérimentaux

Les résultats présentés dans ce paragraphe sont obtenus pour un sous ensemble représentatif de la vidéo «aquaculture en méditerranée». La figure 3 montre les résultats obtenus (chaque plan est représenté par une imagette manuellement sélectionnée). Les 21 plans sont regroupés en 5 hyper-scènes globalement homogènes. Il est à noter que si les trois dernières semblent moins homogènes, cela vient du fait qu'elles contiennent des plans subissant des variations de contenu important non représenté par les imagettes.

6. Conclusions et perspectives

Nous avons présenté dans cet article, une nouvelle méthode permettant de structurer une vidéo. Pour réduire la quantité d'information à traité, nous avons utilisé une représentation des plans en mosaïque 1D construites directement à partir des vecteur de mouvement MPEG2. Ces mosaïques définissent alors différentes classes, selon leurs similarité. Deux mesures de similarité : global et par compensation des mouvements sont utilisés.

Les résultats expérimentaux obtenus sur une vidéo complexe sont encourageants. Nous proposons dans la suite de ce travail d'augmenter le nombre de distance en particulier en ajoutant une catégorie de distances sémantiques et une distance audio qui corroborerait ou invaliderait les décisions prises grâce aux mesures vidéo. Une stratégie de décision multicritère serait alors utilisé. Nous proposons de plus de nous intéresser à un critère de mesure de la qualité de la classification obtenue : en effet, calculer les pourcentages de faux positif/vrai positif suppose que l'on connaisse a priori la structure que l'on souhaite obtenir ce qui est difficile en raison du caractère subjectif et sémantique des regroupements.

Remerciements

Ce travail est réalisé dans le cadre du projet RNTL DOMUS VIDEUM, dans lequel nous sommes associés plus particulièrement à William Dupuy et Dominique Barba de l'équipe IVC de l'IRCCyN et à Jenny Benois du LABRI. Nous remercions tous ces collaborateurs pour leur aide précieuse.

Références

[1] I. Koprinska and S Carrato, "Temporal Video Segmentation: A Survey", *Signal Processing : Image Comunication*. V.16(2001), pp.477-500.

[2] P. Aigrain, H.J. Zhang, and D. Petkovic.- Content-based representation and retrieval of visual media : a state-of-the-art review

[3] H. Nicolas "New Methods for dynamic mosaicking". *IEEE Transactions on Image Processing*, February 2001.

[4] M. Irani and P. Anandan, "Video indexing based on mosaic representation " *IEEE Trans. on PAMI*, 86(5):905--921, May 1998.

[5] W.Dupuy, J.Benois -Pineau, D.Barba, "Outils pour l'analyse et l'indexation vidéo basée sur l'approche du signal 1D dans le domaine de la transformée Mojette", RFIA'2002, Angers, France, pp 337-386, 8-10 January 2002

[6] I. Koprinska and S Carrato, "Temporal Video Segmentation: A Survey"

[7] Classification and Regression Trees Leo Breiman, Charles J. Stone, Richard A. Olshen, Jerome H. Friedman

[8] J.C.M. Lee, Q. Li, and W. Xiong. Automatic and Dynamic Video Manipulation. In B. Furht, editor,Handbook of Multimedia Computing, 1998

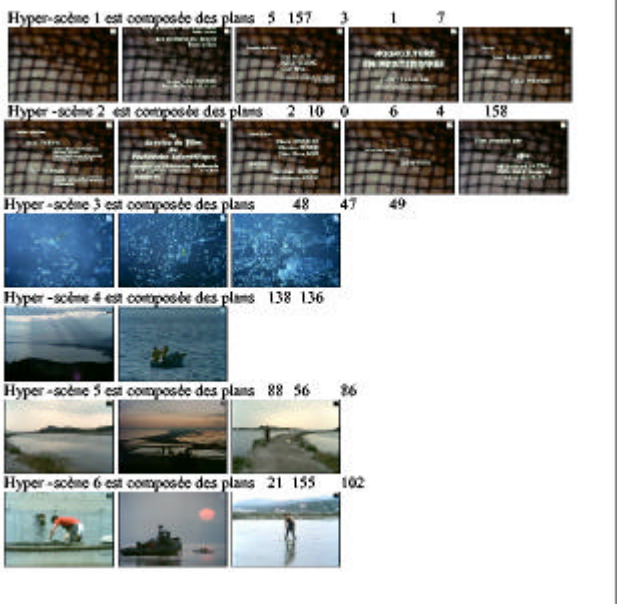

FIG. 3 : Résultat de classification.