Création de résumés de vidéos appliquée à la recherche par l'exemple

M. GUIRONNET[1], D. PELLERIN[1] et P. LADRET[1]

[1]LIS, Laboratoire des Images et des Signaux, INPG, 46 avenue Félix Viallet, 38031 Grenoble Cedex

Mickael.Guironnet@lis.inpg.fr

Résumé – Cet article décrit une méthode pour créer des résumés de vidéos possédant deux niveaux de résolution (intra-plan et inter-plan), à partir d'un descripteur couleur ou mouvement. Notre approche, applicable à tout autre type de descripteurs, consiste à extraire des images clés par regroupement de similarité suivant l'index considéré. Le regroupement non supervisé est réalisé par l'algorithme c-moyens flous (fuzzy c-means). Les résumés de vidéos ainsi obtenus permettent d'effectuer une recherche par l'exemple par combinaison d'index. La fusion est basée sur la théorie des ensembles flous qui définit la notion d'intersection des ensembles flous par l'utilisation d'une t-norme. Les résultats obtenus montrent que la recherche par l'exemple est améliorée par la combinaison des index.

Abstract – This article describes a method to create video summaries, on two levels of resolution (intra-shot and inter-shot), from a color or motion descriptor. Our approach, applicable to all types of descriptors, consists in extracting keyframes by similarity clustering according to the considered index. Unsupervised clustering is carried out by the fuzzy c-means algorithm. Thus the obtained video summaries allow to achieve a query by example by combination of indexes. Fusion is based on the theory of fuzzy sets which defines the notion of intersection of the fuzzy sets by the use of a t-norme. The obtained results show that the query by example is improved by the combination of indexes.

1. Introduction

La quantité d'informations audiovisuelles s'est accrue de façon spectaculaire avec l'apparition de l'Internet à haut débit et de la télévision numérique. Retrouver un document parmi cette masse d'information est devenu une tâche complexe et difficile. L'indexation de vidéos tente alors de faciliter l'accès automatique et rapide à l'information dans de grandes bases de vidéos. Une description pertinente de la vidéo est une tâche difficile. La méthode simple qui consiste à associer manuellement des mots clés est la plus répandue. Cependant, la recherche de vidéos est limitée à ces mots clés et ne permet pas de retrouver une vidéo par son contenu audiovisuel. Pour ces raisons, beaucoup de chercheurs essaient d'enrichir la description de la vidéo. Les travaux actuels [1][2] emploient généralement des critères de bas niveau pour décrire une séquence d'images, mais encore très peu [3] abordent le problème de la combinaison de ces critères.

Nous présentons dans cet article une méthode de création de résumés de vidéos à deux niveaux de résolution. L'approche est indépendante du type de descripteurs et repose sur un algorithme de regroupement par similarité non supervisé : « fuzzy c-means » (FCM). Les résumés de vidéos ainsi obtenus permettent d'effectuer une recherche par l'exemple par combinaison d'index. La fusion d'index hétérogènes, réalisée grâce à la théorie des ensembles flous, est générale et pourrait être appliquée à un nombre quelconque de descripteurs.

Dans la section 2, nous présentons les deux descripteurs (couleur et mouvement) qui vont permettre de représenter le contenu de chaque image de la vidéo. La section 3 explique la méthode de construction de résumés de vidéos. Dans la section 4, des résultats sont exposés sur le regroupement par similarité. La section 5 illustre la technique de fusion d'index pour améliorer la recherche par l'exemple. Enfin la dernière section conclut l'article.

2. Extraction de l'information couleur et mouvement

Pour caractériser la vidéo, nous avons choisi d'extraire des informations couleur et mouvement. En effet, la couleur est un descripteur qui offre un bon compromis entre la qualité des résultats obtenus et la complexité. Le mouvement, quant à lui, est indispensable pour représenter le contenu dynamique de la vidéo.

La couleur : Parmi les nombreux descripteurs couleurs existants [4][5], nous avons retenu l'histogramme couleur qui offre une grande simplicité. L'espace couleur choisi est l'espace YCbCr, très employé dans les formats de compression MPEG ½. Il est quantifié uniformément en 8x8x8 bins.

Le mouvement : Le descripteur mouvement est encore peu utilisé en indexation vidéos. Les méthodes existantes d'analyse du mouvement sont généralement complexes et coûteuses en calculs. De nombreux travaux [6] supposent alors une transformation affine entre deux images successives. Parfois il est préféré une mesure d'activité [7], moins coûteuse en calculs, pour décrire le contenu dynamique. Nous avons choisi une méthode récente d'estimation du mouvement à base d'ondelettes [8] qui permet de caractériser le mouvement par un jeu de coefficients. L'approche permet une représentation multi-échelle et compacte du mouvement. La connaissance fine du

mouvement n'étant pas nécessaire pour l'indexation vidéos, son estimation est réalisée sur des images sous-échantillonnées (imagette 72x88 pixels) pour accélérer les traitements. La méthode offre une signature du mouvement sous la forme de 162 coefficients (81 coefficients pour chaque composante de vitesse).

3. Méthode de construction de résumés vidéo

Pour pouvoir naviguer dans une base de vidéos ou rechercher un extrait de film dans une séquence, nous proposons de construire un résumé vidéo possédant deux niveaux de résolution. Le premier niveau de résolution (résolution fine) est obtenu par une première étape d'extraction d'images clés au sein de chaque plan de la vidéo par regroupement de similarité (Figure 1). Le plan représente une portion de vidéo filmée continûment sans effets spéciaux ni coupure. Plusieurs techniques performantes de découpages d'une vidéo en plans existent [9], nous supposerons donc dans cet article que les plans sont connus. Le second niveau de résolution (résolution plus grossière) est obtenu par une étape de regroupement par similarité des images clés précédentes.

L'algorithme de regroupement par similarité choisi est le « fuzzy c-means » (FCM). Il présente l'avantage de fournir à chaque élément (ou vecteur caractéristique extrait des images) un degré d'appartenance par rapport à chaque groupe formé. De plus, la théorie des ensembles flous est bien adaptée à la manipulation et la fusion d'informations hétérogènes. Ces propriétés sont exploitées lors de la phase de recherche par l'exemple pour combiner les index.

Soient $\{x_1,x_2,\ldots,x_n\}$ les vecteurs caractéristiques en terme de couleur ou de mouvement, où n correspond au nombre d'images dans le plan considéré. Le principe de l'algorithme est de minimiser une fonction objective :

$$J_m=\sum_{i=1}^{n}\sum_{k=1}^{K}(u_{ki})^m \; d^2(x_i,v_k) \qquad (1)$$

où u_{ki} est le degré d'appartenance de l'élément x_i au groupe k, K est le nombre de groupes cherchés, v_k est le centre de gravité du groupe k, d(·,·) est la distance euclidienne et m est un paramètre supérieur à 1 (m=1.25). Le poids de l'exposant m détermine la quantité de flou dans la partition. Le flou de la partition augmente avec m.

Afin d'obtenir une recherche non supervisée de groupes, l'index de Xie et Beni [10] est utilisé. Il fournit une mesure de validité sur la bonne compacité et séparabilité des groupes. La méthode consiste à déterminer le nombre de groupes K optimal en évaluant un critère (2) suivant la partition obtenue avec l'algorithme FCM.

$$\nu=\left\{(1/K)\sum_{k=1,K}\sigma_k^{\,2}\right\}/\{D\min\}^2 \qquad (2)$$

$$\sigma_k^{\,2}=\sum_{i=1,n}u_{ik}^{\,m}\,d(x_i,v_k) \qquad (3)$$

Plus la valeur du critère ν est faible, meilleure est la partition des données. Le nombre de groupes K cherché varie de 2 à K_{max}, K_{max} étant fixé suivant la longueur du plan étudié. Nous retenons la partition qui a le critère ν le plus faible. Nous choisissons le nombre de groupes K égal à 1 seulement si la taille du plan est inférieure à un seuil fixé (25 images) ou si la variance des éléments contenus dans le plan est inférieure à un seuil prédéfini. Ainsi cette étape (étape 1) permet d'extraire comme images clés, les images les plus proches des centres de gravité des groupes (Figure 1).

L'étape 1 fournit un résumé vidéo intra-plan qui pourrait être suffisant pour accomplir une recherche par l'exemple. Pour accélérer les traitements en réduisant la taille du résumé et donc le nombre de comparaisons à effectuer, nous proposons d'appliquer à nouveau la méthode précédente pour regrouper les images clés (étape 2).

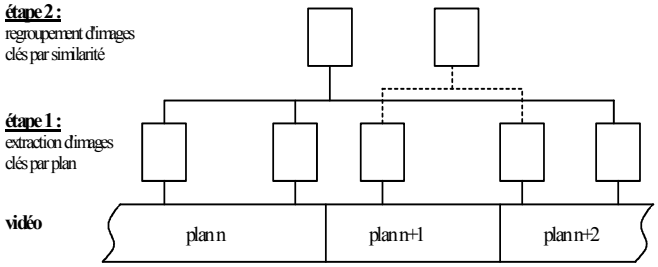

FIG 1 : schéma de principe de la méthode

4. Résultats expérimentaux

Nous avons évalué notre méthode sur un extrait de la séquence « The Avengers ». Cet extrait, composé de 5754 images organisées en 100 plans, possède une grande diversité de contenus. La figure 2 présente un exemple de deux regroupements réalisés grâce à l'information couleur. Les images de droite, issue de l'extraction d'images clés au niveau des plans (étape 1), ont été regroupées lors de la 2e étape. Les images de gauche correspondent aux images les plus proches des centres de gravité des groupes et sont considérées comme images clés. Les images regroupées ont la particularité de décrire la même unité de lieu et ne proviennent pas forcément de mêmes plans.

FIG 2 : exemple de deux regroupements réalisés grâce à l'information couleur (après la 2ème étape de la méthode). Pour chaque regroupement, l'image de gauche représente le centre de gravité du groupe et les images suivantes sont les images clés appartenant à ce groupe et obtenues lors de la 1ère étape. U indique le degré d'appartenance de l'image au groupe et Pl correspond au numéro du plan auquel elle appartient.

Avec l'information mouvement, la méthode extrait des images clés fonction de la dynamique de la scène considérée. Ainsi, les coefficients d'ondelettes que nous avons sélectionnés permettent d'estimer le mouvement dominant dans les images. Le mouvement dominant révèle l'activité globale et donc informe du niveau d'action (séquence lente, film d'action) mais il fournit aussi une direction privilégiée. La figure 3 présente un exemple de regroupement réalisé grâce à l'information mouvement. On observe que le groupe correspond à une translation d'ensemble de la gauche vers la droite et présente une forte activité avec une course poursuite de voitures.

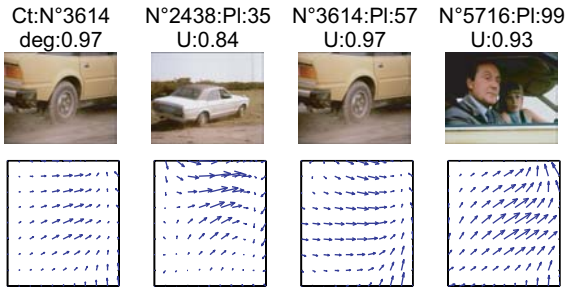

FIG 3 : exemple d'un regroupement réalisé grâce à l'information mouvement (après la 2ème étape de la méthode). L'image de gauche représente le centre de gravité du groupe et les images suivantes sont les images clés appartenant à ce groupe et obtenues lors de la 1ère étape. U indique le degré d'appartenance de l'image au groupe et Pl correspond au numéro du plan auquel elle appartient.

Après la 1ère étape, l'extrait de vidéo (5754 images) est résumé par 191 images clés avec l'index couleur et par 247 images clés par l'index mouvement. La 2e étape qui réunit les images clés du 1er niveau aboutit à 50 images clés avec la couleur et 50 images clés avec le mouvement.

5. Application à la recherche par l'exemple

Une application telle que la recherche par l'exemple permet de juger de l'efficacité de la méthode proposée pour créer des résumés de vidéos. Il s'agit de comparer une image requête aux différents groupes constitués par la méthode. La figure 4 illustre le principe de recherche par l'exemple. Le descripteur couleur ou mouvement est d'abord extrait de la requête (phase 1). Il est ensuite comparé aux centres de gravité des groupes de la 2e étape (phase 2). Les trois premiers groupes, c'est-à-dire les plus proches de la requête, sont sélectionnées puis l'image requête est comparée aux images clés que contiennent ces groupes (phase 3). Le plan qui correspond à l'image clé la plus proche est considéré comme le plus proche de la requête. Un degré de ressemblance (4) est alors associé à chaque image clé appartenant aux 3 groupes sélectionnés :

$$u_k = \exp\left(\frac{-\mathrm{d}(x,v_k)^2}{\underset{i}{median}\,(\mathrm{d}(x,v_i))^2}\right) \qquad (4)$$

où x est le vecteur caractéristique de la réquête, v_k est le vecteur caractéristique de l'image clé k, $d(\cdot,\cdot)$ est la distance euclidienne. Comme les vecteurs caractéristiques des index peuvent avoir des ordres de grandeur très différents, la distance dans l'équation 4 est normalisée par le médian.

Si les critères couleur et mouvement sont utilisés lors de la requête, une intersection au sens des ensembles flous est réalisée entre les plans proches selon la couleur et le mouvement. Une t-norme est utilisée pour généraliser la notion de "et" logique (phase 4). Il s'agit de sélectionner les plans qui apparaissent comme plan proche selon la couleur et le mouvement. La t-norme de Lukasiewicz (5) est alors appliquée au degré de confiance de chacun des plans.

$$\max\left(u_{ci}+u_{mi}-1,0\right) \qquad (5)$$

où u_{ci} est le degré de ressemblance du plan i selon l'index couleur et u_{mi} est le degré de ressemblance du plan i selon l'index mouvement. Ainsi le plan le plus proche de la requête en terme de mouvement et de couleur sera celui qui aura le degré de ressemblance le plus grand.

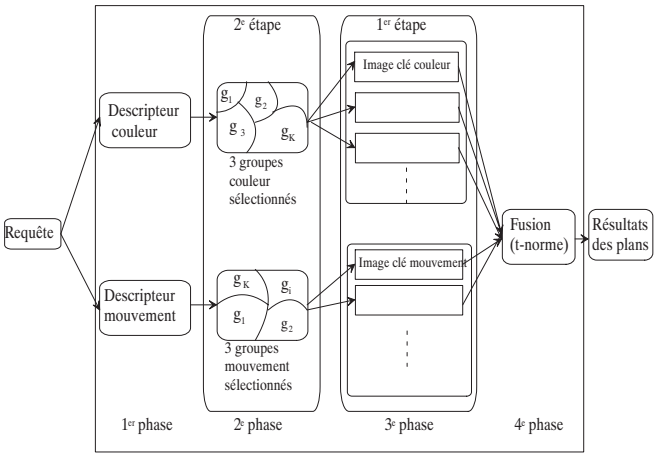

FIG 4 : principe de la recherche par l'exemple.

Afin de l'illustrer et montrer l'intérêt de la méthode de fusion des index, un exemple est présenté sur les figures 6 et 7. Il s'agit d'un cas où le descripteur couleur seul ne permet pas de retrouver le plan auquel appartient la requête. Il en est de même pour le descripteur mouvement. En revanche, la fusion des index permet de retrouver le plan en calculant la t-norme entre les deux ensembles couleur-mouvement.

FIG 5 : exemple d'une requête couleur où le descripteur couleur seul ne suffit pas à retrouver le plan auquel appartient l'image. L'image de gauche correspond à la requête, les images suivantes représentent les images clés les plus proches en terme de couleur. U indique le degré de ressemblance de l'image clé à la requête et Pl correspond au numéro du plan auquel elle appartient.

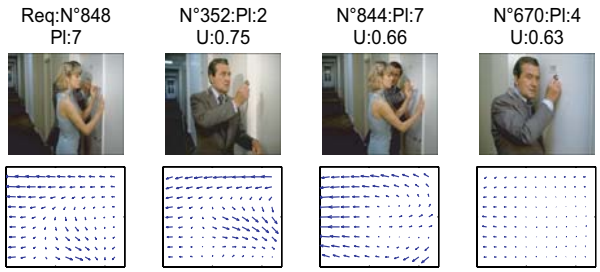

FIG 6 : Exemple d'une requête mouvement où le descripteur mouvement seul ne suffit pas à retrouver le plan auquel appartient l'image. L'image de gauche correspond à la requête, les images suivantes représentent les images clés les plus proches en terme de mouvement. U indique le degré de ressemblance de l'image clé à la requête et Pl correspond au numéro du plan auquel elle appartient.

Cette méthode a été testée sur l'extrait de la vidéo. Nous avons effectué plusieurs fois (5 réalisations) un tirage aléatoire de 20 images requêtes se trouvant dans l'extrait et nous avons vérifié que les plans auxquels elles appartiennent sont bien retrouvés.

Nous avons observé que le résumé couleur offre de bons résultats avec 92% de plans retrouvés. L'histogramme couleur est un descripteur global qui fournit en effet une bonne signature lorsque le nombre d'images n'est pas trop important.

Le résumé mouvement, moins efficace, permet de retrouver 48% des plans. Cet écart dans les résultats provient du fait que le mouvement ne renseigne que sur l'intensité de mouvement et sa direction privilégiée. Par exemple, les scènes sans mouvement ne permettent pas d'être retrouvés efficacement puisque différents plans peuvent contenir une succession d'images sans mouvement et donc posséder la même signature de mouvement.

Enfin, la combinaison couleur et mouvement, facilement réalisée grâce à l'utilisation des ensembles flous, permet d'améliorer les résultats avec 94% de plans retrouvés. Le faible gain apporté ici par la combinaison avec le mouvement provient du fait que la vidéo utilisée est de taille moyenne et qu'ainsi l'information couleur seule fournit déjà de bons résultats. Pour des vidéos plus longues, l'information couleur est moins pertinente et le gain apporté par la combinaison avec le mouvement devrait être plus importante.

6. Conclusion

Nous avons présenté une méthode de construction de résumés de vidéos, selon différents index, avec deux niveaux de résolution. Deux index, couleur et mouvement, ont été décrits. L'index couleur est représenté simplement par son histogramme. Le mouvement, estimé par une méthode reposant sur la théorie des ondelettes, est caractérisé de manière compacte et multi-échelle par un jeu de coefficients. La création de résumés repose sur l'algorithme de regroupement par similarité non supervisée (FCM). Le 1er niveau (intra-plan) extrait des images clés à l'intérieur de chaque plan, puis le 2^{e} niveau (inter-plan) assemble les images clés pour obtenir un résumé vidéo de taille plus réduite. La recherche par l'exemple, testée sur les résumés vidéos, montre que la combinaison des index (couleur, mouvement) améliore les résultats.

A présent, nous allons évaluer et adapter la méthode pour des vidéos de taille plus importante. Nous envisageons pour cela d'introduire des descripteurs plus élaborés, notamment pour la couleur (descripteur local), et pour gagner en efficacité de remplacer la 2^{e} étape par une technique de classification hiérarchique.

Références

[1] B. S. Manjunath, J. Ohm, V. V. Vasudevan, et A. Yamada. *Color and texture descriptors*. IEEE Trans on Circuits and Systems for Video Technology, vol. 11, 6 June 2001.

[2] S. Jeannin et A. Divakaran. *Visual Motion Descriptors*. IEEE Trans on Circuits and Systems for Video Technology, vol 11, 6 June 2001.

[3] G. Sheikholeslami, W. Chang, et A. Zhang. *SemQuery: Semantic Clustering and Querying on Heterogeneous Features for Visual data*. IEEE Trans on Knowledge and Data Engineering, 14(5):988-1002, Sept/Oct 2002.

[4] A. Mufit Ferman, S. Krishnamachari, A. Murat Tekalp, M. Abdel-Mottaleb et R. Mehrotra. *Grou*p-Of-*Frame/Picture Color Histogram Descriptors For Multimedia Applications.* in IEEE International Conference on Image Processing, September. 2000.

[5] S. Krishnamachari et M. Abdel-Mottaleb. *Compact Color Descriptor for Fast Image and Video Segment Retrieval.* in IS&T/SPIE Conference on Storage and Retrieval of Media Databases 2000, Jan 2000.

[6] A.Kokaram et P.Delacourt, *A new global motion estimation algorithm and its application to retrieval in sports events*. in the IEEE workshop on Multimedia Signal Processing, 2001.

[7] K. A. Peker, A. A. Alatan, A. N. Akansu. *Low-Level Motion Activity Features for Semantic Characterization of Video*. IEEE International Conference on Multimedia and Expo (II), 2000.

[8] E. Bruno et D. Pellerin. *Modélisation du mouvement global sur une base d'ondelettes : application à l'indexation de séquences vidéo.* Gretsi, 2001.

[9] R. Ruiloba, P. Joly, S. Marchand-Maillet et G. Quénot. *Towards a Standard Protocol for the Evaluation of video-to-shots Segmentation Algorithms*. CBMI, 1999.

[10] X. L.Xie et G. Beni. *A validity measure for fuzzy clustering*. IEEE Trans. Pattern. Anal. Match. Intell, 13(8) 841-847,1991.