Dix-septième colloque GRETSI, Vannes, 13-17 septembre 1999

Détection de volets pour l'indexation de vidéo par le contenu

Didier Zugaj et Patrick Bouthemy
IRISA/INRIA
Campus universitaire de Beaulieu, 35042 Rennes Cedex, France
Tel : 02 99 84 71 00, Fax : 02 99 84 71 71,
e-mail : dzugaj@irisa.fr, bouthemy@irisa.fr

Résumé – Nous nous intéressons au partitionnement temporel de vidéo, qui est une étape nécessaire dans l'analyse du contenu d'un document audiovisuel en vue de son indexation et de son exploitation. L'objectif est plus particulièrement la détection de volets qui sont des transitions progressives fréquemment utilisées dans le montage de vidéo et le changement de scènes. La difficulté de ce problème réside dans la grande variété de ces effets spéciaux, tant du point de vue géométrique (volet horizontal, diagonal, vertical, zoom), que de l'effet généré (frontière franche, graduelle, enroulement), qui complique notablement leurs détections. L'originalité de la méthode proposée consiste à exploiter la distribution spatiale des pixels non conformes au mouvement dominant estimé entre paires d'images successives. Les résultats expérimentaux obtenus sur de nombreux exemples réels montrent d'une part la capacité de la méthode à détecter une grande variété de volets, et d'autre part à être robuste aux fausses alarmes dans le cas de scènes complexes.

Abstract – This paper is concerned with the temporal segmentation of video programs into shots with a view to content-based indexing and browsing. More precisely, we focus on a specific type of progressive transitions called wipes. They represent special effects frequently used in video editing to accompany scene changes. The main difficulty is that a wipe can exhibit different patterns. The originality of the proposed method consists in exploiting the spatial distribution of pixels that do not conform to the dominant image motion estimated between successive images of the video. The proposed method has been validated by experiments carried out on MPEG-1 videos containing a wide variety of documents (news, series, documentaries, etc...). The detection of wipes is fast, accurate and the false alarm rate is very low in spite of the complexity of most processed scenes.

1 Introduction

L'accès rapide à un passage précis d'un document audiovisuel numérique nécessite que ce document ait été indexé, c'est à dire que des informations essentielles décrivant son contenu aient été extraites et annotées. Une première étape consiste alors à structurer temporellement le document en plans. Un plan est une séquence temporelle homogène au sens cinématographique, comme par exemple un plan de commentateur TV ou un plan de scène d'intérieur. Le raccord entre deux plans, effectué lors du montage du document, peut typiquement relever d'un des trois types de transitions suivants. Le raccord le plus simple est le cut par lequel on passe sans transition d'un plan à l'autre. Le fondu consiste à passer d'une séquence d'images à une autre séquence, en combinant les images de chacune d'elles pendant un intervalle de temps prédéfini, au travers d'une fonction de mélange des intensités. Le volet est construit en faisant varier progressivement la proportion des deux plans de part et d'autre de la transition en terme d'occupation spatiale (dans le cas le plus simple suivant un axe horizontal ou vertical). Cette transition peut par exemple produire un effet de page que l'on tourne, qui explique son utilisation abondante dans les journaux télévisés notamment. La détection de volets reste très peu abordée jusqu'à présent bien que nécessaire pour tout système d'indexation vidéo [2, 5].
L'état de l'art sur la détection des changements de plans [1, 5, 6, 7], montre que la plupart des techniques sont mises en défaut par la présence de mouvements de caméra importants ou bien lorsque des objets mobiles significatifs apparaissent dans les séquences d'images.
L'approche que nous privilégions pour détecter les volets consiste à exploiter la distribution spatiale des pixels non conformes au mouvement dominant estimé entre paires d'images successives.
L'article est organisé de la manière suivante. Nous rappelons tout d'abord brièvement le principe de la détection de changements de plan que nous avions préalablement développée [4]. Nous décrivons ensuite en détail notre méthode de détection de volets, et présentons des résultats expérimentaux obtenus sur de nombreux exemples réels comportant des scènes complexes.

2 Détection de changement de plan

La détection des transitions entre plans repose sur la notion de cohérence temporelle du mouvement dominant qui existe entre images d'un même plan.
Un modèle paramétrique du mouvement dominant, qui dans le cadre de cette étude est un modèle affine 2D, est estimé entre paires d'images successives. L'utilisation d'un M-estimateur dans un schéma incrémental multi-résolution assure l'objectif de robustesse de l'estimation en présence de mouvements secondaires [3]. L'analyse de l'évolution temporelle du nombre de pixels conformes au mouvement dominant ainsi estimé, que l'on nomme *support*, conduit au partitionnement de la séquence d'images en plans [4]. Un test cumulatif de Hinkley permet de détecter les variations significatives du support, et aussi per-

met de localiser avec précision les instants de début et de fin des transitions. Cette technique est particulièrement efficace pour détecter les fondus.

3 Détection des volets

Nous allons conserver le principe de l'exploitation de l'information de mouvement et donc de l'estimation robuste du mouvement dominant, mais au lieu de s'intéresser au support, nous allons cette fois considérer son complémentaire et expliciter une information spatio-temporelle liée à la géométrie particulière des volets.

3.1 Différence temporelle entre histogrammes normalisés

La répartition spatiale des "outliers" (points non conformes au mouvement dominant) sur une image, reflète par nature la construction du volet. Ceci est illustré à la figure 1(a,b). Pour caractériser simplement et efficacement cette situation, nous projetons la carte binaire des outliers sur l'axe horizontal et l'axe vertical de l'image, ce qui permet de manipuler des signaux monodimensionnels qui contiennent les régions d'intérêt recherchées -voir la figure 1(c)-.
On considère des images de taille N lignes par M colonnes, et on note $H_h(t,s)$, resp. $H_v(t,s)$, les histogrammes normalisés des projections horizontale, resp. verticale, des cartes binaires d'outliers de taille N x M à l'instant t, où $s \in [1,..,M]$ pour H_h, et $s \in [1,..,N]$ pour H_v.
Afin de rendre la méthode la plus générique possible face à la grande diversité du contenu de chaque plan, nous avons étudié la répartition des cartes des outliers et leurs évolutions pour différentes situations. La situation la plus favorable est la détection de volet entre deux plans fixes. En effet les profils des histogrammes des projections sont alors très discriminants, et l'identification du volet n'est pas ambigüe.
La lisibilité des projections devient plus délicate lorsque par exemple les deux plans, de part et d'autre du volet, présentent des mouvements dominants différents, ou encore lorsqu'un des deux plans contient de nombreux éléments en mouvement, comme l'illustre la figure 2. On observe toutefois dans l'ensemble des situations évoquées, un "front de propagation" sur les histogrammes successifs des projections des cartes des outliers. On considère, pour caractériser ce front, les valeurs absolues des différences temporelles entre histogrammes successifs :

$$\begin{cases} \Delta H_h(t,s) = |H_h(t-1,s) - H_h(t,s)| \\ \Delta H_v(t,s) = |H_v(t-1,s) - H_v(t,s)| \end{cases} \quad (1)$$

La figure 3 montre la consistence temporelle de $\Delta H_h(t,s)$ au cours d'un volet de type horizontal.

3.2 Mesure de corrélation

Pour la détection des volets, il est alors naturel d'évaluer les corrélations $C_h(t,k)$, resp. $C_v(t,k)$ entre les différences d'histogrammes successives $\Delta H_h(t-1)$ et $\Delta H_h(t)$ pour les projections horizontales, et respectivement $\Delta H_v(t-1)$ et $\Delta H_v(t)$ pour les projections verticales. On calcule $C_h(t,k)$ comme suit:

$$\begin{cases} \frac{1}{M-k} \sum_{s=0}^{M-k-1} \Delta H_h(t-1,s+1)\Delta H_h(t,s+k+1), k \geq 0 \\ \\ \frac{1}{M-|k|} \sum_{s=0}^{M-|k|-1} \Delta H_h(t-1,s-k+1)\Delta H_h(t,s+1), k < 0 \end{cases} \quad (2)$$

$C_v(t,k)$ est obtenu de manière analogue, k est l'indice de translation.
Les corrélations ci-dessus sont proportionnelles d'une part à la hauteur du "front de propagation" relevé par les différences d'histogrammes, mais aussi dépendantes de la structure géométrique que celui-ci véhicule entre images successives au cours d'un volet.
Il faut toutefois prendre en compte le cas de scènes complexes, la carte des outliers peut alors n'être pas structurée spatialement de manière claire, et la densité des outliers être importante. De ce fait, les corrélations entre projections successives peuvent prendre des valeurs élevées sans impliquer nécessairement la présence de volet. On pondère alors les valeurs de corrélation par la variance de leur distribution, en introduisant les moments du second ordre $m_{2h}(t)$, resp. $m_{2v}(t)$, calculés comme suit:

$$\begin{cases} m_{2h}^2(t) = \frac{1}{\lambda_h} \sum_{i=1}^{2M-1} [i - m_{1h}(t)]^2 C_h^*(t,i), \\ m_{2v}^2(t) = \frac{1}{\lambda_v} \sum_{i=1}^{2N-1} [i - m_{1v}(t)]^2 C_v^*(t,i), \\ \text{où:} \\ m_{1h}(t) = \frac{1}{\lambda_h} \sum_{i=1}^{2M-1} i C_h^*(t,i), \\ m_{1v}(t) = \frac{1}{\lambda_v} \sum_{i=1}^{2N-1} i C_v^*(t,i), \\ \lambda_h = \sum_{i=1}^{2M-1} C_h^*(t,i), \quad \lambda_v = \sum_{i=1}^{2N-1} C_v^*(t,i) \\ \text{et:} \\ C_h^*(t,i) = C_h(t,i-M), \quad i = 1,..,2M-1, \\ C_v^*(t,i) = C_v(t,i-N), \quad i = 1,..,2N-1. \end{cases} \quad (3)$$

La figure 4 illustre sur deux exemples typiques la distribution des corrélations, avec les valeurs associées des moments m_{2h}. Nous avons choisi, respectivement, le cas d'une scène complexe à l'intérieur d'un plan comportant de nombreux éléments en mouvement et un mouvement global de la caméra, et celui d'une transition de type volet. On observe pour le volet une distribution des corrélations très étroite, alors que dans le cas de la scène complexe la distribution est très étendue et indique que les projections des outliers ne conservent pas la même structure spatio-temporelle.
Finalement, nous utilisons pour détecter les volets les mesures suivantes:

$$\begin{cases} \text{, }_h(t) = \max_{k=1,..,2M-1} \frac{C_h^*(t,k)}{1 + m_{2h}(t)} \\ \\ \text{, }_v(t) = \max_{k=1,..,2N-1} \frac{C_v^*(t,k)}{1 + m_{2v}(t)} \end{cases} \quad (4)$$

Les instants de début et fin de volet sont obtenus en seuillant , $_h(t)$ et , $_v(t)$. Les valeurs des seuils ont été fixées pour l'ensemble de nos expérimentations respectivement à 1.10^{-3} et 5.10^{-3}.

4 Résultats

Des expérimentations ont été effectuées sur des séquences vidéo au format MPEG-1 fournies par l'Institut National de l'Audiovisuel (INA) pour lesquelles on dispose d'un partitionnement temporel de référence. Ces séquences représentent environ trois heures de programmes comprenant des journaux télévisés, documentaires, et séries.
Nous montrons à la figure 5 les résultats de détection de volets sur une partie d'un journal télévisé, pour lesquels nous reportons les instants de début et la durée des volets dans le tableau ci-dessous. Les instants de référence fournis par l'INA ont été obtenus manuellement. On les confronte aux résultats de notre méthode de détection automatique. La détection des volets est précise, le taux d'erreur est très faible compte tenu de la complexité des scènes observées. On obtient une fausse alarme et une non détection (cas d'un volet combiné avec un fondu) sur une séquence de 22000 images comportant douze volets.

Début et durée des volets			
Valeurs de référence		Valeurs calculées	
349	14	348	16
533	14	532	16
863	16	864	16
1081	16	1081	17
7309	22	7310	22
7935	23	7949	12

Tableau 1: Résultats de localisation des volets.

5 Conclusion

Nous nous sommes intéressés au problème de la détection des transitions progressives de type volet souvent employées dans le montage de vidéos. Nous avons justifié et décrit les différentes étapes de notre approche. La méthode exploite la distribution spatiale des pixels qui ne sont pas conformes au mouvement dominant estimé sur les images successives de la vidéo. Nous mettons en évidence au travers d'exemples de volets, la structure géométrique particulière exhibée par les projections des cartes des outliers sur les axes verticaux et horizontaux de l'image. Le critère de détection choisi est basé sur une mesure de corrélation entre les différences d'histogrammes successifs associés à ces projections. La méthode a été validée sur plusieurs heures de vidéos, la détection est rapide, précise, et robuste, il existe en effet très peu de fausses alarmes en dépit de la complexité de la plupart des scènes traitées.
Remerciements: Les auteurs remercient le département Innovation, à la Direction de la Recherche de l'INA, pour avoir fourni les supports vidéos sur lesquels ont porté les expérimentations et dont une partie des résultats sont illustrés dans cet article. Ce travail est partiellement financé par le projet européen Esprit EP24956 DIVAN.

References

[1] F. Idris et S. Panchanathan. Review of image and video indexing techniques. In *Jal of Visual Communication and Image Representation*, 8(2):146-166, juin 97.

[2] M. Wu, W. Wolf et B. Liu. An algorithm for wipe detection. In *Proc. 5th IEEE Int. Conf. on Image Processing, ICIP'98*, pages 893-897 Vol.3, Chicago, octobre 98.

[3] J.M. Odobez et P. Bouthemy. Robust multiresolution estimation of parametric motion models. *Jal of Visual Communication and Image Representation*, 6(4):348-365, décembre 1995.

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

[5] P. Aigrain et P. Joly. The automatic real-time analysis of film editing and transition effects and its applications. *Computer & Graphics*, 18(1):93-103, 1994.

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

[7] H.J. Zhang, A. Kankanhalli et S.W. Smoliar. Automatic partitioning of full-motion video. *Multimedia Systems*, 1:10-28, 1993.

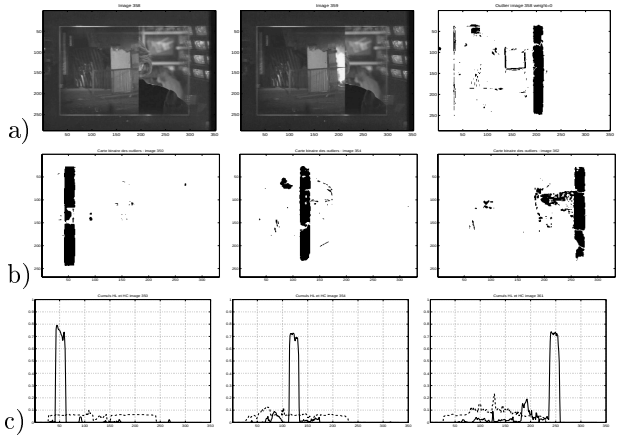

FIG. 1: Géométrie particulière des "outliers" dans une transition de type volet: (a) Images successives à $t_0 + 10$ et $t_0 + 11$, et carte des outliers correspondante, (b) Carte des outliers durant la transition à $t_0 + 2, t_0 + 6, t_0 + 14$, (c) Projections de la carte des outliers sur l'axe horizontal $H_h(t)$ (trait plein), et sur l'axe vertical $H_v(t)$ (en pointillé).

FIG. 2: Evolution de la carte des outliers dans une transition de type volet: (a,b) Un des deux plans contient de nombreux éléments en mouvement, (c,d) Les deux plans possèdent des mouvements dominants différents.

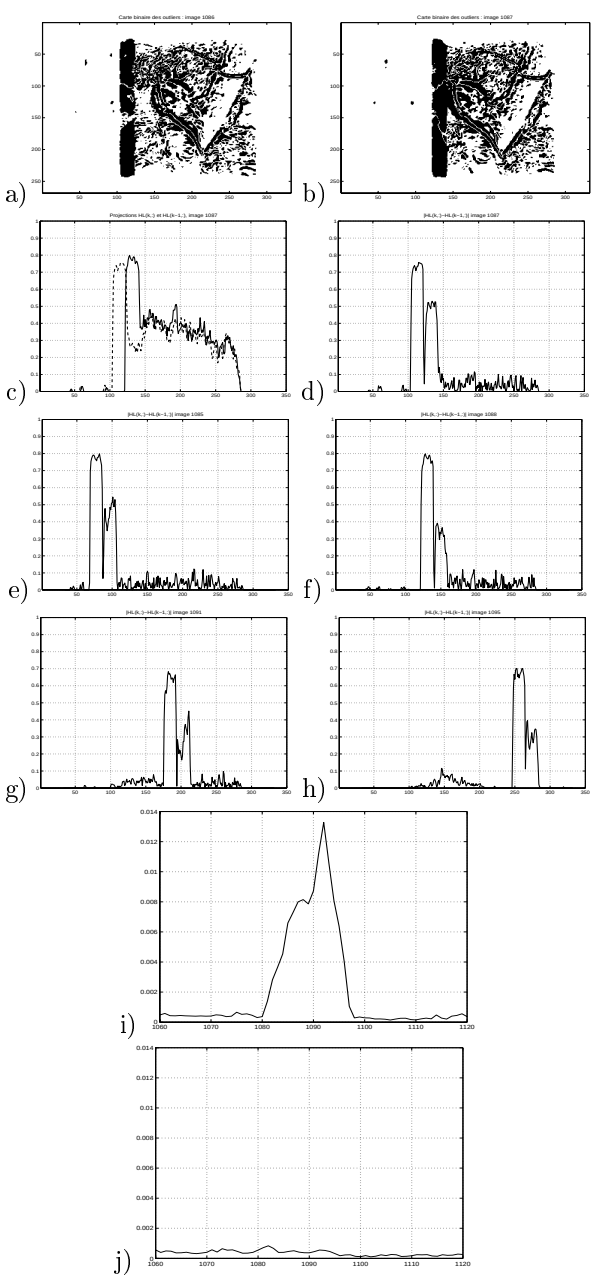

FIG. 3: Calcul des valeurs absolues des différences temporelles entre histogrammes sur l'exemple de la figure 2*(c)*: (a,b) Cartes des outliers pour deux instants successifs; (c) Projections sur l'axe horizontal aux instants $H_h(t-1)$ (en pointillé), et $H_h(t)$ (trait plein), (d) Valeurs absolues des différences des histogrammes de la fig.3c. Le pic correspondant à la position courante du volet est ainsi mis en évidence; (e,..,h) Valeurs absolues des différences au cours de la transition. Le front de propagation apparaît et se déplace pendant la transition; (i,j) Extrema de corrélation , $_h(t)$, , $_v(t)$ sur l'intervalle de temps de 40 images successives.

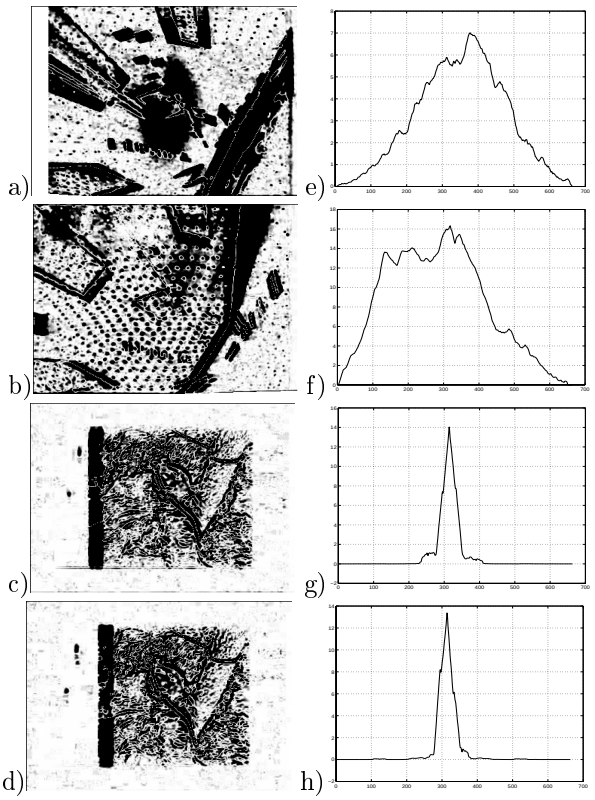

FIG. 4: Profil de la distribution des corrélations: (a,b) Cartes des outliers dans le cas de scènes complexes à l'intérieur d'un plan; (c,d) Cartes des outliers pendant une transition de type volet; (e,f,g,h) Corrélations "horizontales" associées $C_h^*(t,k)$. Les variances $m_{2h}(t)$ ont respectivement pour valeurs: 117.72, 129.84, 27.91, 26.99.

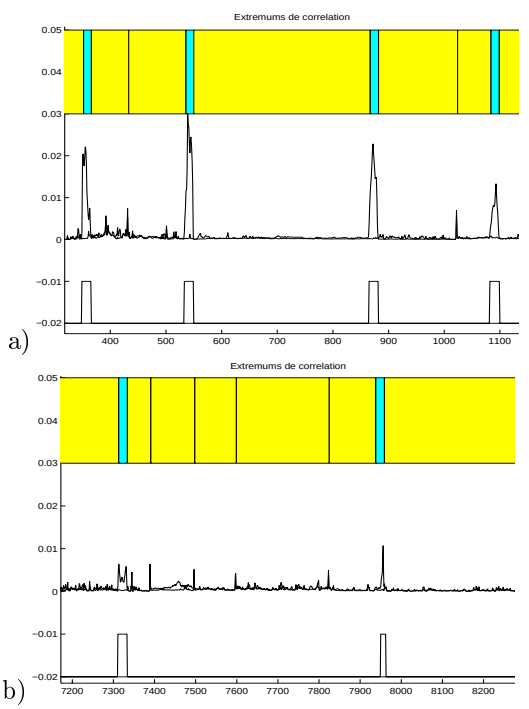

FIG. 5: Résultats de la détection de volets: (a,b) de haut en bas : Segmentation manuelle de référence fournie par l'INA (les séparations indiquées par un seul trait vertical correspondent à des cuts), Mesures , $_h(t)$, , $_v(t)$, Résultats de la détection après seuillage.