Combinaison de descripteurs flous de couleur et d'activité pour le résumé de vidéos

Combining color and activity fuzzy descriptors for video summaries

M. Guironnet[1] D. Pellerin[1] 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

Dans cet article, nous introduisons deux nouveaux descripteurs de bas niveau (mouvement et couleur). Le descripteur mouvement caractérise l'activité et s'appuie sur un algorithme d'estimation de mouvement par ondelettes. Le descripteur couleur, quant à lui, est représenté par un histogramme flou. A partir de ces index, nous proposons une méthode pour créer des résumés de vidéos possédant deux niveaux de résolution (intra-plan et inter-plan). Notre approche, applicable à tout autre type de descripteurs, consiste à extraire des images clés par regroupement de similarité suivant l'index considéré. Les premier et deuxième niveaux sont respectivement réalisés par l'algorithme de type "fuzzy c-means" et l'algorithme de classification hiérarchique ascendante (CHA). Enfin, nous montrons comment les résumés de vidéos 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.

Mots Clef

Indexation de vidéos, descripteur d'activité, histogramme couleur flou, ensembles flous, algorithme "fuzzy c-means", classification hiérarchique ascendante.

Abstract

In this paper, two new low level descriptors (motion and color) are introduced. The motion descriptor characterizes the dynamic content or activity and is built on a robust optical flow estimation algorithm based on wavelets. The color descriptor is represented by a fuzzy histogram. We propose a method to create video summaries with two resolution levels (intra-shot and inter-shot), from these indexes. Our approach, applicable to all types of descriptors, consists in extracting keyframes by similarity clustering according to the considered index. The first and second levels are respectively carried out by the fuzzy c-means algorithm and the ascending hierarchical classification algorithm. Thus the obtained video summaries allow to achieve a query by example based on a 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

Keywords

video indexing, activity descriptor, fuzzy color histogram, fuzzy sets, fuzzy c-means, ascending hierarchical classification.

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 extrait de vidéos 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. 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 deux nouveaux descripteurs : un histogramme couleur flou représentant l'information couleur contenue dans chaque image et un descripteur d'activité reposant sur une méthode récente d'estimation du mouvement à base d'ondelettes [4].
A partir de ces index, nous proposons une méthode de création de résumés de vidéos à deux niveaux de

résolution. Le premier niveau extrait des images clés grâce à un algorithme de regroupement par similarité non supervisé : "fuzzy c-means" (FCM). Le second niveau consiste à rassembler les images clés par une technique de classification hiérarchique et est similaire aux travaux définis dans [5]. L'approche est indépendante du type de descripteurs. 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.

La section 2 décrit l'état de l'art des descripteurs couleur et mouvement. A partir de l'algorithme d'estimation du mouvement présenté brièvement dans la section 3, un descripteur d'activité est expliqué dans la section 4. L'extraction du descripteur couleur est ensuite exposée dans la section 5. Puis, la section 6 développe la méthode de construction de résumés de vidéos. Dans la section 7, des exemples de résumés de vidéos sont présentés. La section 8 illustre la technique de fusion d'index pour améliorer la recherche par l'exemple. Enfin la dernière section conclut l'article.

2 Etat de l'art des descripteurs couleur et mouvement

Beaucoup de travaux ont été réalisés ces dernières années sur l'indexation, le résumé et la recherche de vidéos. La majorité des approches se compose de deux étapes. La première étape consiste à segmenter la vidéo en plusieurs unités de base appelées plan. Le plan représente une portion de vidéo filmée continûment sans effets spéciaux ni coupure. Plusieurs techniques ont été proposées pour la segmentation en plan telles les méthodes basées sur la différence d'histogrammes [6], sur la différence de mouvement [7] ou sur le suivi des contours [8]. La deuxième étape a pour but de décrire le contenu des plans selon différents niveaux. La plupart des approches utilise des critères de bas niveaux comme la couleur, la texture ou le mouvement. En revanche, l'extraction du contenu sémantique comme les évènements ou les objets [9,10] restent quelque chose de complexe et demande une connaissance a priori sur le type de vidéos.

Parmi les index de bas niveaux, un des plus utilisés est la couleur en raison de sa simplicité et de ses performances. L'index couleur fournit en effet une bonne perception de similarité entre différentes images. Le descripteur couleur est le plus souvent réalisé à partir de l'histogramme couleur [11,12,13]. Les variantes proviennent du choix de l'espace couleur et du nombre de bins par composantes. Néanmoins ces approches ont l'inconvénient de fournir le même poids aux pixels qu'ils soient au centre ou proches du bord du même bin. En effet, la comparaison entre deux histogrammes couleur extraits d'images similaires peut engendrer une distance élevée en raison d'un changement d'éclairage, d'un glissement de bins lors de la quantification ou lorsque les images sont bruitées.

D'autres types de descripteurs couleur [1] issus de la norme MPEG-7 se développent comme l'extraction des couleurs dominantes, la distribution spatiale des couleurs et la répartition locale de la couleur. Mais la plupart des travaux actuels n'exploitent pas les évènements spatio-temporels pour représenter une vidéo. Pour répondre à ce genre de problème, Ferman et al. [13] définissent une méthode simple pour décrire un groupe d'images. Ils définissent un histogramme moyen lissé appelé "alpha-trimmed average histogram". L'avantage est une faible contribution des données aberrantes.

La caractérisation du mouvement à l'intérieur d'un plan est nécessaire pour décrire le contenu dynamique des vidéos. Jeannin et al. [2] montrent que le mouvement est une information pertinente pour l'indexation de vidéos. Celui-ci peut provenir du mouvement de la caméra, du mouvement des objets ou d'une combinaison des deux. Des méthodes simples décrivent l'activité par une différence d'images [14] ou par l'algorithme du block-matching [15,16]. Le descripteur d'activité capture des notions intuitives d'intensité de mouvement (séquence lente, film d'action…). En général, le contenu vidéo s'étend sur toutes les gammes, de la forte à la faible activité. D'autres techniques [17,18,19] reposant sur un modèle paramétrique considèrent le mouvement dominant entre deux images successives. Elles permettent de caractériser les types de mouvement de caméra. Un descripteur sur le mouvement des objets est aussi défini à partir d'un modèle paramétrique dans [20]. Il montre son intérêt pour caractériser le mouvement arbitraire des objets dans une vidéo. En revanche, Fablet et al [21] présentent une méthode non paramétrique du mouvement qui repose sur une modélisation statistique de distributions de mesures locales de mouvements.

3 Algorithme d'estimation de mouvement

Cette section va décrire une méthode d'estimation du mouvement entre deux images successives qui conduira ensuite à un descripteur d'activité (section 4).

La détermination du flot optique est une étape importante et conditionnera les performances du descripteur mouvement. La méthode d'estimation que nous avons choisie [4] caractérise le mouvement par un jeu de coefficients. Elle permet une représentation multi-échelle et compacte du mouvement.

Soit $V_j(p_i, t)$ l'estimation du flot optique au pixel $p_i=(x_i, y_i)$ entre les images $I(p_i,t)$ and $I(p_i,t+1)$ au niveau de résolution j. Le mouvement est approché par une combinaison linéaire de fonctions d'échelle.

$$V_j(p_i,t) = \sum_{k_1,k_2=0}^{2^{j-1}} \theta_{j,k_1,k_2}\, \Phi_{j,k_1,k_2}(p_i) \qquad (1)$$

Φ représente la fonction d'échelle, dans notre cas, une fonction B-Spline de degré 1 avec 3 niveaux de résolution

(figure 1). Les indices j, k_1, k_2 représentent respectivement le niveau d'échelle, le décalage horizontal et vertical. Les coefficients d'échelle, une fois estimés, modélisent le flot optique entre deux images successives. De plus, connaissant la décomposition du champ des vitesses à un niveau de résolution j, on peut déterminer tous les coefficients d'échelle à un niveau de résolution inférieur. En supposant la conservation de la luminance, l'algorithme estime de façon itérative et robuste les coefficients θ en minimisant une fonction objective:

$$E=\sum_{p_i}\rho(I(p_i+V_j(p_i,t+1)-I(p_i,t),\sigma)) \qquad (2)$$

La fontion $\rho(\cdot,\sigma)$ est une fonction qui pondère les données en fonction de l'erreur (M-estimateur de Geman-McClure). La minimisation de la fonction objective (équation 2) est décrite dans [22].

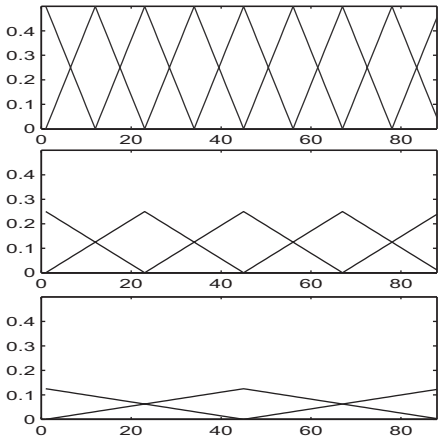


Figure 1 : Fonction B-Spline de degré 1 suivant trois niveaux de résolution: en haut, j=2 , au milieu j=1 et en bas j=0

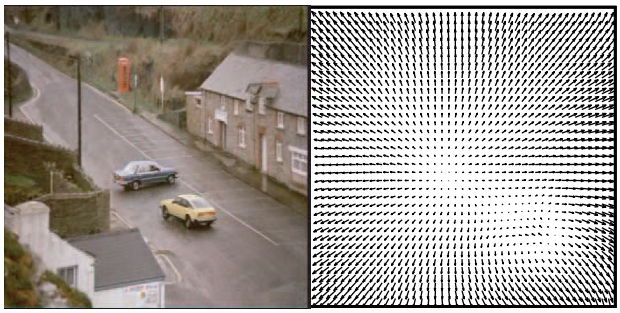
La connaissance fine du mouvement n'étant pas nécessaire pour l'indexation de vidéos, son estimation est réalisée sur des images sous-échantillonnées (imagette 72x88 pixels) pour accélérer les traitements. La méthode offre une signature du mouvement sous la forme de 162 coefficients (81 coefficients pour chaque composante de vitesse). La figure 2 montre un exemple d'estimation du mouvement. Les deux grilles 9*9 de coefficients d'échelle permettent de reconstituer le flot optique. Il correspond à la superposition du mouvement de la voiture et du zoom de la caméra.

-3,2	-2,3	-1,5	-0,6	0,2	1,1	2,0	1,7	1,4
-3,1	-2,1	-1,0	-0,4	0,3	1,0	1,8	2,2	2,7
-3,1	-1,8	-0,6	-0,1	0,3	0,9	1,5	2,7	3,9
-2,6	-1,8	-1,1	-0,3	0,6	1,2	1,7	2,8	3,9
-2,1	-1,8	-1,6	-0,4	0,8	1,4	1,9	2,9	3,8
-2,3	-1,8	-1,3	-0,5	0,3	0,4	0,4	2,3	4,2
-2,6	-1,9	-1,1	-0,7	-0,3	-0,7	-1,0	1,8	4,6
-2,9	-2,0	-1,2	-0,4	0,4	0,9	1,4	1,9	2,4
-3,2	-2,2	-1,2	0,0	1,2	2,5	3,8	2,0	0,2

(a)

-2,9	-2,7	-2,4	-2,5	-2,7	-2,5	-2,3	-2,5	-2,6
-2,2	-2,0	-1,8	-2,0	-2,1	-2,0	-1,9	-1,9	-2,0
-1,5	-1,4	-1,3	-1,4	-1,5	-1,5	-1,4	-1,4	-1,4
-0,7	-0,7	-0,8	-0,8	-0,9	-0,8	-0,8	-0,8	-0,8
0,1	-0,1	-0,3	-0,3	-0,2	-0,1	-0,1	-0,1	-0,1
0,7	0,5	0,4	0,5	0,5	0,1	-0,3	0,4	1,1
1,2	1,2	1,2	1,2	1,2	0,3	-0,6	0,9	2,4
1,8	1,8	1,8	1,8	1,8	1,6	1,4	2,1	2,8
2,3	2,4	2,4	2,4	2,3	2,9	3,4	3,3	3,1

(b)



(c) (d)

Figure 2 : Exemple d'estimation du mouvement. (a) Grille des coefficients d'échelle θ_x et (b) grille des coefficients d'échelle θ_y au niveau de résolution j=2. (c) Image de la séquence "The avengers" et (d) flot optique estimé

4 Construction d'un descripteur d'activité

L'extraction de l'activité dans les plans fournit une information sur le contenu spatio-temporel des vidéos. Chaque plan peut être représenté par un degré d'activité. En effet, les films d'action possèdent en général beaucoup de plans avec une forte activité alors que les journaux télévisés sont plutôt caractérisés par une faible activité. La distribution spatiale de l'activité apporte aussi une information appropriée pour la classification de vidéos. Il est intéressant de savoir si le mouvement est présent sur toute l'image ou au contraire localisé sur une partie de celle-ci. Comme les coefficients d'échelle caractérisent le mouvement sur chaque région de l'image, ils offrent directement une représentation compacte de la distribution du mouvement.

Nous proposons un descripteur d'activité qui découle de la notion intuitive d'intensité de mouvement. Cette information peut être directement symbolisée par l'amplitude des coefficients d'échelle. En effet, le degré d'activité dépend de la norme des vecteurs vitesses. Dans notre approche, l'amplitude est donnée par les coefficients d'échelle suivant l'équation (3).

$$\text{amplitude}=\sqrt{\theta_x^{\,2}+\theta_y^{\,2}} \qquad (3)$$

A partir des grilles θ_x et θ_y de chaque paire d'images, nous obtenons une grille d'amplitude. Connaissant la décomposition du champ des vitesses au niveau de résolution j=2, on détermine tous les coefficients d'échelle au niveau de résolution j=1 (figure 1). Dans notre implémentation, la grille d'amplitude est une grille 5 par 5. Chaque amplitude est ensuite labellisée suivant 3 ensembles flous : forte activité, activité moyenne et faible activité. Les fonctions d'appartenance sont représentées sur la figure (3). Nous obtenons alors 3 grilles à partir d'une grille d'amplitude puisque chaque coefficient est transformé en un degré d'appartenance suivant les 3 ensembles flous considérés. Ces 3 grilles décrivent l'activité locale de mouvement.

Il est possible, par exemple pour visualiser l'activité spatiale, de revenir à une seule grille en utilisant la méthode du centre de gravité. Elle consiste à calculer la moyenne des valeurs modales des ensembles, pondérés par leur degré d'appartenance. Une valeur modale d'un ensemble flou est par définition une valeur x telle que $\mu(x)=1$. Dans notre implémentation, les valeurs modales des 3 ensembles flous (faible, moyenne et forte activité) sont respectivement 0, 2 et 4. Par exemple, si l'amplitude d'un coefficient à des degrés d'appartenance de 0.2, 0.8 et 0 respectivement pour les 3 ensembles flous (faible, moyenne et forte activité) alors l'amplitude résultante vaudra 1.6 (0.2*0+0.8*2+0*4). A partir de cette description, nous pouvons aussi caractériser une activité globale entre deux images successives en calculant la cardinalité de chacun des 3 ensembles flous (figure 4). L'ensemble possédant la plus grande cardinalité renseignera sur le niveau d'activité.

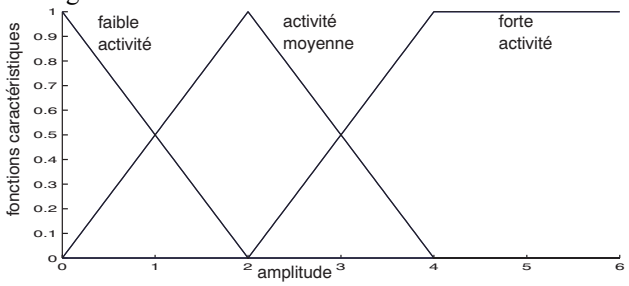


Figure 3 : Fonctions caractéristiques de descripteur de l'activité en fonction de l'amplitude du mouvement

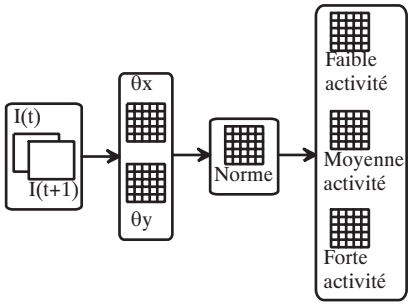


Figure 4 : Principe de construction du descripteur activité

5 Construction d'un descripteur couleur

L'indexation image par l'histogramme couleur est efficace pour rechercher des images ou des vidéos dans une base. Néanmoins, les différents espaces couleur (RGB, HSV, YCbCr, …) ne présentent pas tous les mêmes propriétés. En effet, l'espace RGB n'est perceptuellement pas uniforme et dépend des conditions d'éclairage. L'espace HSV lui est souvent préféré parce qu'il traduit directement les notions intuitives de couleur. La transformation de l'espace RGB à HSV est non linéaire mais inversible. La *teinte H* qui varie de 0 à 2π représente la nuance de couleur et où celle-ci se situe dans le spectre des couleurs. La *saturation S* décrit la pureté de la couleur et varie de 0 à 1. Enfin *la luminance V* correspond à la luminosité de la couleur et elle est représentée par des valeurs de 0 à 255. Le cylindre HSV est souvent quantifié uniformément suivant chacune des composantes et permet la réduction du nombre de couleurs dans l'image. Chaque composante se divise en intervalles réguliers. Comme souligné précédemment, l'inconvénient de cette méthode est qu'elle attribue le même poids aux pixels à proximité du centre d'un bin de ceux se situant aux bords du même bin. L'utilisation des ensembles flous permet de résoudre ce genre de problème en associant aux pixels un degré d'appartenance suivant chaque bin. De plus, Sural et al. [23] ont observé que la saturation détermine la transition entre les couleurs et les niveaux de gris. Lorsque S vaut 0 et V augmente, la couleur va du noir au blanc en passant par tous les niveaux de gris. En revanche si la saturation augmente pour une luminance V et une teinte H données, la couleur perçue change du niveau de gris à la couleur pure indiqué par H. Finalement, pour de faibles valeurs de S, la couleur peut être représentée par un niveau de gris (V) alors que pour des valeurs importantes de S, la couleur peut être apparentée à la teinte (H).

La construction de l'histogramme couleur flou consiste d'abord à déterminer le seuil afin de savoir si un pixel va être représenté par sa teinte (H) ou sa luminance (V). La fonction qui détermine ce seuil dépend de la luminance V

$$\text{seuil}(V) = 0.8 * \left(1 - \frac{V}{255}\right)^2 + 0.2 \qquad (4)$$

Si V est égal à 0, le seuil vaut 1, ce qui signifie que le pixel est bien associé à du noir. En revanche si V vaut 255 et S est supérieur à 0.2 alors le pixel sera considéré comme de la couleur. Ainsi le seuil est une valeur déterminante puisqu'il permet de trancher sur le représentant du pixel (H ou V). Comme le seuil est calculé approximativement par l'équation 4, il est intéressant d'associer un degré d'appartenance à chaque pixel suivant la valeur du seuil. La figure 5 montre comment s'effectue la pondération entre la composante V et la composante H suivant le seuil calculé. Ainsi chaque pixel proche du seuil sera représenté à la fois par sa teinte et sa luminance.

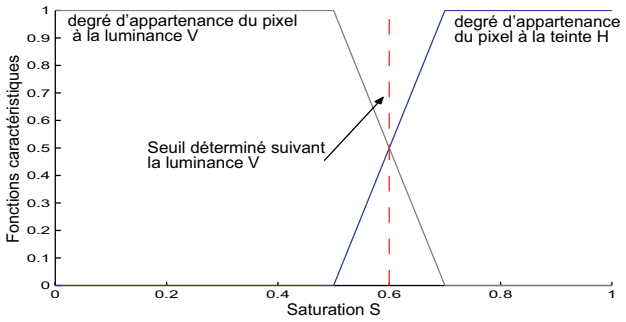


Figure 5 : Fonctions d'appartenance des pixels à chacune des composantes (H et V) suivant la valeur du seuil obtenu par la saturation S.

De même, si un pixel appartient à la composante teinte (couleur), un degré d'appartenance lui est aussi associé. Dans notre implémentation, nous avons considéré 6 bins couleurs : rouge, jaune, vert, cyan, bleu et magenta. La figure 6 montre les fonctions caractéristiques représentant les différents bins.

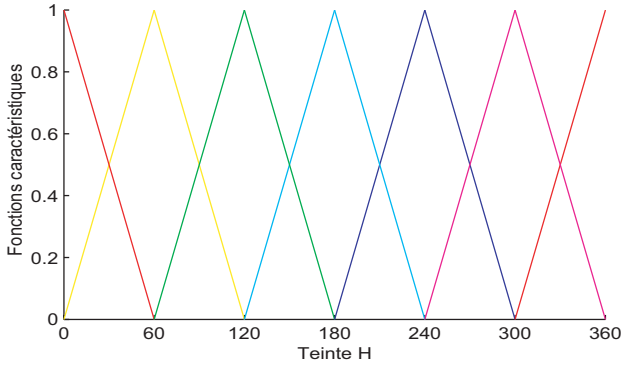


Figure 6 :Fonctions caractéristiques des 6 bins couleurs en fonction de la teinte H.

Les pixels appartenant à la composante luminance sont eux aussi associés à 5 bins dont les fonctions caractéristiques sont montrées sur la figure 7. Nous avons choisi 5 bins symbolisant les différents niveaux de gris : noir, gris-foncé, gris, gris-clair et blanc.

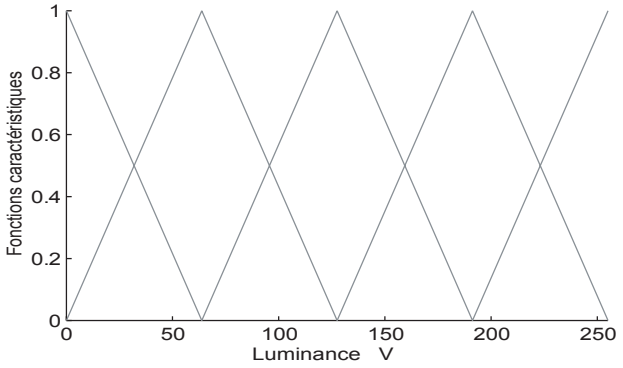


Figure 7 : Fonctions caractéristiques des 5 bins de niveaux de gris en fonction de la luminance V.

Les méthodes basées sur l'histogramme couleur ont l'avantage d'être invariantes aux orientations, à la taille de l'image et robustes aux bruits de l'arrière plan. Cependant, le problème de ce type d'approche est qu'elle ne fournit aucune information sur la répartition spatiale de la couleur. Ainsi deux images possédant des couleurs similaires peuvent être sémantiquement très différentes. Pour pallier ce problème, nous calculons un *histogramme couleur local*. L'image est simplement divisée en 9 blocs (grille 3*3). Suivant les degrés d'appartenance des pixels, l'histogramme couleur est défini sur chacun des blocs par l'union des deux histogrammes selon les composantes luminance et teinte. Ce procédé a l'avantage de conserver l'information spatiale et peut permettre de retrouver l'histogramme global en effectuant une moyenne des histogrammes obtenus sur ces blocs.

6 Méthode de création 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 8). Comme plusieurs techniques performantes de découpages d'une vidéo en plans existent, 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 pour le 1er niveau 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 de données par nature imprécises.

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 :

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

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, $d(\cdot,\cdot)$ est la distance euclidienne et m est un paramètre supérieur à 1 (m=1.5). 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 [24] est utilisé. Il fournit une mesure de validité sur la compacité et la séparabilité des groupes. La méthode consiste à déterminer le nombre de

groupes K optimal en évaluant un critère ν suivant la partition obtenue avec l'algorithme FCM.

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

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

Dmin est la distance minimum entre les centres de gravité des groupes. 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 8).
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 aux images clés extraites l'algorithme de classification hiérarchique ascendante (étape 2).
La méthode de regroupement hiérarchique ascendant considère les éléments (vecteurs caractéristiques représentant les images clés du 1er niveau) comme une partition initiale. A chaque itération, une nouvelle classe est formée par agglomération des 2 classes les plus proches au sens de la distance choisie entre les groupes. La distance utilisée dans notre implémentation est la distance au voisin le plus éloigné car elle offre une mesure efficace pour former des groupes compacts. Soient A et B deux groupes, d_{ij} la distance entre deux individus (vecteurs caractéristiques), la distance entre A et B est :

$$d(A,B) = \max_{i \in A, j \in B} (d_{ij})$$

La fusion des groupes est arrêtée lorsque la distance entre les groupes est supérieure à un seuil préfixé (1 pour la couleur et 0.01 pour le mouvement). Si l'utilisateur souhaite au contraire imposer le nombre d'images constituant le résumé vidéo, le critère d'arrêt est alors le nombre de groupes souhaités.

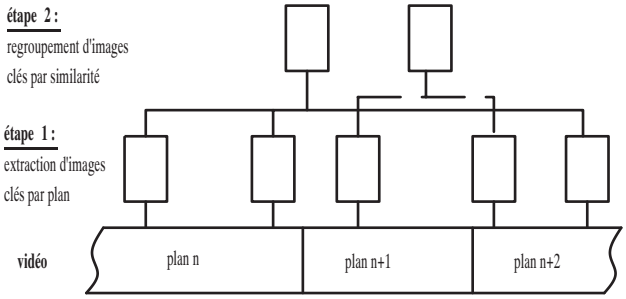


Figure 8 : Principe de construction de résumés de vidéos

7 Exemples de résumés de vidéos

Nous avons évalué notre méthode sur un extrait de la séquence "The Avengers". Cet extrait, composé de 38973 images organisées en 370 plans, possède une grande diversité de contenus. La figure 9 présente un exemple de regroupements réalisés grâce à l'histogramme couleur flou local. Les images les plus proches des centres de gravité sont considérées comme images clés. Les images de droite, issues 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 clés de la 2e étape. Les images regroupées ont la particularité de décrire la même unité de lieu et ne proviennent pas forcément de mêmes plans.

Figure 9 : Exemple de trois regroupements réalisés grâce à l'information couleur (après la 2e 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. Pl correspond au numéro du plan auquel l'image clé appartient.

Avec l'information mouvement, la méthode extrait des images clés fonction de la dynamique de la scène considérée. La figure 10 présente un exemple de regroupement réalisé grâce à l'information mouvement. Les images possédant les descripteurs d'activité les plus proches des centres de gravité sont considérées comme images clés. Cependant l'activité locale affichée à chaque image correspond au centre de gravité du groupe. L'activité locale est représentée par une grille formée de différents niveaux de gris. Le blanc correspond à une

forte activité et le noir à une faible activité. Le 1er regroupement a la particularité de posséder une activité localisée sur une partie de l'image autour des personnages. Par contre, le 3e regroupement présente une forte activité avec une course de voitures. Néanmoins, l'endroit où l'activité est la plus faible se situe sur le véhicule. C'est le paysage qui est en mouvement puisque la caméra traque la voiture.

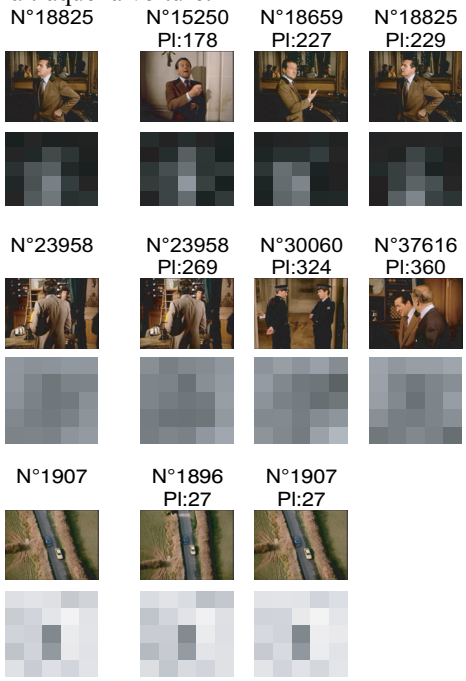


Figure 10 : Exemple de trois regroupements réalisés grâce à l'activité (après la 2e étape de la méthode). Pour chaque regroupement, l'image de gauche représente le centre de gravité du groupe et les images suivantes sont les images clés appartenant à ce groupe et obtenues lors de la 1ère étape. Pl correspond au numéro du plan auquel l'image clé appartient.

Bien que l'activité va améliorer les résultats obtenus par le descripteur couleur (section 8), il ne peut pas être utilisé pour regrouper des plans avec la même unité de lieu. En effet, cet index décrit le comportement dynamique au sein d'un plan. Par définition, l'activité peut être similaire dans des plans de contenu sémantique complètement différent (figure 11). Il suffit de considérer par exemple 2 plans de contenu différent ne comportant pas de mouvement.

Figure 11 : Exemple d'un regroupement réalisé grâce à l'activité (après la 2e 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. Pl correspond au numéro du plan auquel l'image clé appartient.

Après la 1ère étape, l'extrait de vidéo (38973 images) est résumé par 741 images clés avec l'index couleur et par 519 images clés par l'index mouvement. La 2e étape qui réunit les images clés du 1er niveau aboutit à 320 images clés avec la couleur et 264 images clés avec le mouvement.

8 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. L'objectif va être de retrouver le plan auquel appartient l'image requête. Il s'agit de comparer une image requête aux différents groupes constitués par la méthode. Comme évoqué dans la section précédente, le descripteur d'activité renseigne sur le mouvement et il ne permet pas le regroupement de plans possédant la même unité de lieu. La recherche s'effectue alors en deux étapes. Le descripteur couleur de l'image requête est d'abord considéré seul et permet de sélectionner un groupe d'images clés proche de la requête. Ce groupe d'images est ensuite comparé à la requête en considérant le descripteur d'activité.

Le descripteur couleur extrait de la requête, est comparé aux centres de gravité des groupes de la 2e étape. Les quatre premiers groupes, c'est-à-dire les plus proches de la requête, sont sélectionnés puis l'image requête est comparée aux images clés que contiennent ces groupes. Le plan qui correspond à l'image clé la plus proche est considéré comme le plus proche de la requête. Un degré de ressemblance (équation 5) est alors associé à chaque image clé appartenant aux 4 groupes sélectionnés :

$$u_k = \exp\left(\frac{-\,\mathrm{d}(x,v_k)^2}{\underset{i}{median}\,(\,\mathrm{d}(x,v_i)\,)^2}\right) \qquad (5)$$

où x est le vecteur caractéristique de la requête, 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 5 est normalisée par le médian.

Afin de séparer les images effectivement proches de la requête de celle qui ne le sont pas, les images sélectionnées en terme de couleur sont ensuite comparées avec le descripteur d'activité. Un degré de ressemblance est aussi calculé pour cet index (équation 5).

Une t-norme est utilisée pour généraliser la notion de "et" logique. Il s'agit de sélectionner les plans qui apparaissent comme plan proche selon la couleur et le mouvement. La t-norme de Lukasiewicz (équation 6) est alors appliquée au degré de ressemblance de chacun des plans.

$$\max\left(u_{ci}+u_{mi}-1,0\right) \quad (6)$$

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. Le plan qui est le plus proche de la requête en terme de mouvement et de couleur est celui qui a le degré de ressemblance le plus grand.

Afin d'illustrer la méthode de fusion des index, un exemple est présenté sur la figure 12. Il s'agit d'un cas où le descripteur couleur ne permet pas de retrouver le plan auquel appartient la requête. Les images de droite représentent les images similaires à la requête selon l'index couleur.

Figure 12 : Exemple d'une requête couleur où le descripteur couleur 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 l'image clé appartient.

A partir de cet ensemble d'images, le descripteur d'activité est comparé à la requête. La figure 13 montre que le plan auquel appartient la requête est retrouvé.

Figure 13 : Exemple d'une requête où le descripteur d'activité est employé sur les images sélectionnées par la couleur. L'activité permet de retrouver le plan auquel appartient la requête. U indique le degré de ressemblance de l'image clé à la requête et Pl correspond au numéro du plan auquel l'image appartient.

Enfin l'application de la t-norme permet d'éliminer les images éloignées en terme de couleur et de mouvement. Le résultat est présenté sur la figure 14.

Figure 14 : Exemple d'une requête où le descripteur couleur et mouvement sont utilisés. U indique le degré de ressemblance de l'image clé à la requête et Pl correspond au numéro du plan auquel l'image clé appartient.

Conclusion

Nous avons présenté dans cet article deux nouveaux descripteurs de bas niveau : un histogramme couleur flou local et un descripteur d'activité local. Le descripteur d'activité s'appuie sur un algorithme d'estimation de mouvement par ondelettes alors que l'histogramme couleur caractérise directement les notions intuitives de couleur. A partir de ces deux index, nous avons défini une méthode de construction de résumés de vidéos possédant deux niveaux de résolution. La création de résumés repose sur l'algorithme de regroupement par similarité non supervisée. Le premier niveau (intra-plan) extrait des images clés à l'intérieur de chaque plan (algorithme FCM), puis le deuxième niveau (inter-plan) assemble les images clés pour obtenir un résumé vidéo de taille plus réduite (algorithme CHA). La recherche par l'exemple, testée sur les résumés vidéos, montre que la combinaison des index (couleur, mouvement) améliore les résultats.

L'évaluation de la méthode doit maintenant être approfondies sur une base contenant plusieurs heures de vidéos. Nous envisageons aussi l'introduction d'autres descripteurs comme par exemple la direction du mouvement.

Bibliographie

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

[5] E Veneau, R Ronfard et P Bouthemy, "From video shot clustering to sequence segmentation", *Fifteenth International Conference on Pattern Recognition ICPR'2000*, Barcelone, September 2000.

[6] D. Zhang, W. Qi, H. Zhang: "A New Shot Boundary Detection Algorithm", *IEEE Pacific Rim Conference on Multimedia 2001*: 63-70, 2001.

[7] E. Bruno and D. Pellerin, "Video shot detection based on temporal linear prediction of motion", *In Proc of the IEEE International Conference on Multimedia and Exposition (ICME),* Lausanne, Switzerland, August 2002.

[8] W.J. Heng et K.N. Ngan, "An object-based shot boundary detection using edge tracing and tracking", *Journal of Visual Communications and Image Representation*, Academic Press, U.S.A., Vol. 12, No. 3, pp. 217-239, September 2001.

[9] G. Xu, Y.-F. Ma, H.-J. Zhang et S. Yang, "Motion based event recognition using HHM", *In Proc. ICPR*, Quebec, 2002.

[10] J. Fan, X. Zhu, L. Wu, "Automatic model-based semantic object extraction algorithm", *IEEE Trans. on Circuits and Systems for Video Technology*, 11(10) 1073-1084, October 2001.

[11] 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.

[12] J. R. Ohm, et al., "A Multi-Feature Description Scheme for Image and Video Database Retrieval", *Multimedia Signal Processing*, IEEE 3 rd Workshop, pp.123-128, 1999.

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

[14] A. Albiol, L. Torres, and E. J. Delp, "Video preprocessing for audiovisual indexing", *In* Proc of the International Conference on Acoustics, Speech and Signal Processing*,* vol 4, Orlando, pp 3636-3639, 2002.

[15] L. Yuan, W. Gao, W. Zeng, "A Novel Approach to Video Indexing and Retrieval Using Motion Activity", *ACCV2002: The 5th Asian Conference on Computer Vision*, Melbourne, Australia, pp.50-55 January 2002.

[16] 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)*, *ICME'00*, 2000.

[17] J.-M. Odobez, P. Bouthemy, "Robust multiresolution estimation of parametric motion models", *Journal of Visual Communication and Image Representation*, 6(4):348-365, December 1995.

[18] A. Smolic, T. Sikora and J.-R. Ohm, "Direct Estimation of Long-Term Global Motion Parameters Using Affine and Higher Order Polynomial Models", *Proc. PCS'99, Picture Coding Symposium*, Portland, OR, USA, April 1999.

[19] F Comby, O Strauss, MJ Aldon, "Possibilty theory and rough histograms for motion estimation in a video sequence", *Int. Workshop on Visual Form*, Capri, May 2001.

[20] T.Zaharia et F.Preteux, "Parametric motion models for video content description within the MPEG-7 framework", *SPIE Conf. Nonlinear Image Processing and Pattern Analysis*, San Jose, January 2001.

[21] R. Fablet, P. Bouthemy et P. Perez, "Non-parametric motion characterization using causal probabilistic models for video indexing and retrieval", *IEEE Trans. on Image Processing*, April 2002.

[22] E. Bruno et D. Pellerin, "Global motion model based on B-spline wavelets: application to motion estimation and video indexing", *In Proc. of the 2nd Int. Symposium. on Image and Signal Processing and* Analysis ISPA'01, Pula, Croatia, June 2001.

[23] S. Sural, G. Qian and S.Pramanik, "A histogram with perceptually smooth color transition for image retrieval", *Fourth International Conference on Computer Vision, Pattern Recognition and Image Processing CNP 2002*, Durham, North Carolina, pp. 664-667, March 2002.

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