Estimation Robuste du Mouvement Global au sein de Séquences Vidéo Haute Définition après une Estimation du Mouvement basée Tubes Spatio-Temporels

O. Brouard[1] F. Delannay[1] V. Ricordel[1] D. Barba[1]

[1]Université de Nantes – Laboratoire IRCCyN – Équipe IVC

Polytech' Nantes, rue Christian Pauc, 44306 Nantes, France

{olivier.brouard, fabrice.delannay, vincent.ricordel, dominique.barba}@univ-nantes.fr

Résumé

Les méthodes d'estimation du mouvement global (EMG) sont efficaces pour la segmentation et/ou le suivi d'objets vidéo. En effet, les méthodes de segmentation d'objets basées sur les informations de mouvement nécessitent de connaître le mouvement global de la vidéo pour le compenser. Dans cet article, nous proposons une méthode pour estimer le mouvement global de séquences vidéo Haute Définition (HD). D'abord, nous développons une méthode adaptée d'estimation multi-résolution du mouvement (EMRM) pour obtenir un champ de vecteurs de mouvement cohérent. À partir de ces vecteurs de mouvement, nous estimons les paramètres de mouvement d'un modèle affine caractérisant le mouvement global du plan vidéo.

Mots clefs

Estimation du mouvement global, estimation multi-résolution du mouvement, tube spatio-temporel, vidéo HD.

1 Introduction

Le nouveau standard de codage vidéo H.264/MPEG-4 AVC [1] développé par le « Joint Video Team » de ISO/IEC MPEG et ITU-T « Video Coding Expert Group » vise à atteindre une réduction du débit de 50% pour une qualité équivalente comparée aux autres standards existants. Cette meilleure efficacité de compression est obtenue via un ensemble de spécifications relatives notamment aux multiples modes de prédiction, aux multiples images références et à une meilleure précision des vecteurs de mouvement. Cependant, cette réduction de débit est obtenue au prix d'une augmentation de la complexité et rend les applications temps réel délicates. On peut aussi constater les effets (papillonnement) du manque de cohérence au cours du temps du codage des objets de la vidéo.

Afin d'exploîter pleinement les spécifications offertes par le codeur H.264, une rapide pré-analyse des vidéos en amont du codeur serait judicieuse. Au prix d'un léger décalage temporel du codage, l'idée serait de fournir ensuite au codeur une information afin de réduire la combinatoire du choix des modes de prédiction, choisir les images références et assurer la cohérence temporelle du codage. En effet, une des stratégies possibles est d'utiliser des paramètres de codage adaptés à un objet vidéo donné tout au long de sa *durée de vie* le long du plan vidéo. Par objet vidéo, nous désignons une forme spatio-temporelle caractérisée par une couleur, une texture et un propre mouvement qui diffère du mouvement global du plan (c'est-à-dire, typiquement du mouvement dû au capteur). Dans ce papier nous nous intéressons, au niveau de la pré-analyse, à l'outil capable de détecter les objets de la vidéo.

Afin de pouvoir suivre des objets spatio-temporels dans une séquence vidéo, ils doivent être segmentés. Dans la littérature, plusieurs sortes de méthodes sont décrites. Ces différentes méthodes utilisent des informations spatiales et/ou temporelles [2, 3, 4] pour segmenter les objets. Dans le cas d'informations temporelles, il est nécessaire de connaître le mouvement global au sein du plan pour réaliser une segmentation correcte. Horn et Schunck [5] proposent de déterminer le flux optique entre deux images. Cela permet ainsi de connaître les objets en mouvement. Les paramètres du modèle de mouvement entre deux images successives peuvent également être estimés [6]. Une fois le modèle de mouvement connu, le mouvement global est compensé et seuls les objets « réellement » en mouvement subsistent avec l'information de leur mouvement.

Pour notre méthode, nous utilisons une information de mouvement par macrobloc pour un groupe d'images (GdI) pour estimer le mouvement global. Cette information de mouvement doit refléter autant que possible le *mouvement réel* (c'est-à-dire, obtenir les vecteurs de mouvement correspondants aux mouvements réels des objets de la scène) du macrobloc. Pour cela, nous développons une méthode adaptée d'estimation du mouvement qui utilise plusieurs images références pour estimer le mouvement du bloc courant, avec l'hypothèse d'un mouvement uniforme le long du GdI visant à lisser les vecteurs de mouvement obtenus. La seconde étape de notre méthode est l'EMG. Nous utilisons un modèle affine et l'estimation robuste des paramètres du modèle est réalisée à partir de ces vecteurs.

Dans la partie suivante, nous présentons notre méthode d'EMRM. Dans la partie 3, nous décrivons le calcul des paramètres du mouvement global. Finalement, nous montrons les résultats obtenus dans la partie 4 et concluons.

2 Estimation multi-résolution du mouvement

Pour obtenir une information de mouvement plus corrélée avec le mouvement réel des objets de la séquence vidéo, nous utilisons plusieurs images de référence et considérons un mouvement uniforme entre les images. Le temps de fixation du système visuel humain est d'environ $200ms$ [7] et comme la prochaine génération de télévision HD utilisera une définition de 1920×1080 pixels avec un taux de cinquante images par seconde, nous retenons un GdI constitué de neuf images ($180ms$). Ainsi nous assurons la cohérence du mouvement au sein du GdI sur une durée significative perceptuellement. L'image courante est située au centre du GdI. Ainsi, quatre images passées et quatres images futures entourent celle-ci. Nous utilisons les informations de ces neuf images pour contraindre l'estimation de mouvement et obtenir des vecteurs de mouvement plus lisses. En pratique, nous retenons cinq images pour évaluer les vecteurs de mouvement, comme cela est illustré à la figure 1. Nous considérons un mouvement uniforme, ainsi apparaît la notion de tube entre les images. Un tube suit et aligne un macrobloc donné sur plusieurs images. Le champ de vecteurs de mouvement contraint ainsi obtenu est plus homogène et donc plus corrélé avec le mouvement réel.

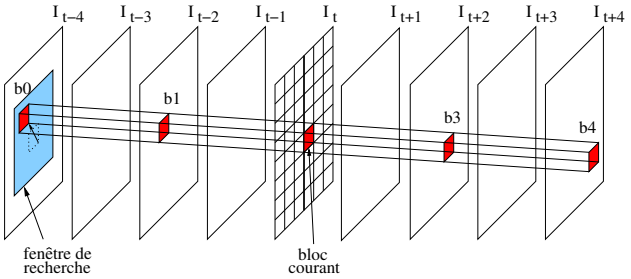


Figure 1 – *Tube spatio-temporel.*

Comme nous utilisons des images HD, nous proposons une EMRM [8] afin d'accélérer celle-ci. Les images HD sont filtrées spatiallement et sous-échantillonnées par un facteur six (d'abord, les images sont sous-échantillonnées par un facteur deux et ensuite par un facteur trois). Avant chaque étape de sous-échantillonnage, un filtre passe-bas adéquat est appliqué afin d'éviter tout recouvrement de spectres.
À partir des images filtrées et sous-échantillonnées, nous réalisons l'estimation du mouvement. Chaque bloc est simultanément comparé aux blocs potentiellement correspondants des images précédentes et futures, comme cela est illustré à la figure 1. L'erreur globale, EQM_G, est obtenue par la somme des quatre erreurs quadratiques moyennes (EQM) chacune entre le bloc courant et ses blocs correspondants des images passées et futures (voir éq. 1).

$$EQM_G = \sum_k EQM_k, \quad k = -4, -2, +2, +4. \qquad (1)$$

EQM_k prend en compte les trois composantes YUV de chaque bloc (voir éq. 2),

$$EQM_k = \frac{\sum_{i,j=0}^{N-1}(C_Y(i,j)-Rk_Y(i+\lambda_k.m,j+\lambda_k.n))^2}{N\times N}+ \frac{\sum_{i,j=0}^{N-1}(C_U(i,j)-Rk_U(i+\lambda_k.m,j+\lambda_k.n))^2}{N\times N}+ \frac{\sum_{i,j=0}^{N-1}(C_V(i,j)-Rk_V(i+\lambda_k.m,j+\lambda_k.n))^2}{N\times N}, \qquad (2)$$

avec $\lambda_{-4} = 4$, $\lambda_{-2} = 2$, $\lambda_2 = -2$ et $\lambda_4 = -4$. (m, n) est le vecteur de mouvement entre l'image courante I_t et l'image précédente I_{t-1}. C_Y, C_U, C_V, Rk_Y, Rk_U et Rk_V représentent respectivement les trois composantes YUV de l'image courante et celles de l'image utilisée comme référence pour l'estimation de mouvement, avec des blocs de taille $N \times N$ (typiquement 16×16). Le vecteur de mouvement retenu est celui qui minimise EQM_G entre le bloc courant et les blocs correspondants du tube des quatre images. Les vecteurs de mouvement sont estimés à la plus faible résolution, ensuite ils sont multipliés par un facteur approprié à la résolution supérieure afin d'être utilisés comme point de recherche initial pour la recherche du vecteur de mouvement à la résolution donnée.

3 Estimation du mouvement global

Précédemment nous avons réalisé l'estimation du mouvement, les tubes obtenus reflètent plus fidèlement le mouvement réel des objets. La prochaine étape est d'estimer les paramètres du modèle du mouvement global du plan vidéo à partir du champ de vecteurs de mouvement obtenu (un vecteur par tube).
Plusieurs modèles existent pour représenter le mouvement global d'un plan vidéo. Un champ de vecteurs de mouvement par bloc est moins dense et moins précis que des informations de mouvement basées pixel. Dans ce contexte il n'est pas nécessaire d'utiliser un modèle trop complexe, nous utilisons donc un modèle affine à six paramètres pour représenter le mouvement global (voir éq. 3) :

$$\begin{pmatrix} V_x \\ V_y \end{pmatrix} = \begin{pmatrix} a_1 & a_2 \\ a_3 & a_4 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} + \begin{pmatrix} t_x \\ t_y \end{pmatrix}, \qquad (3)$$

où a_i $(i = 1...4)$, t_x et t_y sont respectivement les paramètres de déformation et de translation. V_x et V_y sont les composantes horizontale et verticale rapportées pour chaque bloc, (x, y) étant la position du bloc. Pour estimer les paramètres du modèle de mouvement, nous adaptons la méthode d'accumulation des vecteurs de mouvement décrite par Coudray [9]. Ce dernier calcule plusieurs histogrammes afin d'obtenir les paramètres du mouvement global. Cette méthode repose sur le cumul de dérivées orientées (gradients) des vecteurs de mouvement. Ce cumul est réalisé au sein d'histogrammes qui permettent ainsi d'extraire les paramètres relatifs au déplacement majoritaire. Le mode majoritaire de chaque histogramme représente la valeur du paramètre global. Les vecteurs de mouvement sont d'abord compensés en utilisant les quatre

paramètres de déformation, la dernière étape est l'accumulation des vecteurs de mouvement ainsi compensés dans un dernier histogramme à deux dimensions. Les paramètres de translation du modèle sont alors identifiés à partir de celui-ci (les calculs seront détaillés dans la partie 3.2).

3.1 Indices de confiance

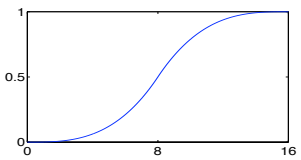
Pour un macrobloc donné, le vecteur de mouvement obtenu à partir de l'estimation de mouvement basée tube minimise l'EQM. Si le macrobloc courant est situé dans une région uniforme, l'appareillement de blocs n'est pas fiable et les macroblocs situés dans le tube qui minimisent l'EQM, ne sont pas nécessairement les macroblocs qui appartiennent à l'objet « réel ». Ainsi, le vecteur de mouvement associé au tube ne reflète pas le mouvement réel du bloc. Pour être robustes, de tels vecteurs de mouvement ne doivent pas contribuer à l'EMG de la vidéo. La contribution des vecteurs de mouvement doit donc être pondérée en fonction du contenu spatial des macroblocs du tube. Les vecteurs de mouvement de macroblocs relatifs à des zones orientées donnent eux des informations plus fiables sur le mouvement réel que ceux associés à des zones uniformes.
D'où l'idée d'utiliser l'activité spatiale du tube pour le qualifier. Pour calculer l'activité spatiale des macroblocs, nous utilisons les gradients spatiaux. Plus le gradient spatial d'un macrobloc est élevé, plus nous pouvons donner du poids au vecteur de mouvement de ce macrobloc.
Nous utilisons les deux gradients spatiaux moyens, $\overline{\Delta V}$ et $\overline{\Delta H}$, qui sont respectivement le gradient vertical moyen et le gradient horizontal moyen. À partir du calcul de ces gradients, un macrobloc peut-être identifié comme étant d'une zone uniforme, d'une zone moyennement texturée ou d'une zone fortement texturée.
Un macrobloc identifié comme étant d'une zone fortement texturée, peut l'être seulement dans une seule direction, c'est-à-dire, que l'un des gradients est élevé et l'autre faible. Si le mouvement global est une translation dans la même direction (verticale ou horizontale) que la zone texturée, les vecteurs de mouvement des macroblocs situés dans cette région ne sont pas fiables. C'est pourquoi nous distinguons les deux gradients spatiaux. En pratique, la confiance accordée aux deux composantes des vecteurs de mouvement est calculée de façon appropriée en fonction du gradient spatial $\Psi(\overline{\Delta H})$ et $\Psi(\overline{\Delta V})$. La fonction Ψ pour obtenir les indices de confiance fonction des gradients spatiaux s'écrit de la façon suivante :

$$\Psi(x) = \begin{cases} (\frac{x}{8})^3/2, & x \leq 8 \\ 1 - \Psi(16 - x), & 8 < x < 16 \\ 1 \ \ sinon \end{cases} \quad (4)$$

La figure 2 donne les indices de confiance obtenus pour une image de la séquence vidéo HD *Shields*. On remarque que plus le gradient spatial est élevé, plus l'indice est proche de 1 et au contraire pour un gradient faible, il est proche de 0.

3.2 Estimation robuste des paramètres du mouvement global

L'information de base utilisée pour estimer le mouvement global est un champ de vecteurs de mouvement (un vecteur par tube). Nous estimons les paramètres du modèle à partir du champ de vecteurs de mouvement de la manière suivante :



(a) Fonction de confiance.


(b) Image de la séquence vidéo *Shields*.


(c) Indices pour les gradients horizontaux (blanc = 1, noir = 0).


(d) Indices pour les gradients verticaux (blanc = 1, noir = 0).

Figure 2 – *Indices de confiance des vecteurs de mouvement.*

$$\begin{array}{c} a_1 = \frac{\partial V_x}{\partial x}, \ \ a_4 = \frac{\partial V_y}{\partial y}, \ \ a_2 = \frac{\partial V_x}{\partial y}, \ \ a_3 = \frac{\partial V_y}{\partial x}, \\ t_x = V_x - a_1.x - a_2.y, \ \ t_y = V_y - a_3.x - a_4.y. \end{array} \quad (5)$$

Le mouvement global est souvent celui de la caméra qui peut être relativement complexe. Étant donné qu'un zoom ou une rotation affectent l'estimation des paramètres de translation, l'EMG est réalisée en deux étapes. D'abord, nous calculons les paramètres de déformation pour chaque vecteur de mouvement. Chaque dérivée calculée donne une « hypothèse » unitaire pour l'un des paramètres du mouvement. Pour connaître l'hypothèse majoritaire qui a la plus grande probabilité de représenter le paramètre du mouvement global, les hypothèses unitaires sont toutes accumulées dans un histogramme. Chacune d'entre elles contribue proportionnellement en fonction de l'indice de confiance associé aux vecteurs de mouvement. Les données sont accumulées dans l'histogramme, pondérées par une distribution gaussienne (voir éq. 6) afin de regouper les hypothèses proches (typiquement $\sigma = 3$).

$$G(x) = \sqrt{\frac{1}{2\pi\sigma^2}} e^{-x^2/2\sigma^2}. \quad (6)$$

La localisation du mode majoritaire de cet histogramme indique la valeur majoritaire retenue pour le paramètre du mouvement global. Pour affiner la localisation du maximum, une méthode des moindres carrés est utilisée pour estimer la pente autour de la position du maximum. Une fois les paramètres de déformation identifiés, les vecteurs de mouvement sont compensés et les résidus ne correspondent alors qu'aux mouvements de translation. Ces derniers sont accumulés dans un histogramme à deux dimensions en utilisant une distribution gaussienne (voir éq. 7). Chaque distribution gaussienne est pondérée par le minimum des indices de confiance ($\Psi(\overline{\Delta H})$ et $\Psi(\overline{\Delta V})$).

$$min(\Psi(x), \Psi(y)).G(x,y) \ \ avec \ \ G(x) = \frac{1}{2\pi\sigma^2} e^{-\frac{x^2+y^2}{2\sigma^2}}, \quad (7)$$

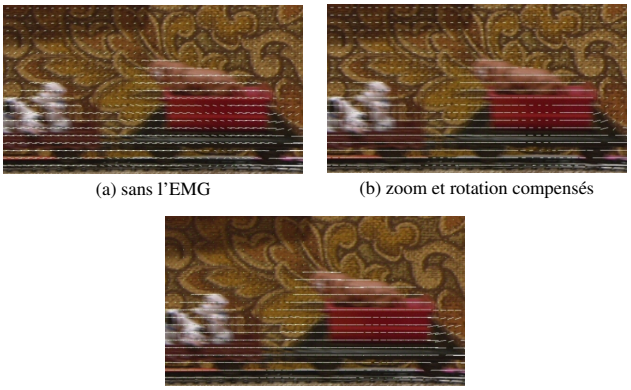
où x et y représentent respectivement les composantes compensées des vecteurs de mouvement. Les valeurs des paramètres de translation t_x et t_y sont obtenues en localisant le mode majoritaire dans l'histogramme à deux dimensions. À ce niveau, nous disposons des paramètres du modèle de mouvement global caractéristique du GdI.

4 Résultats expérimentaux

Nous avons utilisés une séquence HD 1080p (*Blue sky*) et deux séquences HD 720p (*Shields* et *New mobil calendar*) du SVT « corporate development technology » [10]. Ces vidéos présentent différents mouvements de caméra :

– *Blue sky* : rotation et translation ;
– *Shields* : translation horizontale et zoom avant ;
– *New mobil calendar* : translations verticale et horizontale et zoom arrière.

La figure 3 montre les vecteurs de mouvement après les différentes étapes de l'EMG. La première image (a) illustre les vecteurs de mouvement bruts obtenus par notre méthode d'EMRM. Les deux dernières images (b) et (c) contiennent les vecteurs de mouvement compensés après les deux étapes de l'EMG. Nous pouvons voir que le mouvement global est correctement compensé puisque les vecteurs de mouvement du fond sont presque tous nuls, seuls les vecteurs de mouvement du train se déplaçant subsistent et reflètent son mouvement de translation horizontale. Afin



(a) sans l'EMG (b) zoom et rotation compensés

(c) mouvement global compensé

Figure 3 – *Vecteurs de mouvement de la sequence vidéo New mobil calendar.*

de comparer les résultats numériques de notre méthode, nous le faisons avec les résultats obtenus avec le logiciel robuste Motion2D [11]. Motion2D procède à partir d'un champ de vecteurs de mouvement dense et précis et opère une estimation robuste des paramètres de l'EMG (c'est actuelleemnt la méthode de référence pour l'EMG). Comme Motion2D calcule les paramètres du mouvement global entre deux images consécutives, nous combinons les paramètres obtenus pour neuf images consécutives pour les comparer à ceux obtenus avec notre estimateur (qui travaille avec un GdI de neuf images). Les figures 4 et 5 illustrent les paramètres de déformation pour deux vidéos synthétiques contenant seulement une rotation ou un zoom, donc nous disposons des paramètres effectifs globaux utilisés. Nous pouvons voir que notre méthode d'EMG se comporte correctement, mais les résultats obtenus avec Motion2D sont plus proches des paramètres de déformation effectifs utilisés pour créer les séquences vidéo. En effet, Le logiciel Motion2D estime les paramètres du modèle de mouvement global pour une restriction temporelle et spatiale plus fines (estimation du mouvement basée pixel entre deux images successives) et ainsi, ceux-ci sont plus précis. Dans les graphiques suivants (figures 6, 7 et 8), nous comparons les résultats obtenus sur séquences réelles avec notre méthode (•) et Motion2D (+). Sur l'axe des abscisses est représenté le numéro du GdI (un indice de GdI est égal à neuf images).

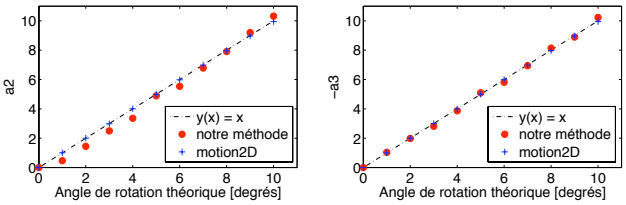


Figure 4 – *Paramètres globaux de déformation pour une rotation synthétique.*

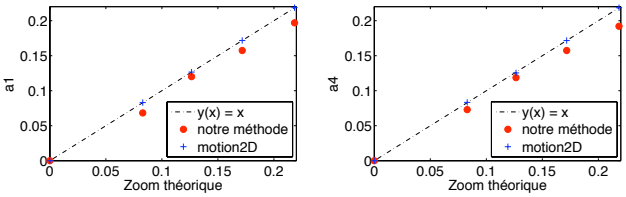


Figure 5 – *Paramètres globaux de déformation pour un zoom synthétique.*

Nous présentons ici seulement les résultats significatifs, les paramètres globaux estimés proches de zéro via les deux méthodes ne sont pas illustrés. Pour la séquence vidéo *Blue sky* qui ne contient qu'une rotation, les paramètres sont estimés correctement par notre méthode et Motion2D. Dans la séquence vidéo *Shields*, un homme est présent dans la scène et se déplace. Malgré ce mouvement local, les paramètres sont correctement estimés par les deux méthodes. Pour la dernière séquence, *New mobil calendar*, les résultats obtenus diffèrent entre les deux méthodes. En effet, notre méthode obtient de meilleurs résultats que Motion2D. Les mouvements translationnels du train et du calendrier sont estimés comme une rotation globale de toute la scène par le logiciel Motion2D. Cette estimation plus robuste est due à notre estimation du mouvement initiale (EMRM) plus cohérente.

5 Conclusion

Dans cet article, nous avons présenté une méthode robuste d'estimation du mouvement global (EMG) pour

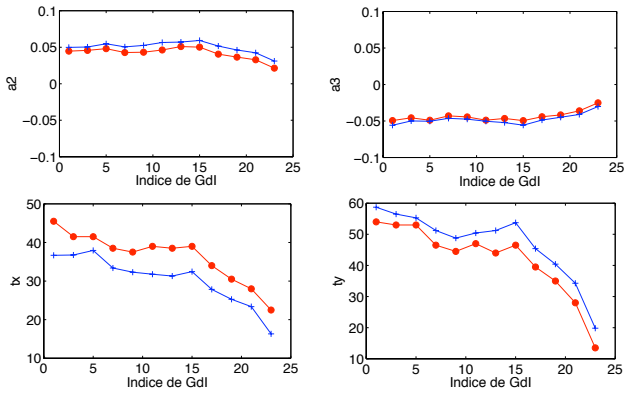


Figure 6 – *Paramètres estimés du mouvement global pour la séquence Blue sky (Motion2D (+), notre méthode (•)).*

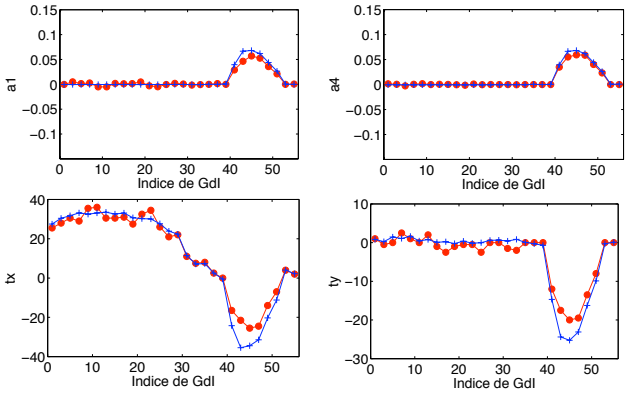


Figure 7 – *Paramètres estimés du mouvement global pour la séquence Shields (Motion2D (+), notre méthode (•)).*

des vidéos HD. D'abord, une estimation multi-résolution du mouvement utilisant plusieurs images références est réalisée, afin d'obtenir un champ de vecteurs de mouvement cohérent (un vecteur par tube spatio-temporel) pour le groupe d'images (GdI). Le GdI contient en effet neuf images, avec l'hypothèse que le mouvement d'un objet est uniforme le long de celui-ci, ceci afin de lisser le champ de vecteurs de mouvement obtenu. Ensuite, l'EMG est robustifiée car ces vecteurs de mouvement sont pondérés en fonction de l'activité spatiale du tube. Nous estimons exactement les paramètres du mouvement global d'un modèle affine à partir de ces vecteurs de mouvement pondérés. Ainsi, nous obtenons les paramètres du mouvement global de la vidéo. Les résultats obtenus montrent que notre méthode d'EMG est plus robuste que le logiciel Motion2D lorsque le plan vidéo contient des objets importants qui ont des mouvements différents de celui de la caméra (mouvement global).

6 Remerciements

Cette recherche a été effectuée dans le cadre du projet ArchiPEG financé par l'ANR (convention N°ANR05RIAM01401).

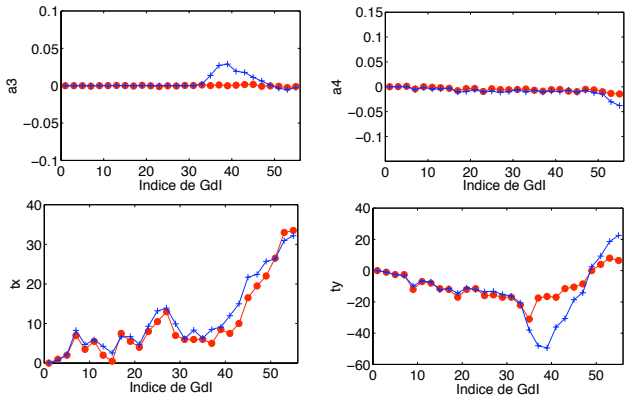


Figure 8 – *Paramètres estimés du mouvement global pour la séquence New mobil calendar (Motion2D (+), notre méthode (•)).*

Références

[1] I. E. G. Richardson. *H.264 and MPEG-4 video compression : Video Coding for Next-Generation Multimedia.* Chippenham, Septembre 2003.

[2] Remi Megret et Daniel DeMenthon. A survey of spatio-temporal grouping techniques. Rapport technique, Research report CS-TR-4403, LAMP, University of Maryland, USA, Octobre 2002.

[3] Fatih Porikli et Yao Wang. Automatic video object segmentation using volume growing and hierarchical clustering. *EURASIP Journal on Applied Signal Processing, ISSN*, 3 :442 – 453, Mars 2004.

[4] Y. Wang, J. F. Doherty, et R. E. Van Dyck. Moving object tracking in video. Washington DC, USA, Octobre 2000. IEEE Applied Imagery Pattern Recognition Workshop.

[5] B.K.P. Horn et B.G. Schunck. Determining optical flow. *Artificial Intelligence*, 17(1 – 3) :185 – 203, Août 1981.

[6] J. M. Odobez et P. Bouthemy. Robust multiresolution estimation of parametric motion models. *Journal of Visual Communication and Image Representation*, 6 :348–365, Décembre 1995.

[7] O. Le Meur, P. Le Callet, et D. Barba. A coherent computational approach to model bottom-up visual attention. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(5) :802 – 817, Mai 2006.

[8] S. Péchard, P. Le Callet, M. Carnec, et D. Barba. A new methodology to estimate the impact of H.264 artefacts on subjective video quality. Dans *Proceedings of the Third International Workshop on Video Processing and Quality Metrics for Consumer Electronics, VPQM*, Scottsdale, 2007.

[9] R. Coudray et B. Besserer. Global motion estimation for mpeg-encoded streams. Singapore, Republic of Singapore, Octobre 2004. IEEE International Conference on Image Processing, ICIP 2004.

[10] SVT. Overall-quality assessment when targeting wide xga flat panel displays. Rapport technique, SVT corporate development technology, 2002. ftp ://ftp.ldv.e-technik.tu-muenchen.de/pub/test_sequences/.

[11] IRISA. Motion2D. Rapport technique. http ://www.irisa.fr/Vista/Motion2D/.