Suivi et indexation des objets dans des séquences vidéo avec la mise à jour par confirmation rétrograde

Amal MAHBOUBI[1] Jenny BENOIS-PINEAU[2] Dominique BARBA[1]

[1]IRCCYN / EPUN
rue Christian Pauc La chantrerie BP 50609
NANTES 44306
Tél. +33 2 40 68 32 36 Fax +33 2 40 68 32 32
amahboub@ireste.fr , dbarba@ireste.fr

[2] LABRI, CNRS UMR 5800
Université de Bordeaux-1
33405 TALENCE
Tél. +33 5 56 84 84 24 Fax +33 5 56 84 66 69
Jenny.benois@labri.fr

Résumé

Dans ce papier, nous proposons une approche pour le suivi temporel des séquences vidéo dans le cadre d'applications de type MPEG4. La stratégie que nous employons se base au départ sur une segmentation spatio-temporelle interactive de la scène en entités d'objets normalisées appelées VOPs (Video Object Plane). Le maintient d'un découpage cohérent de la vidéo au cours du temps est réalisé par la détection et l'indexation des nouvelles régions d'une part et d'autre part le suivi des régions persistantes (prédiction et affinage). Néanmoins les mouvements forts dans des séquences à taux d'échantillonnage faible ou encor un mouvement important de la caméra ont pour conséquence des erreurs de segmentation, des zones de découvrement du fond peuvent être assignées à de nouveaux VOPs (traînées d'objets en mouvement). C'est pourquoi une étape de confirmation sémantique s'impose. Cette sous-pape de sécurité est réalisée par une mise à jour en arrière des nouvelles régions (indifféremment dans la forme ou dans le fond) en fonction d'une évolution contrôlée du fond de départ dit *certain*.

1 Introduction

La stratégie que nous employons dans la construction de la segmentation spatio-temporelle est de type segmentation-estimation elle s'apparie avec les applications de type MPEG4, en effet après une segmentation spatiale de la scène en régions homogènes au sens de la couleur [1], l'utilisateur désigne les régions d'intérêts (VOPs) ce qui permet d'avoir une classification sémantique quasi certaine de la trame du début de la séquence vidéo. Cette classification est schématise par l'utilisation de trois classes FOND pour le fond FORM pour l'objet et INDEF pour les zones incertaines, plus le masque de l'utilisateur (sélection) est précis plus petite est cette zone d'incertitude INDEF. Une fois les régions spatiales classées en termes d'objets, une première estimation des paramètres de mouvements est effectuée sur la segmentation spatiale, nous pouvons dés lors construire notre représentation hiérarchique de la scène en effectuant une fusion basée sur la bonne compensation de mouvement dans chaque classe [2]. Le suivi consiste à localiser précisément les objets dans la vidéo à chaque instant, en s'appuyant sur les informations du passé (prédiction), sur celles du présent (ajustement de la projection) [3] et confirmation de la classification sémantique ce qui renverrait un reflet fidèle de l'évolution des VOPs au cours du temps.

2 Schéma du suivi

Le schéma proposé se base sur l'enchaînement des 7 étapes suivantes :

- la projection (prédiction) de la segmentation S^t dans le plan de l'image suivante I^{t+1}, ce qui a pour résultat la segmentation à l'instant suivant notée S^{t+1} ;
- l'ajustement spatial de cette première version de S^{t+1} ;
- traitement des zones d'occultation ;
- découpage homogène des régions dont l'erreur de compensation de mouvement est élevée ;
- estimation du mouvement ;
- gestion automatique de la sémantique : classification des nouvelles régions au sens « FOND », « FORM », « INDEF » ;
- fusion des nouvelles régions en respectant la cohérence sémantique suivant un critère de bonne compensation du mouvement.

Le suivi temporel de la segmentation spatio-temporelle permet la construction de la segmentation S^{t+1} à l'instant de temps courant connaissant la segmentation S^t de l'instant précédant et les images I^t et I^{t+1}.

Bien que les travaux traitant de la problématique du suivi temporel ne sont pas récents, ce sujet reste pour lors non résolu entièrement. La contribution de la présente étude est l'introduction de la gestion automatique du contenu (la prise en compte de la sémantique objet/fond dans la vidéo) au sein des étapes classiques d'un schéma de suivi en avant. Cet aspect 'sémantique' est le nouveau défi du suivi automatique. En effet les travaux de [4,5] se basent sur une méthode semi-automatique pour l'extraction des objets vidéo. Dans un premier temps ils font appel à l'opérateur humain pour valider la

segmentation de certaines trames de la séquence (les I-frame) et traitent le restant des trames (P-frame) de façon automatique, la décision d'affiliation de chaque région au VOP ou au fond est prise en fonction de la sémantique de l'image précédente.

Notre méthode de gestion automatique du contenu se décompose en trois phases :
1 Initialisation ;
2 Extraction des nouveaux objets et leurs indexation ;
3 Validation de l'étiquetage.

1 L'initialisation porte sur la première trame de la séquence. Sachant que les objets d'intérêt peuvent être partiellement statiques, l'interaction de l'utilisateur pour leurs extraction semble nécessaire. Affin de minimiser cette interaction, la tâche de l'utilisateur consiste à encercler grossièrement l'objet d'intérêt, le masque ainsi obtenu guide le processus spatio-temporel de fusion des régions basée mouvement.

2 Les nouvelles régions issues du traitement des zones d'occultation sont classifiées en fonction de la similarité de leur texture avec le voisinage spatial et de la disposition de leurs supports dans le passé.
Les nouvelles régions issues du découpage homogène au sens de bonne compensation du mouvement sont classées en fonction de l'évolution de leur mouvement. Ici on se base sur l'hypothèse que des nouveaux objets sont caractérisés par une forte valeur de l'activité différentielle. Ainsi les nouvelles régions sont classées en terme de FOND/FORM/INDEF, les régions FORM sont indexées à leur VOP.

3 La validation de la classification consiste à maintenir et affirmer la classification au cours du temps, lors du suivi temporel il s'agit de confirmer la segmentation courante par les éléments sûrs de son passé.

Le lecteur est invité à consulter [6] pour plus de détails sur les étapes 1 et 2, alors que l'étape 3 est décrite dans la section 3.

3 Confirmation sémantique

Notre méthode de gestion automatique du contenu sémantique de la vidéo se heurte à des erreurs d'estampillage dès que des mouvements brusques surviennent. C'est pourquoi nous introduisons une confirmation sémantique qui réside en la mise à jour des nouvelles étiquettes que ce soit dans la forme ou dans le fond. Cette démarche s'inspire des travaux de [7,8,9,10] sur la création d'images panoramiques (mosaïque). Ces travaux s'inscrivent dans le cadre des applications liées à la manipulation d'objets vidéo ou à la manipulation de scènes vidéo vu que la création de mosaïque permet de créer un résumé d'une séquence vidéo. Les travaux de [8] optimise le critère de création d'image panoramique dynamique, alors que [9] propose d'utiliser un support adapté mouvement pour s'affranchir des contraintes sur le mouvement de caméra. Pour assurer une meilleure construction de la mosaïque [10] propose de minimiser l'erreur d'alignement entre l'image mosaïque partielle et l'image courante à rajouter.
Comme l'étiquetage des régions dans l'image de départ est assisté par l'utilisateur, la zone de « FOND » est 'certaine'. Cette zone est le complémentaire des régions étiquetées comme « FORM » ou « INDEF ».

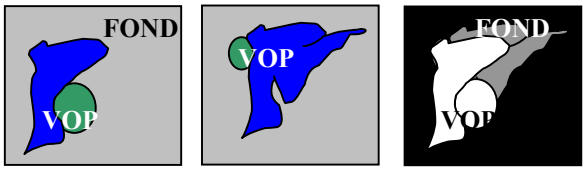
3.1 La projection du fond certain

Pour confirmer la segmentation à l'instant de temps courant le fond 'certain' est projeté et superposé sur la carte de segmentation courante. Cette confirmation est réalisée au cours du temps de façon périodique en supposant « certaine » la segmentation confirmée à l'instant de la confirmation précédente.
Soit *t0* l'instant de départ ou le fond est certain à l'instant de confirmation *t*, nous avons d'une part le fond issu du suivi et d'autre part le fond certain projeté à l'instant *t* noté *Fond_remonte*. Soit un point *i* de coordonnées (x_{i0}, y_{i0}) à l'instant *t0*, à l'instant de confirmation *t* ce point *i* a pour coordonnées (x_{ik}, y_{ik}) avec
$x_{ik} = x_{i0} + dx_{i1} + \ldots + dx_{ik-1}$ et
$y_{ik} = y_{i0} + dy_{i1} + \ldots + dy_{ik-1}$ ou les dx_{ij}, dy_{ij} sont les valeurs du vecteur de déplacement du point *i* considéré à l'instant *j*.
Ce vecteur de déplacement est choisi en fonction de la localisation spatiale du pixel *i* à l'instant *j* :
- si l'étiquette de la région à laquelle le pixel *i* appartient à l'instant *t* est FOND alors les calculs sont fait avec le vecteur de déplacement de cette région FOND (la zone en noir sur la Figure 1-c).
- Si l'étiquette de la région à laquelle le pixel *i* appartient à l'instant *t* est non FOND alors c'est le vecteur de déplacement de la région fond dominante à cet instant qui est pris en considération (la zone en gris sur la Figure 1-c).

Notons que les coordonnées dans le plan image sont des valeurs entières alors que nous manipulons des valeurs flottantes, pour s'affranchire des positions inter-pixelaires nous interpolons ces coordonnées aux entiers les plus proches.

Le *Fond_remonte* est la projection des coordonnées des pixels du fond 'certain' de départ à l'instant courant avec conservation de la luminance des pixels depuis l'image de départ.
La Figure 2 illustre le processus de construction de *Fond_remonte*. Dans la Fig.2-a nous pouvons

observer le Masque du fond 'certain' à l'instant de départ t=3, la Fig.2-b représente le masque du *Fond_remonte*, on peut observer la carte de segmentation dans la Fig-2-c (noire pour le FOND, blanc pour la FORM et gris pour INDEF).



-a- -b- -c-

Segmentation t_0 Segmentation t_n Fond_remonte

Figure 1 Projection du fond certain.

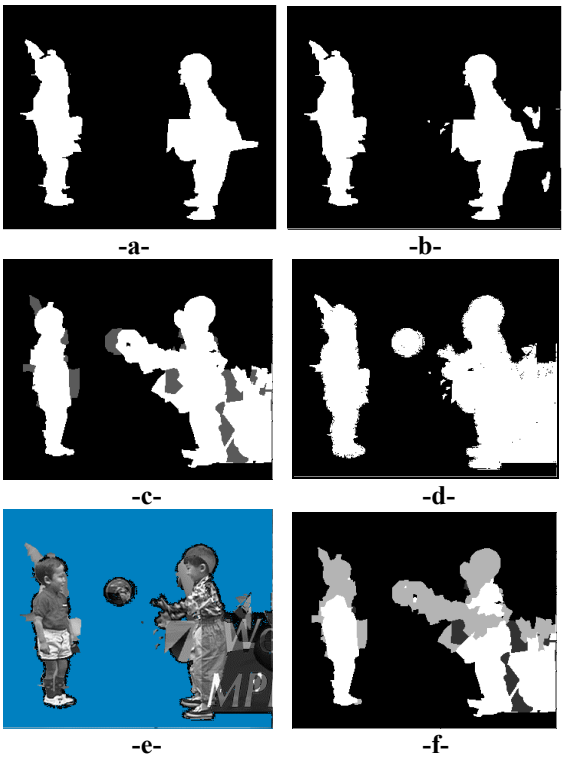


Figure 2 Exemple de projection du fond certain -séquence Children-

3.2 La validation de la segmentation

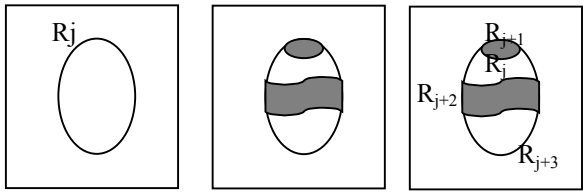
Une fois le *Fond_remonte* obtenu (Fig. 2-b), nous pouvons corriger les erreurs de segmentation survenu au cours du suivi (fausse fusion erreur d'étiquetage). Nous construisons l'image de DFD (Fig. 2-d et 2-e) qui est la différence de luminance entre l'image originale à l'instant *t* et la luminance de *Fond_remonte*. Les zones ou l'on observe une forte DFD (les zones en blanc dans Fig. 2-d) correspondent aux répartitions de l'objet alors que les zones de DFD faible (les zones en noire dans Fig. 2-d) représentent le fond.

Pour chaque région R_j^k nous vérifions s'il y a correspondance des classes dans les deux fonds, au quel cas aucune étape supplémentaire n'est à effectuer alors que nous mettons à contribution la confirmation de l'étiquetage par la mise à jour par confirmation rétrograde du fond de départ certain dans le cas d'étiquetage multiple : sur la Fig 2-f les zone en gris sont les zones dites ambiguës (étiquetage multiple).

Deux étapes sont à distinguer, la validation des VOPs et la validation du fond.

La validation dans l'objet

Les erreurs de segmentation dans l'objet sont essentiellement dues au fausses fusions, ou l'on se retrouve avec une partie du FOND étiqueté FORM. L'étiquetage multiple ici correspond à FORM dans la carte de segmentation et FOND par le biais de la DFD. Pour lever l'ambiguïté une re segmentation est effectué dans la région concerné, il en résulte de ce découpage sémantique l'éclatement de l'ancienne région FORM en deux partie FORM et FOND ou chacune d'elles contient une ou plusieurs régions suivant la répartition des pixels dans l'image DFD. La Figure 3 illustre le principe de ce procédé : en Fig. 3-a la région Rj issue du suivi, en Fig. 3-b l'étiquetage par la DFD les zones en gris sont les nouvelles régions à extraire de Rj, dans la figure 3-c le résultat de la correction dans l'objet. La Figure 4 présente l'effet de la validation dans l'objet sur la carte de segmentation.



-a- -b- -c-

Figure 3 Principe du découpage sémantique

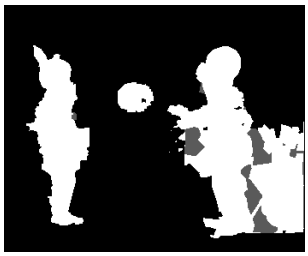
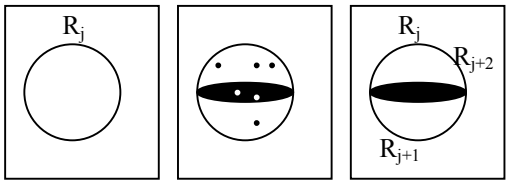


Figure 4 La carte des étiquettes après la validation VOP

La validation dans le fond

La validation dans le fond est utilisé pour remédié aux fausses affiliations de l'objet dans le fond. Ici la région ambiguë à pour étiquette FOND dans la carte de segmentation alors qu 'elle est FORM par la DFD. Le même procédé d'éclatement est suivi que lors de la segmentation dans l'objet (Figure 3).

Il est à noter qu'un raffinement est effectuer lors de la correction des étiquettes. En effet les pixels isolés et les petites régions issues de l 'éclatement d'une région ambiguë ne sont pas automatiquement affilié à la classe indiqué par la DFD mais à la région englobante. La Figure 5 illustre un exemple de ce filtrage.



-a- -b- -c-

Figure 5 Filtrage des nouvelles régions issues du découpage sémantique

4 Résultats

La méthode proposé est testée sur les séquences « Children » et « Akiyo » issues du corpus de test de MPEG4. Les premiers résultats sont encourageants (remonté du fond, détection des régions divergentes, correction dans l'objet). La figure 6 illustre un exemple de correction des erreurs de segmentation par mise à jour en arrière du fond de départ à l'instant de confirmation, jusqu'à présent nous nous somme pas encor confronte au correction de la segmentation dans le fond néanmoins les résultats prometteurs obtenu pour la correction dans l'objet devraient être atteint.

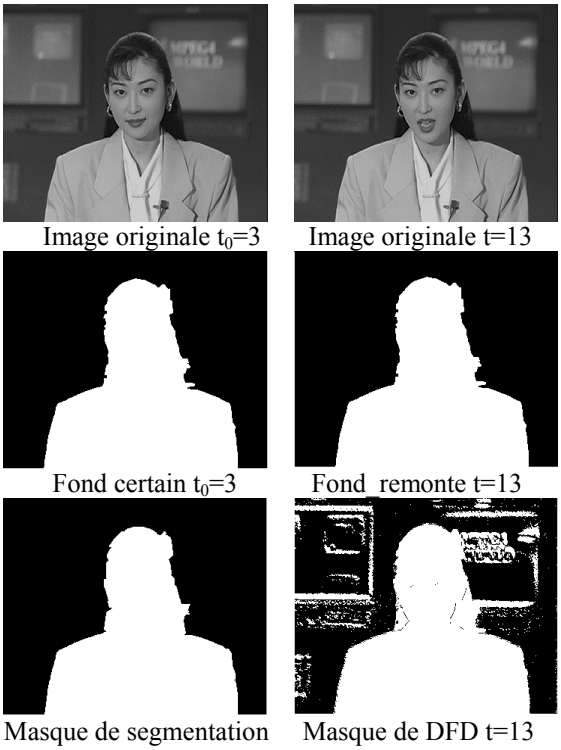


Image originale t_0=3 Image originale t=13

Fond certain t_0=3 Fond remonte t=13

Masque de segmentation Masque de DFD t=13

Figure 6 Effet de la validation par mise à jour arrière -séquence Akiyo-

5 Bibliographie

[1] A. Mahboubi, J. Benois-Pineau, D. Barba "Segmentation spatiale couleur des images par une approche morphologique et hierarchique", CORESA, Poitiers, France, 19-20 Oct 2000.

[2] J. Benois-Pineau, F. Morier, D. Barba, H. Sanson "Hierarchical segmentation of video sequences for content manipulation and adaptive coding", Signal Processing 66 pp. 181-201, 1998.

[3] L. Bonnaud, C. Labit, J. Konrad « Interpolative coding of image sequences using temporal linking of motion-based segmentation », Proc. IEEE Int. Conf. Acoustic Speech Signal Processing, Detroit, Michigan,USA, mai 1995.

[4] C. Toklu, A. Murat Tekalp, A. Tanju, « Simultaneous alpha map generation and 2-D mesh tracking for multimedia application» Proceeding of ICIP'97, , 1997, Vol. 1, pp 113-116.

[5] C.Gu, M-C Lee, "Semantic video object tracking using region-based classification", Proceedings of ICIP'98, Chicago, Illinois, USA, 4-7 octobre 1998, Vol. 2, pp 643-647.

[6] A. Mahboubi, J. Benois-Pineau, D. Barba " Tracking of objects in video scenes with Time varying content ",WIAMIS'2001, Tampere, Finland, 16-17 May 2001, pp. 101-105.

[7] H. Nicolas, « Mosaic reprentation and video object manipulations for post-production applications » Proceedings of ICIP'98, Chicago, Illinois, USA, 4-7 octobre 1998, Vol. 2, pp 451-455.

[8] H. Nicolas, « Critère optimal pour la création de séquences d'images panoramiques » Symposium of Gretsi'99, Vannes, France, 13-17 Septembre 1999, Vol. 2, pp 503-506.

[9] S. Peleg, B. Rousso, A. Rav-Acha, A. Zomet, « Mosaicing on Adaptive Manifolds » IEEE Transactions on pattern analysis and machine intelligence, vol.22, n°10, pp 1144-1154,Octobre 2000.

[10] H. Wallin , C. Christopoulos, A. Smolic, Y. Abdeljaoued, T. Ebrahimi, « Robust mosaic construction algorithm » ISO/IEC JTC1/SC29/WG11 MPEG00/M5698, Noordwijkerhout, The Netherlands, March 2000.

Ces travaux sont supportés par le projet RNRT OSIAM.