Extraction d'attributs de mouvement et classification hiérarchique pour l'indexation vidéo par le contenu

Ronan Fablet[1] et Patrick Bouthemy[2]

[1]IRISA / CNRS
[2]IRISA / INRIA
Campus universitaire de Beaulieu,
35042 Rennes Cedex, France
e-mail : rfablet@irisa.fr, bouthemy@irisa.fr
Tel : (33) 2.99.84.25.23 Fax : (33) 2.99.84.71.71

Journées ORASIS'99, Aussois, France

Résumé

Nous proposons dans cet article une approche originale pour la caractérisation du mouvement apparent dans le contexte de l'indexation vidéo par le contenu. Nous avons formalisé et exploitons une analyse statistique des distributions de cooccurrences temporelles de mesures locales de mouvement pour extraire des attributs de mouvement globaux. Ceci nous permet d'appréhender une large gamme de situations de mouvement. Une procédure de classification hiérarchique ascendante exploite ces descripteurs afin de déterminer une représentation arborescente pertinente de la base de séquences vidéo traitée. Cette hiérarchie permet ensuite de réaliser aisément la recherche d'exemples similaires à une requête en terme de propriétés de mouvement. Nous présentons des résultats de classification et de recherche obtenus sur un ensemble de séquences d'images réelles. [1]

1 Introduction

La création et l'exploitation sans cesse croissantes d'images dans de nombreuses applications (météorologie à partir d'images satellites, surveillance du trafic routier, imagerie médicale, archives audiovisuelles comprenant des documentaires, des journaux télévisés, des œuvres cinématographiques, l'édition multimédia, ...), rendent nécessaire le développement de nouvelles méthodes de structuration et de consultation de ces bases. La réalisation d'une indexation adéquate en relation avec le contenu de ces documents audiovisuels en est un des aspects cruciaux. Nous sommes dans notre cas concernés par les documents vidéo. Bien que de nombreux travaux soient actuellement dédiés à l'indexation vidéo, [1, 9, 12, 23], il reste difficile d'identifier les informations pertinentes au regard d'une requête donnée, en raison de la complexité de l'interprétation des images et de l'analyse des scènes dynamiques.

Dans le cas particulier de l'indexation vidéo, il est nécessaire de réaliser en premier lieu la segmentation temporelle de la vidéo en plans élémentaires, [3, 6, 22], ce qui conduit également à identifier des situations classiques de prise de vue (plan fixe, traveling, zoom, ou plan panoramique), [3, 18, 19, 21, 22]. Dans un second temps, le contenu de chaque plan peut être caractérisé. Cet aspect a motivé des travaux récents concernant la construction de mosaïques

1. Ces travaux ont été partiellement financés par l'Association Franco-Israélienne pour la Recherche Scientifique (AFIRST)

d'images, [13], l'interprétation du mouvement des objets dans le cas d'une caméra fixe, [4], ou encore la détection et de suivi des objets mobiles, [7, 8]. La plupart de ces méthodes exploitent une segmentation au sens du mouvement qui repose soit sur des modèles de mouvement paramétriques 2D, soit sur l'estimation de champ dense de vitesse. Néanmoins, elles peuvent se révéler inadaptées pour certaines classes de séquences, en particulier dans le cas de mouvement non-structurés de rivières, de feuillage, de foules, ... , (*cf* Fig. 1). De plus, dans le contexte de l'indexation vidéo, il semble aussi pertinent de fournir une interprétation globale du mouvement et de discriminer des types de mouvement généraux, sans segmentation préalable au sens du mouvement ou sans estimation complète du mouvement. Ces remarques soulignent le besoin de développer de nouvelles approches afin de proposer une description directe et globale du mouvement, [2, 15, 17, 20].

Nous proposons dans cet article une approche originale pour l'indexation et la consultation de bases de vidéo en relation avec l'information de mouvement. Nous utilisons des descripteurs globaux du mouvement décrits dans nos précédents travaux, [2], qui reposent sur une analyse statistique des cooccurrences temporelles de mesures de mouvement, locales et non-paramétriques. Ces index de mouvement sont introduits dans un schéma de classification hiérarchique ascendant pour déterminer une représentation pertinente d'un ensemble de séquences d'images. Cette structure hiérarchique repose sur une notion de similarité dans l'espace des attributs de mouvement. Par conséquent, elle peut être aisément exploitée pour réaliser une recherche efficace en relation avec une requête exprimée.

Cet article est organisé comme suit. La partie 2 introduit les idées directrices de nos travaux. Dans la partie 3, nous décrivons brièvement la méthode d'extraction des attributs de mouvement globaux. La partie 4 présente la structure d'indexation arborescente construite ainsi que la procédure de recherche dans la base de données d'exemples similaires à une requête. Dans la partie 5, nous présentons des résultats de classification et de recherche. Enfin, nous concluons en partie 6.

2 Problématique

Nous nous intéressons ici à la caractérisation de séquences d'images par l'information de mouvement. Cette problématique n'a de sens que si nous disposons d'une segmentation temporelle préalable de la séquence traitée. Par conséquent, nous extrayons dans un premier temps les plans élémentaires des séquences vidéos en exploitant la technique présentée en [3]. Elle consiste à étudier les variations temporelles de la taille normalisée du support associé au mouvement dominant dans l'image. Dans la suite, nous utiliserons indifféremment les termes de plan ou de séquence. Pour chaque plan extrait, nous cherchons à caractériser de façon globale la distribution spatio-temporelle du mouvement en vue de définir un système d'indexation vidéo par le contenu.

Il est à noter que l'analyse de texture suit la même approche pour étudier la distribution des niveaux de gris. En particulier, les mesures de cooccurrences fournissent des outils efficaces pour la description des textures en termes d'homogénéité, de contraste ou de granularité, [10]. Nous avons cherché à adapter l'extraction d'attributs de cooccurrence à l'analyse du mouvement apparent. Des travaux préliminaires dans cette direction, [15], ont conduit à définir la notion de texture temporelle, opposée aux mouvements périodiques ou rigides, et associée aux mouvements fluides. Ainsi, les mouvements de rivière, de flammes, de foules ou de feuillage peuvent être considérés comme des exemples de texture temporelle (*cf* Fig. 1).

Dans [20], des exemples de synthèse de textures temporelles proches des séquences originales sont présentés. Cependant, cette approche est limitée à ce type particulier de séquences

a)  b) 

FIG. 1 – *Exemples de textures temporelles : a) feuillage b) feu (image provenant du* MIT*).*

d'images et ne peut être étendue aux mouvements rigides ou périodiques. Dans [17], des attributs de texture temporelle sont extraits à partir de la description des surfaces générées par les trajectoires spatio-temporelles des contours des objets mobiles dans la scène. Toutefois, cette méthode repose sur une détection préalable des contours mobiles par un simple seuillage de la dérivé temporelle de l'intensité, qui est connue pour être très bruitée.

L'utilisation des cooccurrences nécessite de définir des cartes de mesures locales pour la séquence traitée. L'estimation de champ dense peut fournir de telles grandeurs locales. Néanmoins, elle se révèle coûteuse en temps de calcul et peu fiable dans les cas de contenus dynamiques complexes. Par conséquent, nous préférons considérer une mesure de mouvement locale directement calculée à partir des gradients spatio-temporels de l'intensité. Plutôt que d'utiliser la vitesse normale comme dans [15], nous exploitons une mesure plus fiable décrite dans la partie suivante. Les attributs globaux de mouvement extraits de statistiques de cooccurrences temporelles sont alors utilisés pour définir une méthode d'indexation et de recherche dans une base de vidéos. Dans cette optique, nous déterminons dans un premier temps une structure d'indexation adéquate. Nous mettons à profit une méthode de classification hiérarchique arborescente (CHA), [5, 14]. Elle exploite une norme L_2 dans l'espace des attributs de mouvement et tend à minimiser la variance intra-classe. La représentation hiérarchique obtenue permet de réaliser une recherche efficace et pertinente dans la base de vidéos en réponse à une requête exprimée sous la forme d'une séquence d'images.

3 Extraction des attributs de mouvent globaux

3.1 Mesures locales de mouvement

En considérant l'hypothèse d'invariance de l'intensité le long des trajectoires du mouvement 2D, l'équation de contrainte du mouvement apparent permet d'exprimer la vitesse normale v_n, parallèle au gradient spatial de l'intensité, en un point p, en fonction des dérivés spatio-temporelles de la fonction intensité I, [11]:

$$v_n(p) = \frac{-I_t}{\|\nabla I(p)\|} \tag{1}$$

où $\nabla I = (I_x, I_y)$ est le gradient spatial de l'intensité et $I_t(p)$ la dérivé temporelle de l'intensité.

Si le mouvement apparent est orthogonal au gradient spatial de l'intensité, la quantité v_n est nulle quelle que soit l'amplitude du mouvement. D'autre part, v_n est aussi très sensible au bruit lié au calcul des dérivés de l'intensité. Ainsi, une moyenne pondérée de v_n dans un

voisinage forme une mesure de mouvement plus fiable, [16] :

$$v_{obs}(p) = \frac{\sum_{s\in\mathcal{F}(p)} \|\nabla I(s)\|^2 \cdot |v_n(s)|}{\max(\eta^2, \sum_{s\in\mathcal{F}(p)} \|\nabla I(s)\|^2)} \tag{2}$$

où $\mathcal{F}(p)$ est une fenêtre 3×3 centrée en p, et η^2 une constante prédéterminée relative au niveau de bruit dans les régions uniformes. v_{obs} fournit une mesure locale directement liée au mouvement apparent, aisément calculable et fiable. La perte de l'information de direction n'est pas réellement préjudiciable puisque nous cherchons à définir un typage général du contenu dynamique de la scène observée.

Le calcul des cooccurrences ne peut être réalisé sur un ensemble de mesures continues. Du fait de l'étalement de la distribution des quantités v_{obs}, une simple quantification linéaire sur l'intervalle $[\inf_p v_{obs}(p); \sup_p v_{obs}(p)]$ n'est pas pertinente. Puisqu'il est généralement admis en analyse du mouvement qu'une approche mono-résolution ne permet pas d'appréhender les grands déplacements, nous fixons une borne supérieure au-delà de laquelle la quantité de mouvement n'est plus jugée exploitable. Ainsi, en pratique, nous appliquons une quantification linéaire sur 16 niveaux dans l'intervalle $[0, 4]$. Les quantités de mouvement seront notées $obs(p)$.

3.2 Attributs globaux de mouvement

Dans [15], des cooccurrences spatiales sont associés à la vitesse normale pour classer des exemples en terme de mouvement simple (rotation, divergent) ou de texture temporelle (rivière, feuillage). Dans ce cas, comme les interactions étudiées sont uniquement spatiales, il est seulement possible de caractériser les mouvements stationnaires suivant l'axe temporel. De plus, pour interpréter la structure spatiale du mouvement, il est nécessaire d'étudier plusieurs configurations d'interaction spatiale, ce qui accroît le temps de calcul requis. Par conséquent, nous préférons considérer des cooccurrences temporelles, définies pour une paire de quantités de mouvement (i, j) par :

$$P_{d_t}(i,j) = \frac{\sharp\{(r,s) \in C_{d_t} / obs(r) = i, obs(s) = j\}}{|C_{d_t}|} \tag{3}$$

où obs est la version quantifiée de la mesure locale v_{obs}, et C_{d_t} un système de voisinage temporel défini par $C_{d_t} = \{(r,s)$ à la même position spatiale dans l'image tels que $r \in \text{image}(t)$ et $s \in \text{image}(t - d_t)\}$. À partir de ces matrices de cooccurrence, nous définissons des attributs globaux du mouvement du même type que ceux décrits dans [10] pour la caractérisation de textures :

$$\begin{cases} f^1 = \sum_{(i,j)} P_{d_t}(i,j) \log(P_{d_t}(i,j)) \\ f^2 = \sum_{(i,j)} P_{d_t}(i,j)/[1 + (i-j)^2] \\ f^3 = \sum_{(i,j)} (i-j)^2 P_{d_t}(i,j) \\ f^4 = \left[\sum_{(i,j)} i^4 P_{d_t}(i,j)\right] / \left[\sum_{(i,j)} i^2 P_{d_t}(i,j)\right]^2 - 3 \\ f^5 = \left[\sum_{(i,j)} (i-j)^4 P_{d_t}(i,j)\right] / \left[\sum_{(i,j)} (i-j)^2 P_{d_t}(i,j)\right]^2 - 3 \end{cases} \tag{4}$$

où f^1 est l'entropie, f^2 la moyenne pondérée par l'inverse du contraste, f^3 le contraste, f^4 le kurtosis et f^5 et le kurtosis des différences.

Ces attributs globaux du mouvement sont calculés sur l'ensemble de l'image. Pour interpréter des mouvements non-stationnaires dans le domaine spatial, nous pouvons aisément obtenir une caractérisation au niveau d'une région. En fait, l'extraction de ces attributs peut aussi être réalisée sur des blocs prédéfinis ou sur des régions issues d'une segmentation préalable, puisque nous nous intéressons uniquement aux interaction temporelles. Dans ce cas, le processus de recherche dans la base de vidéo consiste à retrouver des séquences dont certaines régions sont similaires en terme de mouvement à celle de la séquence proposée comme requête.

4 Indexation et recherche basées sur le mouvement

4.1 Indexation par le mouvement

Afin de construire un système efficace d'indexation et de recherche sur un ensemble de séquences vidéo basé sur les attributs de mouvement présentés précédemment, nous devons en premier lieu extraire une représentation appropriée de la base de vidéos. Ceci nous permettra de rechercher aisément des exemples similaires, en terme de propriétés de mouvement, à une vidéo fournie comme requête par l'utilisateur. A cet effet, nous utilisons une méthode de classification permettant de regrouper des vidéos dans des classes pertinentes au sens de la distribution de mouvement.

Nous avons choisi une méthode itérative appelée *classification hiérarchique ascendante* (CHA), [5]. En raison de sa simplicité et de sa nature hiérarchique, elle se révèle particulièrement intéressante dans le contexte de l'indexation vidéo, [14]. Elle vise à extraire un arbre binaire exprimant des similarités entre vidéos dans l'espace des attributs relativement à une métrique donnée, (cf. Fig.2).

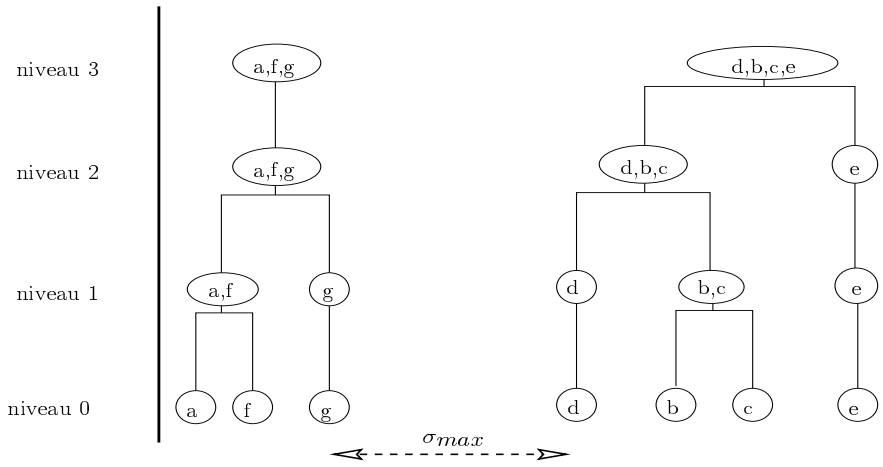


FIG. 2 – *Exemple de classification hiérarchique arborescente à 4 niveaux dans un espace à une dimension pour un ensemble à sept éléments* $\{a, b, c, d, e, f, g\}$. *Dans ce cas à une dimension, chaque élément de la base de données est assimilé à l'attribut qui le représente. Pour chaque niveau de l'arbre, la procédure de classification utilisée consiste à former les paires successives des éléments les plus proches. On introduit en outre une constante* σ_{max} *pour éviter de fusionner les noeuds trop éloignés. Ce cas se présente par exemple pour les noeuds* g *et* d *au niveau* 0 *de l'arbre binaire représenté ci-dessus.*

Pour quantifier la similarité entre deux séquences liée aux propriétés de mouvement, nous exploitons la distance euclidienne dans l'espace des attributs. Toutefois, la nature différente des descripteurs de mouvement ne permet pas leur comparaison directe par le biais d'une norme L_2. Par conséquent, nous procédons au préalable à une normalisation des moments d'ordre supérieur pour les rendre homogènes à des moments d'ordre 1. Ainsi, nous calculons la racine carrée pour l'attribut f^3 et la racine quatrième des attributs f^4 et f^5. Nous associons alors ces nouvelles quantités aux descripteurs f^1 et f^2 pour former un nouveau jeu d'attributs $g = (g^1, g^2, g^3, g^4, g^5)$. Cette normalisation nous permet alors de définir la distance entre deux éléments de la base de vidéo n et m de vecteurs d'attributs respectifs g_n et g_m en exploitant la distance euclidienne :

$$d(n,m) = \left[\sum_{a=1}^{5} (g_n^a - g_m^a)^2 \right]^{\frac{1}{2}} \tag{5}$$

La définition de cette métrique permet d'effectuer simplement une classification à partir de la proximité géométrique dans l'espace des attributs. La procédure CHA vise à construire une représentation hiérarchique minimisant la variance intra-classe et maximisant la variance inter-classe. Considérons deux séquences n et m, l'écart-type intra-classe du noeud x issu de leur fusion est donnée par la distance entre n et m :

$$\sigma_x = d(n,m) \tag{6}$$

La représentation du noeud x dans l'espace d'attributs est associée à la moyenne des vecteurs d'attributs g_n et g_m.

La procédure CHA est initialisée en considérant que chaque élément de la base de séquences traitée constitue une feuille de l'arbre binaire. L'algorithme de construction de la représentation de la base de données procède itérativement en fusionnant des paires de noeuds à un niveau donné pour former les représentants au niveau supérieur. Nous utilisons un algorithme itératif sous-optimal qui assure la minimisation de la variance intra-classe. Ainsi, pour un niveau donné associé à un ensemble de noeuds $\mathcal{N}$, nous cherchons à déterminer la paire $\widehat{(n,m)}$ d'éléments les plus proches :

$$\widehat{(n,m)} = \arg \min_{(k,l) \in \mathcal{N}^2} \sigma_{(k,l)} \tag{7}$$

où $\sigma_{(k,l)}$ représente la l'écart-type intra-classe de la paire (k,l). Son expression est donnée par l'équation (6). Après une première fusion de noeuds $\widehat{(n,m)}_0$, on réalise la recherche d'une nouvelle paire vérifiant l'équation (7) pour l'ensemble $\mathcal{N} \setminus \widehat{(n,m)}_0$. Nous itérons cette procédure de formation de nouvelles paires de noeud tant que l'écart-type de la paire vérifiant (7) est inférieur à une constante σ_{max} fixée. Cette condition revient à interdire la fusion des noeuds trop éloignés dans l'espace des attributs. Elle garantit en outre la cohérence des propriétés de mouvement des classes de séquences formées aux niveaux supérieurs de la représentation. Dans ce cas, les noeuds isolés sont projetés au niveau supérieur de l'arbre.

Dans nos expériences, nous devons fixé une valeur de distance temporelle d_t pour l'extraction des descripteurs de mouvement. Des résultats présentés dans [2] montrent que ce paramètre de calcul des cooccurences a peu d'influence sur les descripteurs de mouvement, qui ne sont pas liés au contraste. Ceci nous permet donc d'effectuer un choix arbitraire. En pratique, nous avons considéré : $d_t = 1$.

4.2 Recherche d'exemples similaires à une requête

Nous souhaitons rechercher des séquences similaires à une requête exprimée sous la forme d'une vidéo. Plus précisément, nous cherchons à mettre en correspondance des séquences relativement à des propriétés de mouvement. En fait, la structure d'indexation décrite précédemment nous fournit un outil hiérarchique efficace pour une recherche d'exemples similaires en terme de mouvement.

Dans un premier temps, nous déterminons la représentation hiérarchique de la base de séquences vidéo. Ensuite, nous traitons la vidéo correspondant à la requête émise afin d'en extraire les attributs de mouvement globaux. De la même manière que précédemment, nous formons le vecteur d'attributs g en normalisant les moments d'ordre supérieur f^3, f^4 et f^5. Nous utilisons alors la norme euclidienne comme fonction de coût.

Au plus haut niveau de l'arbre binaire représentant la base de vidéos, nous sélectionnons le groupe d'éléments le plus proche de la requête au sens de la distance euclidienne entre le vecteur d'attributs normalisés de la requête et celui du centre de gravité d'un élément du niveau considéré dans la représentation hiérarchique. Ensuite, nous procédons de la même façon pour choisir l'élément le plus proche parmi les fils du noeud sélectionné au niveau supérieur. Cette procédure est itérée à travers la structure d'indexation jusqu'à l'obtention d'un nombre donné de réponses ou d'une précision donnée.

5 Résultats

Nous utilisons l'approche présentée précédemment pour traiter une base de séquences. Nous avons attaché une importance particulière au choix d'exemples de vidéo représentatifs de situations de mouvement variées. Notre base de données inclut des textures temporelles, des exemples d'activité de mouvement importante (scènes de basket, d'équitation,...), des mouvements rigides (voitures, trains,...), et des séquences comportant une activité de mouvement faible. Finalement, nous disposons d'une base de 25 séquences d'images (typiquement, chaque vidéo est composée de 10 images).

Tout d'abord, nous appliquons l'algorithme CHA dans l'espace des attributs normalisés (g^1, g^2, g^3, g^4, g^5). En Fig.3, nous donnons la représentation de la base de vidéos dans l'espace des attributs, réduit à l'espace (g^3, g^4, g^5) pour la visualisation. Les quatre classes de séquences du niveau 4 de la hiérarchie sont réellement liées à des contenus dynamiques différents : la classe représentée par le symbole "o" est composée d'exemples de texture temporelle, la classe "x" représente les mouvements non-rigides, et les classes "+" et "." font références à des mouvements rigides avec une activité de mouvement faible ou très faible.

À la Fig. 4, nous présentons des résultats de recherche d'exemples similaires, en terme de propriétés globales du mouvement, à une vidéo qui constitue la requête. Le nombre maximum de réponses est fixé à trois. La procédure de recherche fournit des résultats intéressants. Dans le premier exemple, nous mettons en correspondance différentes vidéo de sport comportant donc une activité de mouvement importante. La deuxième requête envisagée est un plan fixe sur une salle de réunion. La procédure de recherche fournit d'autres exemples de séquences présentant une activité de mouvement très faible.

Nous avons cherché à évaluer de manière plus quantitative le système d'indexation vidéo proposé. Pour ce faire, nous avons suivi la démarche suivante. En premier lieu, nous déterminons une classification a priori de la base de séquences en terme de propriétés de mouvement. Ensuite, nous analysons les trois meilleurs réponses fournies par le système d'indexation et de recherche en considérant chaque élément de la base comme une requête. Pour évaluer la qualité de la recherche

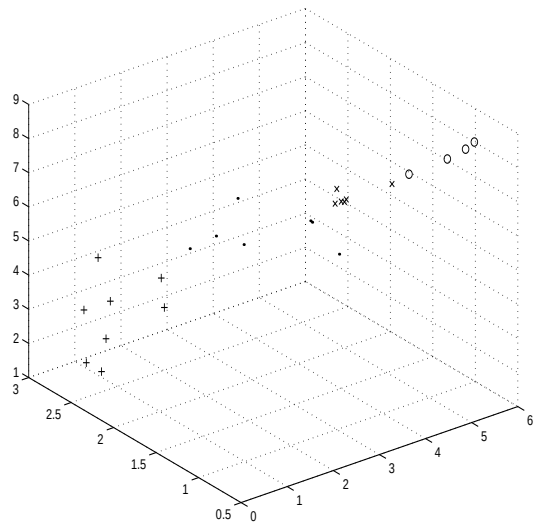
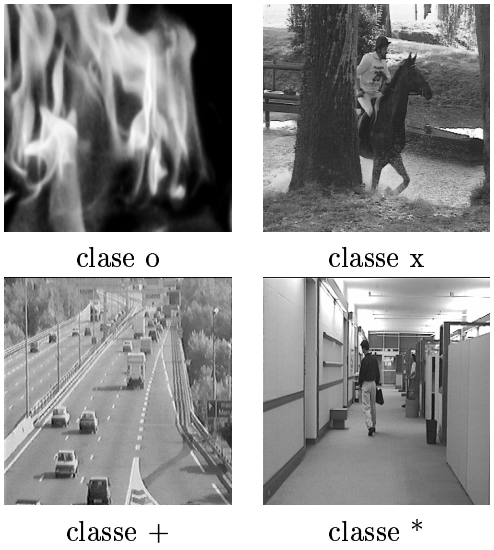


Figure 2.a



clase o | classe x

classe + | classe *

Figure 2.b

FIG. 3 – *Représentation de la base de séquences vidéo obtenue avec la méthode CHA : a) représentation des séquences dans le sous-espace des attributs (g^3, g^4, g^5). Les symboles (+,o,.,*) représentent les index des quatre classes du niveau 4 de l'arbre binaire. b) Exemples représentatifs de chaque classe. Nous présentons la première image de la séquence la plus proche du centre de gravité de chaque classe extraite.*

requête 1 : activité importante | réponse 1 | réponse 2 | réponse 3

requête 2 : activité faible | réponse 1 | réponse 2 | réponse 3

FIG. 4 – *Résultats de recherche d'exemples similaires, en terme de propriétés de mouvement, à une requête exprimée sous la forme d'une vidéo. Nous avons fixé un maximum de trois réponses. Nous présentons la première image de chaque séquence sélectionnée.*

et de la classification, nous considérons deux mesures : nous comptons le nombre de fois que la requête apparaît comme la meilleure réponse et le nombre de fois que la deuxième meilleure réponse appartient à la même classe a priori. Nous avons envisagé les quatre classes suivantes : les séquences comportant respectivement une faible activité de mouvement et des mouvements rigides pour les deux premières classes, les exemples d'activité de mouvement importante pour la troisième classe, et enfin les textures temporelles constituent la quatrième classe de séquences. Même si la procédure d'évaluation retenue reste subjective, elle fournit un premier moyen de validation du système d'indexation vidéo proposé. Les résultats obtenus pour l'ensemble de la base de séquences sont ainsi prometteurs sachant que les critères adoptés sont assez stricts :

requête et meilleure réponse fournie identique (%)	80
taux de bonne classification relativement aux classes a priori (%)	75

TAB. 1 – *Évaluation du système d'indexation vidéo par le contenu de mouvement*

6 Conclusion

Nous avons décrit dans cet article une méthode originale pour l'extraction d'attributs de mouvement globaux et son application à l'indexation de vidéo par le contenu. Notre approche repose sur une analyse de statistiques du second ordre de distributions temporelles de mesures locales de mouvement. Nous exploitons une classification hiérarchique ascendante pour déterminer une représentation de la base de séquences d'images sous la forme d'un arbre binaire. Nous avons obtenu des résultats très encourageants en terme de classification des séquences et de recherche d'exemples similaires relativement à un requête exprimée sous la forme d'une vidéo. Dans nos prochains travaux, nous chercherons à déterminer un ensemble optimal de descripteurs adapté à différents types de contenu dynamique et à évaluer notre système d'indexation sur une base de vidéos plus importante.

Références

[1] P. Aigrain, H.J. Zhang, et D. Petkovic. Content-based representation and retrieval of visual media : a state-of-the-art review. *Multimedia Tools and Applications*, 3(3):179–202, novembre 1996.

[2] P. Bouthemy et R. Fablet. Motion characterization from temporal cooccurrences of local motion-based measures for video indexing. In *Proc. Int. Conf. on Pattern Recognition, ICPR'98*, Brisbane, août 1998.

[3] P. Bouthemy et F. Ganansia. Video partioning and camera motion characterization for content-based video indexing. In *Proc. 3rd IEEE Int. Conf. on Image Processing, ICIP'96*, Lausane, septembre 1996.

[4] J.D. Courtney. Automatic video indexing via object motion analysis. *Pattern Recognition*, 30(4):607–625, avril 1997.

[5] E. Diday, G. Govaert, Y. Lechevallier, et J. Sidi. Clustering in pattern recognition. In *Digital Image Processing*, pages 19–58. J.-C. Simon, R. Haralick, eds, kluwer edition, 1981.

[6] U. Gargi, R. Kasturi, and S. Antani. Performance characterization and comparison of video indexing algorithms. In *IEEE int. Conf. on Computer Vision and Pattern recognition, CVPR'98*, pages 559–565, Freiburg, Germany, juin 1998.

[7] M. Gelgon et P. Bouthemy. Determining a structured spatio-temporal representation of video content for efficient visualization and indexing. In *Proc. 5th European Conf. on Computer Vision, ECCV'98*, Freiburg, juin 1998.

[8] M. Gelgon, P. Bouthemy, and T. Dubois. A region tracking technique with failure detection for an interactive video indexing environment. In *Proc. Visual'99*, Amsterdam, juin 1999.

[9] B. Gunsel, A. Murat Tekalp, et P.J.L. van Beek. Content-based access to video objects : temporal segmentation, visual summaruzation and feature extraction. *Signal Processing*, 66:261–280, 1998.

[10] R.M. Haralick, K. Shanmugan, et I. Dinstein. Textural features for image classification. *IEEE Trans. on Systems, Man and Cybernetics*, 3(6):610–621, novembre 1973.

[11] B. Horn et B. Schunk. Determining optical flow. *Artificial Intelligence*, 17:185–203, 1981.

[12] F. Idris et S. Pandranathan. Review of image indexing techniques. *Jal of Visual Communication and Image Representation*, 8(2):146–166, juin 1997.

[13] M. Irani et P. Anandan. Video indexing based on mosaic representation. *IEEE Trans. on PAMI*, 86(5):905–921, mai 1998.

[14] R. Milanese, D. Squire, et T. Pun. Correspondence analysis and hierarchical indexing for content-based image retrieval. In *Proc. 3rd IEEE Int. Conf. on Image Processing, ICIP'96*, Lausanne, septembre 1996.

[15] R. Nelson et R. Polana. Qualitative recognition of motion using temporal texture. *CVGIP : Image Understanding*, 56(1):78–99, juillet 1992.

[16] J.M. Odobez et P. Bouthemy. Separation of moving regions from background in an image sequence acquired with a mobile camera. In *Video Data Compression for Multimedia Computing*, chap. 8, pages 295–311. H. H. Li, S. Sun, and H. Derin, eds, Kluwer, 1997.

[17] K. Otsuka, T. Horikoshi, S. Suzuki, et M. Fujii. Feature extraction of temporal texture based on spatiotemporal motion trajectory. In *Proc. Int. Conf. on Pattern Recognition, ICPR'98*, Brisbane, août 1998.

[18] N.V. Patel et I.K. Sethi. Video shot detection and characterization for video databases. *Pattern Recognition*, 30(4):607–625, avril 1997.

[19] G. Sudhir and J.C.M. Lee. Video annotation by motion interpretation using optical flow streams. *Jal of Visual Communication and Image Representation*, 7(4):354–368, déc. 1996.

[20] M. Szummer et R.W. Picard. Temporal texture modeling. In *Proc. 3rd IEEE Int. Conf. on Image Processing, ICIP'96*, Lausanne, septembre 1996.

[21] N. Vasconcelos et A. Lippman. A Bayesian video modeling framework for shot segmentation and content characterization. In *IEEE Workshop on Content-Based Access of Image and Video Libraries*, San Juan, Puerto-Rico, juin 1997.

[22] W. Xiong et J. C.H. Lee. Efficient scene change detection and camera motion annotation for video classification. *Computer Vision and Image Understanding*, 71(2):166–181, août 1998.

[23] H.J. Zhang, J. Wu, D. Zhong, et S. Smoliar. An integrated system for content-based video retrieval and browsing. *Pattern Recognition*, 60(4), avril 1997.