

# Extraction spatio-temporelle d'objets dans la vidéo HD dans le domaine des ondelettes sous le paradigme de l'indexation primaire

C. Morand<sup>1</sup>

J. Benois-Pineau<sup>1</sup>

J-Ph. Domenger<sup>1</sup>

<sup>1</sup> LaBRI, UMR CNRS/Université Bordeaux 1

351, Cours de la Libération  
33405 Talence Cedex

{morand, jenny.benois, domenger}@labri.fr

## Résumé

*Les nouveaux outils d'analyse et d'indexation basés contenu doivent maintenant tenir compte des caractéristiques de la vidéo Haute Définition. En particulier, les nouveaux standards de compression des images et vidéos HD ((M)JPEG2000) utilisent la transformée en ondelettes et les codeurs par ondelettes 3D sont largement étudiés. Dans le cadre du paradigme de l'indexation primaire que nous avons déjà proposé, nous présentons une technique d'extraction des objets en mouvement dans les séquences utilisant une segmentation spatio-temporelle à basse résolution. Une méthode de projection tenant compte de la structure de la transformée en ondelettes est développée ; elle permet d'obtenir le résultat de la segmentation précédente à pleine résolution tout en en corrigeant les défauts.*

## Mots clefs

Indexation primaire, ondelettes, segmentation spatio-temporelle, vidéo HD.

## 1 Introduction

L'extraction automatique d'objets d'intérêt est une étape essentielle de l'analyse, l'indexation mais aussi la compression des vidéos. Dans le cas de l'analyse et de l'indexation, il s'agit de décrire et d'organiser automatiquement le contenu [1]. Dans le cas de la compression, le but est de définir les zones à coder avec plus de précision, en conformité avec le standard MJPEG2000 [2] et son codage en régions d'intérêt [3].

Dans le cadre de nos travaux de recherche, nous avons adopté une stratégie que nous appelons "paradigme de l'indexation primaire". Il s'agit d'utiliser directement les informations contenues dans le flux compressé sans décodage complet pour effectuer l'extraction automatique des objets d'intérêt. Une telle approche a été développée dans [4] pour les flux MPEG2 (utilisation des vecteurs de mouvement et des coefficients DCT). Nous présentons ici une approche analogue pour les flux d'images animées encodées en MJPEG2000 [2].

Ainsi, nous proposons une segmentation spatio-temporelle

dans le domaine de la transformée en ondelettes discrète (TOD). Les deux aspects estimation de mouvement et segmentation dans le domaine de la TO ont été largement présentés dans la littérature. Ainsi [5, 6] proposent d'estimer le mouvement dans les séquences vidéo en utilisant des bases d'ondelettes adaptées. La segmentation couleur des images fixes dans le domaine des ondelettes est essentiellement focalisée sur la sous-bande LL [7, 8]. La particularité de notre approche consiste à proposer une solution contrainte par le paradigme de l'indexation primaire, à savoir utiliser la transformée adoptée par le codeur normalisé (base 9/7 de Daubechies), pour mettre en collaboration les processus d'indexation et de codage. Nous proposons d'exploiter toute la richesse de la représentation en ondelettes en utilisant non seulement la sous-bande LL à différentes résolutions mais aussi toutes les sous-bandes de haute fréquence afin d'aboutir à une segmentation fine et scalable des objets le long de la pyramide jusqu'à pleine résolution. Par conséquent, l'indexation des objets s'adaptera également à la scalabilité du flux et s'intégrera à la transmission scalable de la vidéo [9]. Par rapport au processus de codage proprement dit, son utilisation du côté codeur est immédiate. Etant intégrée à un codeur de type MJPEG2000 entre les étapes de décomposition en ondelettes et de quantification, elle permettra de définir les régions d'intérêt pour déployer le style de codage correspondant [3, 10].

Le papier est organisé comme suit. Dans la section 2 nous présentons le schéma global de l'extraction spatio-temporelle d'objet. La segmentation à basse résolution est décrite dans la section 3. La section 4 présente notre approche hiérarchique d'extraction d'objets. Les résultats expérimentaux sont montrés dans la section 5. Enfin la section 6 conclut notre travail et présente les perspectives de recherche.

## 2 Schéma général d'extraction hiérarchique d'objets dans le domaine des ondelettes

La figure 1 résume notre approche. Le flux d'entrée est un flux codé par MJPEG2000. Dans notre travail, ce codeur est modélisé par une décomposition en ondelettes sur  $K$  niveaux. Nous faisons de plus l'hypothèse que le flux MJPEG2000 traité ne contient pas d'information de type région d'intérêt. Deux segmentations, une en mouvement et l'autre en couleur, sont effectuées sur la séquence constituée des sous-bandes LL de basse résolution ( $k=K$ ), la segmentation couleur étant contrainte par le mouvement. Ceci constitue la segmentation spatio-temporelle à basse résolution. Ensuite, une étape de projection/ajustement est appliquée successivement sur tous les niveaux de la pyramide pour obtenir les masques des objets à haute résolution. Notre approche permet d'utiliser l'information contenue dans chacune des sous-bandes (LL, LH, HL et HH) dans des zones d'incertitude définies au voisinage des contours des objets.

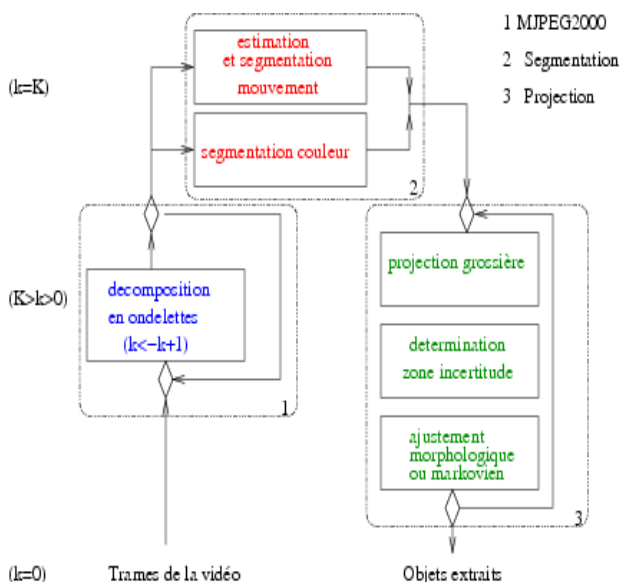


Figure 1 – Schéma général d'extraction d'objets

## 3 Segmentation spatio-temporelle à basse résolution

### 3.1 Segmentation en mouvement

Dans ce travail, nous avons utilisé les résultats de nos recherches antérieures, la méthode d'estimation dans le domaine des ondelettes étant en cours de développement. Le flux MJPEG2000 ne contient pas d'information de mouvement. Aussi, une étape préliminaire consiste à estimer les vecteurs de déplacement. Nous utilisons pour cela l'estimateur de mouvement basé bloc du MSSG (MPEG Software Simulation Group) et nous l'appliquons sur la séquence

constituée des sous-bandes LL de niveau  $K$  (basse résolution). Un exemple de vecteurs de mouvement obtenus est présenté figure 2. A partir de ces mesures et en supposant un modèle à 6 paramètres, le mouvement global de la caméra est estimé par l'estimateur robuste de Tukey [11]. Ceci nous permet d'avoir les valeurs optimales des paramètres grâce au rejet des mesures aberrantes et d'affecter un poids aux blocs en fonction de leur adéquation au modèle de la caméra. Le seuillage de ces poids permet de générer un masque grossier des objets en mouvement (Figure 2). Les poids isolés sont éliminés par filtrage. Pour plus de détails, on pourra se référer à [4].

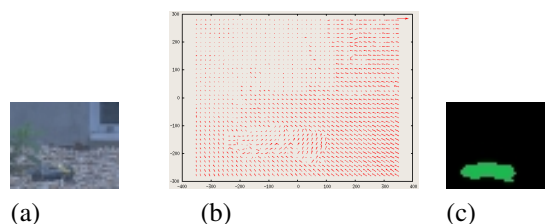


Figure 2 – (a) Image LL basse résolution ©LaBRI (b) Vecteurs de mouvement estimés (c) Masques de Mouvement

### 3.2 Segmentation couleur morphologique

La segmentation présentée dans [4], basée sur un algorithme de ligne de partage des eaux modifié, est appliquée à la sous-bande LL de plus faible résolution ( $k=K$ ); le masque de l'objet en mouvement délimite la surface dans le plan vidéo où elle est appliquée. Dans cette méthode, les pixels sont agglomérés suivant le critère  $C_1(x, y, R)$  donné par

$$C_1 : \|f_{LL}(x, y) - \bar{m}_R\|_{L_1} \leq 3F(\bar{m}_R)\Delta^{(i)} \quad (1)$$

$$F(\bar{m}) = \left| \frac{1}{3} \|\bar{m}\|_{L_1} - 127 \right| + 128 \text{ et } \Delta^{(i+1)} = \Delta^{(i)} + 0.01$$

où  $f_{LL}(s)$  est le vecteur couleur du pixel au site  $s$ ,  $\bar{m}_R$  le vecteur couleur moyen de la région  $R$  et où  $\|\cdot\|_{L_1}$  désigne la norme  $L_1$ . La fonction  $F$  suit le principe de "fonction de sensibilité visuelle simplifiée", qui montre que la différence entre deux niveaux de gris est moins perceptible aux extrémités de la dynamique.  $\Delta^{(i)}$  est un terme incrémental utilisé pour relâcher progressivement le seuil. L'algorithme s'arrête quand tous les pixels ont été affectés à une région.

## 4 Projection et Ajustement

### 4.1 Projection

La segmentation à basse résolution proposée donne déjà une première réponse sur la présence et la localisation de l'objet dans la vidéo. Nous allons maintenant propager cette information aux niveaux de résolution croissante. Pour cela, la segmentation calculée au niveau  $k$  est projetée sur le niveau  $k-1$  et affinée sur les bords des objets (zone d'incertitude) en tenant compte de l'information contenue

dans les sous-bandes du niveau k-1. La projection se déroule en trois temps.

D'abord, le résultat de la segmentation à l'apex de la pyramide (niveau k) est projeté de façon grossière sur le niveau de résolution immédiatement supérieure (niveau k-1) en utilisant le principe de localisation des ondelettes. Cette projection induit des effets de bloc et conduit à une mauvaise segmentation sur les bords des objets. Pour corriger ce défaut, la deuxième étape permet de définir une zone d'incertitude dans laquelle les pixels vont être réaffectés suivant un critère fin au niveau courant (k-1) de la pyramide.

La zone d'incertitude est définie par la différence entre la dilatation et l'érosion du masque grossier de l'objet (résultat de la projection brute). L'élément structurant, présenté dans la figure 3, est 4-connexe et dissymétrique. Cette dissymétrie permet de compenser celle engendrée par la projection brute.

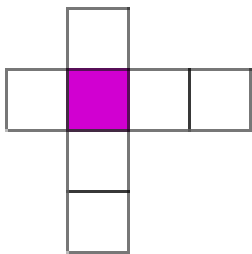


Figure 3 – Détermination de la zone d'incertitude : élément structurant

La troisième étape consiste à affecter chaque pixel de la zone d'incertitude à la région dont il est le plus proche au sens du critère d'attribution. Les critères d'attribution possibles sont définis dans les sections suivantes. La figure 4 récapitule les étapes de la projection.

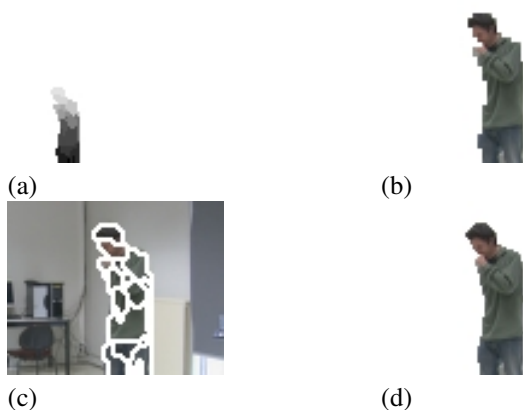


Figure 4 – Etapes de projection (a) masque de régions (niveau k) (b) Projection brute (niveau k-1) (c) Zone d'incertitude (niveau k-1) (d) Ajustement (niveau k-1)

## 4.2 Ajustement morphologique couleur et coefficients de Haute Fréquence

Nous définissons un nouveau critère  $C_2(x, y, R)$  inspiré du critère  $C_1$  (1) et tenant compte de l'information contenue dans les sous-bandes de haute fréquence. Il s'exprime comme suit :

$$C_2 : \|f_{LL}(s) - \bar{m}_R\|_{L1} + T_{HF} \leq 3F(\bar{m}_R)\Delta^{(i)} \quad (2)$$

$$\text{avec } T_{HF} = \sum_{hf=1}^3 \| |f_{hf}(s)| - \hat{f}_{hf}(R) \|_{L1}$$

$$\text{et } \hat{f}_{hf}(R) = \frac{1}{\text{Card}_R} \sum_{s_r \in R} |f_{hf}(s_r)|$$

$T_{HF}$  est un terme permettant de tenir compte de l'information contenue dans les sous-bandes de haute fréquence,  $hf$  étant un indice désignant une de ces sous-bandes ( $hf = 1$  : HL,  $hf = 2$  : LH et  $hf = 3$  : HH).

Du fait du sous-échantillonnage, seule la valeur absolue des coefficients hf est significative. Analysons le terme critère  $C_2$ . Si il n'y a pas de contour et que la région est plate, le terme  $T_{HF}$  est proche de 0, on se retrouve dans la situation du critère  $C_1$  (1). C'est aussi le cas si la région R est une zone texturée et que le pixel (x,y) appartient à cette texture. Dans ce cas,  $\hat{f}$  est une caractérisation simplifiée de la distribution de la texture. Si il y a un contour et que la région R est plate, alors  $T_{HF}$  exprime le contraste du contour. En relaxant le seuil dans l'équation (2), tous les pixels de la zone d'incertitude sont assignés progressivement aux régions avoisinantes au niveau k-1 de la pyramide. En utilisant un critère sur la haute fréquence (2), on ajoute une barrière dans la croissance de région qui améliore la définition des contours, spécialement à la frontière entre deux régions ayant des valeurs moyennes proches.

## 4.3 Ajustement et régularisation des contours des régions par modélisation markovienne

Dans cette partie, nous essayons d'exploiter l'information spécifique contenue dans les sous-bandes de haute fréquence, c'est-à-dire l'information sur les contours horizontaux (LH), verticaux (HL) et diagonaux (HH). Cette information est utilisée dans une régularisation markovienne [12]. Le problème posé est alors équivalent au problème de minimisation d'une somme S de potentiels en chaque pixel de la zone d'incertitude.

$$\min_x \{ S(s_{ZI}) = \sum_{i=1}^3 U_i(x, y; s_{ZI}) \} \quad (3)$$

où x est l'étiquette, y l'observation et  $s_{ZI}$  un site de la zone d'incertitude.

Le potentiel  $U_1$  est le terme d'attache aux données, il est lié aux valeurs de la couleur. Sous l'hypothèse que les couleurs suivent une loi gaussienne multi-variée dans le domaine YUV, on obtient la formulation classique :

$$U_1(x, y; s) = (y(s) - \nu_{x(s)})^t \Sigma_{x(s)}^{-1} (y(s) - \nu_{x(s)})$$

où  $y(s)$  est le vecteur de couleur au site  $s$ ,  $\nu_{x(s)}$  le vecteur couleur moyen de la région d'étiquette  $x(s)$  et  $\Sigma_{x(s)}$  la matrice de covariance de cette même région.

Le potentiel  $U_2$  est le potentiel de clique, il permet de privilégier certaines configurations de voisinage. C'est dans ce cadre que nous allons pouvoir utiliser la notion de direction contenue dans les sous-bandes de haute fréquence.

$$U_2(x; s) = \sum_{c \in C} (1 - \delta(x(s), x(c)) + (2\delta(x(s), x(c)) - 1) |HF|_n^c) \alpha$$

$$\text{avec } |HF|_n = \sqrt{\frac{|HF| - |HF_{min}|}{|HF_{max}| - |HF_{min}|}}$$

C désigne l'ensemble des cliques  $c$  de taille 2 dans le 8-voisinage et  $c$  une clique,  $x(s)$  est l'étiquette à trouver,  $x(c)$  l'étiquette du voisin dans la clique et  $\alpha$  est un facteur fixé par expérimentation.  $|HF|_n^c$  désigne la valeur normalisée du coefficient de haute fréquence associé à la clique  $c$ , c'est-à-dire que pour une clique horizontale (respectivement verticale, diagonale) nous utilisons le coefficient HL (respectivement LH, HH). Pour que le terme soit significatif dans le calcul du potentiel  $U_2$ , nous avons effectué un opération d'étalement de la valeur sur tout le segment  $[0, 1]$ . En plus de la régularisation classique (terme  $(1 - \delta(x(s), x(c))\alpha)$ , le terme  $((2\delta(x(s), x(c)) - 1) |HF|_n^c) \alpha$  pénalise les configurations de 2 étiquettes identiques et dé-pénalise les configurations de 2 étiquettes différentes si le coefficient HF indique la présence d'un contour.

Le potentiel  $U_3$  est le potentiel de création de région. Il permet de retrouver des régions dont une des dimensions est inférieure à  $2^K$ , c'est-à-dire des régions qui disparaissent à basse résolution (niveau  $k=K$  de la pyramide d'ondelettes). Pour cela une nouvelle étiquette  $e$  est considérée ( $e$  ne fait pas partie des étiquettes obtenues grâce à la segmentation basse résolution). On pose alors  $U_1(x, y; s) = 0$  si  $x(s) = e$ . Nous proposons un modèle affine par morceaux (Figure 5) pour définir le potentiel  $U_3$  :

$$U_3(x, y; s) = \begin{cases} a_1 U_{min} + b_1 & \text{si } x(s) \neq e \text{ et } U_{min} \leq U_{minlim} \\ a_2 U_{min} + b_2 & \text{si } x(s) \neq e \text{ et } U_{min} > U_{minlim} \\ a_3 U_{min} + b_3 & \text{si } x(s) = e \text{ et } U_{min} \leq U_{minlim} \\ a_4 U_{min} + b_4 & \text{si } x(s) = e \text{ et } U_{min} > U_{minlim} \end{cases}$$

$$\text{avec } U_{min} = \min_x \{U_1(x, y; s), x(s) \neq e\}$$

$U_{minlim}$  est un seuil réglé expérimentalement.

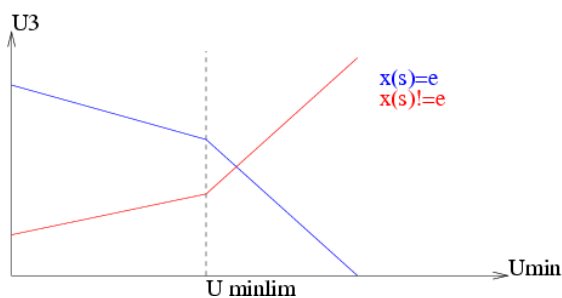


Figure 5 – potentiel  $U_3$

## 5 Résultats

Dans cette partie, nous illustrons les performances de la méthode proposée. Les images utilisées sont extraites du corpus HD (1920x1080 pixels, 25psf) produit au LaBRI. La Figure 5 résume les résultats à pleine résolution pour les différentes méthodes utilisées. Nous avons choisi d'utiliser  $K=4$ . D'une manière générale, l'objet extrait a des contours bien définis. Le défaut de segmentation est principalement situé au niveau des bras du personnage et apparaît dès la segmentation initiale (cercle rouge).

Les Figures 5 (c) et (d) illustrent la régularisation morphologique (paragraphe 4.2). (c) résulte de l'utilisation du critère C1 sur tous les niveaux alors que (d) utilise le critère C2. Nos expérimentations ont montré que le critère C2 est plus efficace aux niveaux intermédiaires, alors que C1 est appliqué au niveau basse résolution de la pyramide. A pleine résolution, il faut aussi appliquer le critère C1 puisque les coefficients des sous-bandes LH, HL et HH ne sont pas disponibles. L'utilisation des coefficients de haute fréquence permet de limiter l'expansion vers le fond ( (d) comparé à (c)).

Les Figures 6 (e) et (f) illustrent la régularisation markovienne (section 4.3) pour les potentiels  $U_1$  et  $U_2$ . (e) est le résultat obtenu sans le terme HF (eq 5). Là encore, l'expansion vers le fond limitée (le détail du carré bleu a été agrandi).

La figure (7) présente l'influence du potentiel de création de régions. Dans ce cas d'une image texturée complexe, la segmentation basse résolution n'a pas pu totalement séparer l'objet du fond. Notre potentiel de création de région a permis de retrouver de l'information perdue (cercle rouge). Cependant, un post-traitement est envisagé pour déterminer si ces nouvelles régions appartiennent ou non à l'objet en mouvement.

## 6 Conclusion

Ainsi, dans ce papier, nous avons proposé une segmentation spatio-temporelle sur la pyramide des ondelettes de Daubechies et scalable avec le flux compressé HD. Nous avons utilisé les coefficients d'ondelettes non quantifiés, l'étude de robustesse par rapport à la quantification étant dans la perspective de ce travail. Les résultats sont prometteurs car la qualité d'extraction de l'objet en pleine résolution est bonne malgré son initialisation très grossière à l'apex de la pyramide. Dans la perspective de ce travail, nous souhaitons adapter notre méthode à la résolution temporelle en inscrivant notre approche dans le cadre du codage en ondelettes 3D compensées en mouvement.

## 7 Remerciements

Ce travail est soutenu par le MENESR, l'ANR dans le cadre du projet ACI ICOS-HD et la région Aquitaine.

## Références

- [1] N. Dimitrova, H.-J. Zhang, B. Shahraray, I. Sezan, et A. Huang, T. and Zakhor. Applications of video content analysis and retrieval. *IEEE Multimedia*, pages 42–55, Juillet 2002.
- [2] ISO/IEC 15444-3 :2002. Technologies de l'information. système de codage d'images jpeg2000. partie 3 : Motionjpeg2000.
- [3] ISO/IEC 15444-1 :2002. Technologies de l'information. système de codage d'images jpeg2000. partie 1.
- [4] F. Manerba, J. Benois-Pineau, et R. Leonardi. Extraction of foreground objects from mpeg2 video stream in rough indexing framework. Dans *SPIE Proc. EI'2004, San José, CA, USA*, Janvier 2004.
- [5] B. Ugur Töreyn, A. Enis Cetin, A. Akscuy, et M. Bilgay Aklan. Moving object detection in wavelet compressed video. *Signal Processing : Image communication*, 20 :255–264, 2005.
- [6] C. Demonceaux et D. Kachi-Akkouche. Motion detection using wavelet analysis and hierarchical markov model. Dans *SCVMA04, Prague, république Tchèque*, Mai 2004.
- [7] C.R. Jung. Multiscale image segmentation using wavelets and watersheds. Dans *IEEE Proceedings of the XVI Brazilian symposium on computer graphics and image processing*, pages 278–283, 2003.
- [8] J.B. Kim et H.J. Kim. Multiresolution-based watersheds for efficient image segmentation. *Pattern recognition letters, Elsevier*, 24 :473–488, 2003.
- [9] C. Tillier, B. Pesquet-Popescu, et M. Van der Schaar. 3-band motion-compensated temporal structures for scalable video coding. *IEEE Transactions on Image Processing*, 15(9) :2545–2557, Septembre 2006.
- [10] C. Christopoulos, J. Askelöf, et M. Larsson. Efficient methods for encoding regions of interest in the upcoming jpeg2000 still image coding standard. *IEEE Signal Processing Letters*, 7(9) :247–249, Septembre 2000.
- [11] P. Kraemer et J. Benois-Pineau. Camera motion detection in the rough indexing paradigm. Dans *TREC Video Retrieval evaluation online proceedings (TRECVID'05)*, Novembre 2005.
- [12] S. Geman et S. Geman. Stochastic relaxation, gibbs distribution and the bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6 :721–741, Novembre 1984.

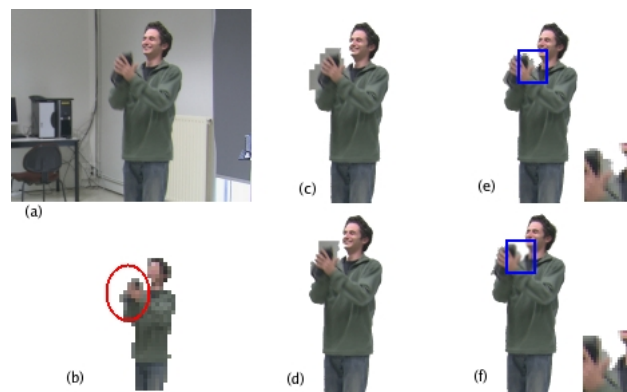


Figure 6 – (a) Image originale ©LaBRI (b) Résultat de la segmentation à basse résolution (grossi x3) (c) Segmentation morphologique LL (d) Segmentation morphologique HF (e) Régularisation markovienne classique (f) Régularisation markovienne HF



Figure 7 – (a) Image originale (b) Basse résolution (c,d) Segmentation obtenue (c) sans et (d) avec utilisation du potentiel  $U_3$  de création de région. Séquence extraite du corpus HD ©LaBRI