Le suivi de *blobs* comme base pour la caractérisation du mouvement dans des séquences audiovisuelles

Rémi Mégret, Jean-Michel Jolion
{megret,jolion}@rfv.insa-lyon.fr
Laboratoire RFV, INSA de Lyon

Résumé

Cet article aborde le problème de l'extraction d'information de mouvement pour l'indexation de documents audiovisuels. Nous proposons une méthode basée sur le suivi de primitives fiables, privilégiant ainsi la dimension temporelle du mouvement. Les primitives considérées sont des blobs *issus de la théorie de l'espace-échelle. Leur comportement temporel est étudié sur des vidéos de type audiovisuelles, validant l'utilisation d'une méthode de suivi adaptée du* multi-hypothesis tracker. *Des exemples commentés illustrent la méthode.*

1 Introduction

Outre les approches par caractérisation globale du mouvement (mouvement dominant [2], ou activité visuelle [8]), la plupart des méthodes existantes pour l'indexation de vidéos par le mouvement [4] donnent la priorité à une structuration spatiale basée sur l'image ou une estimation du mouvement à court-terme (suivi de partitions [6], segmentation des vecteurs MPEG [7], segmentation à l'aide de modèles paramétriques [1]). La dimension temporelle n'est considérée que dans un deuxième temps, par suivi et trajectographie des entités segmentées.

Nous proposons dans cet article une approche alternative basée sur l'extraction de trajectoires par suivi long-terme de primitives fiables. Cette méthodologie s'inscrit dans le cadre de la mise au point de techniques pour l'indexation par le mouvement avec les contraintes suivantes.

Généralité: Le système doit pouvoir traiter tout type de document audiovisuel. En particulier, la résolution et la qualité des vidéos utilisées pour l'indexation sont généralement limitées. De plus, les vidéos peuvent contenir des mouvements de forte amplitude, du flou, des occlusions et des mouvements déformables.

Information long-terme: La caractérisation du mouvement doit rendre compte des différents mouvements indépendants présents, tout au long de leur étendue temporelle. L'analyse du comportement dynamique des objets importe ainsi plus que leur localisation spatiale précise.

Construction incrémentale: Dans une optique d'indexation il n'est pas toujours nécessaire de calculer trop de détails. Il doit être possible de mettre en œuvre des moyens limités pour obtenir des descriptions grossières, et de raffiner lorsque cela s'avère nécessaire.

La figure 1 illustre les étapes principales du processus d'analyse. Dans la suite, nous nous focaliserons sur l'obtention des trajectoires à partir des primitives détectées: des *blobs* de l'espace échelle. Les lecteurs intéressés par leur utilisation pour une représentation simplifiée du mouvement pourrons se reporter à [11].

2 L'espace-échelle comme source de primitives à suivre

Définition La théorie de l'espace-échelle [9] fournit des descripteurs de la structure visuelle de chaque image. L'image analysée est filtrée par convolution avec des filtres gaussiens de tailles croissantes (Nous noterons par la suite σ l'écart-type d'un tel filtre gaussien). Ceci permet de mettre en évidence les structures présentes à différentes échelles. En effet chaque extremum local d'une image filtrée correspond à une zone dont la luminance contraste avec son contexte à l'échelle considérée. Cette zone ainsi que les caractéristiques qui lui sont liées sont appelées un *blob*. Les minima et les maxima sont traités de façon séparés, engendrant ainsi deux types de *blobs*.

Propriétés Ces *blobs* possèdent certaines propriétés intéressantes relativement au suivi et aux contraintes de généralité.

- Compacité: Ils constituent des marqueurs spatialement compacts, qui peuvent être résumés sous la forme d'un centre et de descripteurs de la zone de support.
- Invariance: L'espace-échelle est invariant par rotation ou translation. Tout changement d'échelle de l'image d'origine se traduit par un simple décalage dans l'espace-échelle.

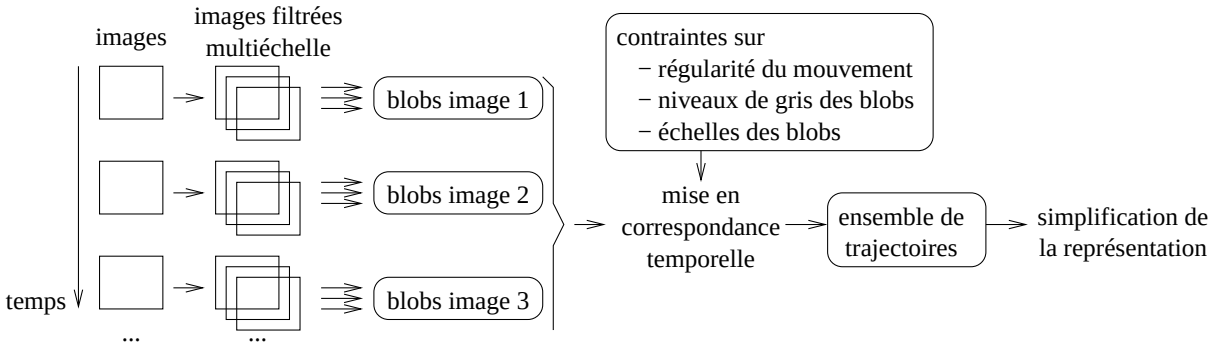

FIG. 1 – *Étapes du processus d'analyse.*

- Robustesse au bruit: Les dégradations de l'image, dues à la compression notamment, sont principalement situées dans les hautes fréquences. La détection de primitives *blobs* étant basée sur un préalable filtrage passe-bas, l'influence du bruit s'en trouve amoindrie.
- Répartition spatiale: Contrairement à des primitives telles que les coins ou les contours, qui peuvent être très concentrées en certaines parties de l'image (présence de texture) et absentes ailleurs, les *blobs* ont une répartition relativement équilibrée, et sont détectés sur toute image présentant un minimum de structure spatiale.

3 Suivi des *blobs* dans des séquences audiovisuelles

L'utilisation de *blobs* en tant que primitives à suivre a été proposée dans [3], où l'accent est porté sur la détection de l'échelle optimale de chacun, sur des images bien adaptées aux détecteurs utilisés. Nous placerons ici le suivi de *blobs* dans le cadre de notre problématique: l'analyse générique du mouvement dans des vidéos audiovisuelles. Pour plus de détails, voir [10].

Comportement dynamique L'étude du comportement dynamique des *blobs* dans des vidéos de type audiovisuel nous a permis de vérifier la stabilité et la persistance des extrema à échelle fixe, au cours des déplacements (voir la figure 2 pour un exemple commenté). Dans des cas particuliers, ces qualités peuvent être perturbées par des évènements de division (un *blob* se séparant en deux *blobs* plus petits) ou de fusion (deux *blobs* se joignent pour former un *blob* plus gros), qui se produisent lorsque l'échelle d'analyse n'est pas adaptée à la taille des entités visuelles détectées. Même dans ces conditions, on constate cependant une bonne stabilité de l'extremum, que nous utiliserons comme primitive à suivre.

Mise en correspondance Pour la mise en correspondance temporelle, nous avons adapté la méthode de suivi multi-hypothèses présentée dans le cadre de suivi de points d'intérêt [5]. Cet algorithme construit des arbres d'hypothèses d'appariements entre plusieurs images successives, et élague celui-ci pour ne conserver que les meilleures hypothèses, afin d'éviter l'explosion combinatoire. Le critère utilisé tient à la fois compte des contraintes d'exclusion (une primitive appartient à une seule trajectoire), et maximise la régularité de la trajectoire ainsi que l'invariance des caractéristiques liées aux primitives.

Les caractéristiques que nous avons utilisées sont très peu nombreuses. En effet pour chaque blob, en plus de la position de son extremum, ne sont considérées que la valeur associée dans l'image filtrée, et son échelle. Seuls sont mis en correspondances les blobs d'une même échelle, mais il serait aussi envisageable d'autoriser des appariements entre échelles voisines, comme dans [3]. Un seuillage n'autorise les appariements que de *blobs* dont les valeurs diffèrent de moins de 32 niveaux de gris, pour des niveaux compris dans $[0 \dots 255]$.

Mise en œuvre La méthode proposée a été implémentée en C++ et testée sur un Pentium 500 Mhz. Les traitements nécessaires peuvent se diviser en deux: la détection des *blobs*, et leur suivi. Pour toutes nos expériences nous avons utilisé des vidéos au format MPEG-1, de taille comparable à 352×288 pixels.

La détection se passe sur les images décodées, et n'implique comme traitement coûteux que des filtrages gaussiens. Ces filtrages étant de taille relativement élevée, nous procédons à une diminution de chaque dimension de l'image d'un facteur deux sans modifier sensiblement les résultats. Pour capturer un large spectre d'échelles nous avons considéré 6 échelles entre $\sigma = 5.6$ et $\sigma = 32$. Cette étape de détection prends alors autour de 2 s par image pour l'ensemble des échelles. Elle produit un nombre de blobs de chaque type (minimum ou maximum) de l'ordre de 250 pour $\sigma = 5.6$, 150 pour $\sigma = 8$ et 50 pour $\sigma = 16$.

L'étape de suivi charge les primitives image par image, créant et effaçant les hypothèses au fur et à mesure de

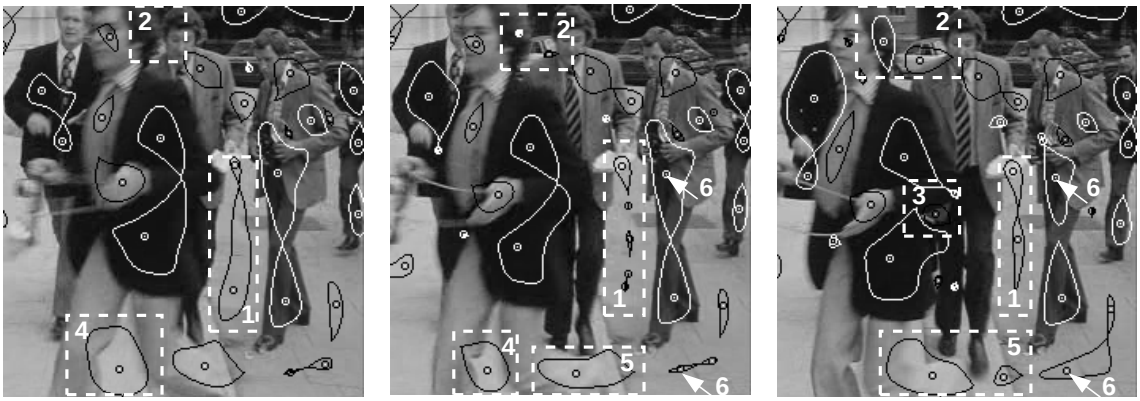

image 6400. image 6401. image 6402.

FIG. 2 – *Illustration du comportement temporel des* blobs *à échelle fixe ($\sigma = 10$). Les rectangles en pointillés blancs mettent en évidence des zones intéressantes.*
(1): Un blob *se divise en plusieurs* blobs*, puis refusionne. (2): La fin de l'occultation entraîne la création de deux petits* blobs *de polarité opposées, qui grossissent par la suite. Le* blob *clair correspond à une zone qui apparaît réellement, alors que le* blob *sombre est créé par réaction au contraste entre la chevelure et la voiture, qui n'existait par auparavant à cause de la chevelure du personnage de derrière. (3): Les* blobs *sombres voisins voient leurs zones de support déformées par l'apparition d'un* blob *clair, sans que l'extremum associé ne soit trop perturbé. (4): Phénomène d'ouverture (en anglais* aperture*): le* blob *clair entre les jambes a le même mouvement apparent qu'elles. (5): Les occultations entraînent aussi des divisions de* blobs*. (6) Flèches blanches: quand un* blob *se divise, l'extremum associé correspond souvent à l'extremum de l'un des* blobs *résultats. Ceci peut également être observé dans les rectangles (1) et (5).*

la progression. Les temps de calcul sont de l'ordre de 2.5 s par image et par type de *blob*, pour l'ensemble des échelles considérées, dans les conditions présentées.
Les facteurs principaux influant sur ce temps sont le nombre de *blobs* considérés, et surtout l'ambiguité d'appariement, c'est à dire le nombre de mises en correspondances possibles pour chaque *blob*. Cette ambiguité est d'autant plus grande que les *blobs* ont des caractéristiques similaires. Ainsi, dans [3], une mesure de corrélation entre *patches* est utilisée. Il est apparu, d'après nos expériences sur vidéos naturelles, qu'un tel critère n'etait pas nécessaire pour les échelles $\sigma >= 5$. Ainsi le seuillage sur la valeur du blob, couplé à la régularité de la trajectoire, se révèle suffisant pour capturer la plupart des déplacements présents (figure 3). Le seuil de 32 niveaux de gris, déterminé empiriquement, a été utilisé avec succès sur toutes les séquences étudiées.

4 Conclusion

Le suivi de *blobs* présenté dans cet article constitue une étape bas-niveau permettant d'extraire à partir des images brutes un ensemble de trajectoires qui représentent une première information de mouvement à long-terme. Le choix des primitives est pour une bonne part dans l'applicabilité de la méthode à des vidéos diverses. Une telle approche nous permettra par la suite d'aborder la construction incrémentale des descripteurs par la sélection a priori des *blobs*, et une construction des trajectoires par ordre décroissant de confiance. Des techniques telles que le clustering de trajectoires [11], peuvent alors être appliquées pour simplifier la représentation. On disposera ainsi de descripteurs adaptés à notre problématique (généralité, information long-terme, caractère incrémental), qui pourront s'insérer au sein d'un système d'indexation par le mouvement.

Références

[1] S. Ayer et H.S. Sawhney. « Layered representation of motion video using robust maximum-likelihood estimation of mixture models and MDL encoding ». Dans *Proc. IEEE Int. Conf. on Computer Vision*, pages 777–784, 1995.

[2] P. Bouthemy, M. Gelgon, et F. Ganansia. « A unified approach to shot change detection and camera motion characterization ». Rapport de Recherche 1148, IRISA, novembre 1997.

[3] L. Bretzner et T. Lindeberg. « On the handling of spatial and temporal scales in feature tracking ». Dans *Proc. 1st Int. Conf. on Scale-Space Theory in Computer Vision*, Springer-Verlag Lecture Notes in Comp. Vis., vol 1252, Utrecht, Pays-Bas, juillet 1997.

[4] R. Brunelli, O. Mich, et C.M. Modena. « A Survey on Video Indexing ». *Journal of Visual Commu-*

(a) (b) (c)

FIG. 3 – *Exemples de trajectoires obtenues par suivi mono-échelle dans des vidéos naturelles Exemples de trajectoires obtenues par suivi mono-échelle dans des vidéos naturelles ((b) et (c) sont tirées du corpus de vidéos de l'INA mis au point pour le GdR ISIS). L'extrémité plus épaisse de chaque trajectoire correspond à la position de l'extremum suivi dans l'image courante. Le reste de la trajectoire représente les positions antérieures, dans le référentiel de la caméra. Afin d'illustrer la stabilité temporelle des* blobs*, ne sont affichées que les trajectoires d'extension temporelle supérieure à un seuil de L images ($L = 30$ *pour (a),* 20 *pour (b),* 10 *pour (c)). Dans tous les cas, l'échelle d'analyse est* $\sigma = 8$.
(a): Séquence coastguard. *La caméra a un mouvement latéral, puis un brusque mouvement vers le haut, pendant que les deux bateaux se croisent. On obtient bien des trajectoires sur chacun des bateaux, ainsi que dans le fond.*
(b): Séquence tête*: le personnage bouge la tête latéralement, alors que la foule derrière défile lentement vers la droite. Ces deux mouvement sont représentés par les trajectoires.*
(c): Séquence photographe*: le personnage en avant-plan se déplace vers la gauche à la vitesse de 15 pixels par image. Plusieurs trajectoires lui sont associées. Le mouvement de caméra vers la gauche est visible sur les autres personnages et sur le fond. Bien qu'aucune trajectoire ne soit associée à cette échelle à la chemise claire du personnage en avant-plan, c'est le cas à l'échelle* $\sigma = 5.4$. *Une analyse multi-échelle est donc nécessaire, permettant la détection de primitives complémentaires.*

nication and Image Representation, 10(2):78–112, juin 1999.

[5] I.J. Cox et S.L. Hingorani. « An Efficient Implementation of Reid's Multiple Hypothesis Tracking Algorithm and its Evaluation for the Purpose of Visual Tracking ». *IEEE Trans. on PAMI*, 18(2):138–150, février 1996.

[6] D. Deng et B.S. Manjunath. « NeTra-V: Toward an Object-Based Video Representation ». *IEEE Trans. on Circuits and Systems for Video Technology*, 8(5):616–627, septembre 1998.

[7] N. Dimitrova et F. Golshani. « Motion Recovery for Video Content Classification ». *ACM Trans. on Information Systems*, 13(4):408–439, octobre 1995.

[8] R. Fablet et P. Bouthemy. « Statistical motion-based retrieval with partial query ». Dans *4th Int. Conf. on Visual Information System, Visual 2000*, Lyon, France, novembre 2000.

[9] T. Lindeberg. *Scale-Space Theory in Computer Vision*. Kluwer Academic Publishers, Pays-Bas, 1994.

[10] R. Mégret et J.-M. Jolion. « Scale-space blobs tracking for video dynamic content representation ». Dans *Int. Workshop on Content-Based Multimedia Indexing (CBMI)*, Brescia, Italie, septembre 2001.

[11] R. Mégret et J.-M. Jolion. « Suivi de blobs de niveaux de gris pour la représentation du contenu dynamique d'une vidéo ». Rapport de Recherche RR-2001-05, RFV, INSA de Lyon, septembre 2001.