Estimation conjointe disparité mouvement pour le codage de séquences vidéo multi-vues

Wided MILED , Ismaël DARIBO , Béatrice PESQUET-POPESCU

TELECOM ParisTech, Département Traitement du Signal et des Images
46 rue Barrault, 75634 Paris Cédex 13, France
{miled,daribo,pesquet}@telecom-paristech.fr

Résumé – Cet article aborde le problème du codage de séquences vidéo multi-vues et présente une nouvelle méthode d'estimation des champs de disparité et du mouvement impliqués dans la séquence. Afin de réduire la complexité et le coût de calcul et améliorer les performances d'estimation de ces champs, une nouvelle technique d'estimation conjointe est adoptée et appréhendée dans le cadre de méthodes variationnelles, en mettant en place un algorithme itératif, utilisant des outils récents d'analyse convexe. Les expérimentations sur des séquences réelles montrent les performances de cette technique aussi bien en termes de reconstruction qu'en termes de cohérence des champs de déplacement estimés.

Abstract – This paper deals with the problem of multiview video coding and describes a new method for recovering disparity and motion fields involved in the video sequence. In order to reduce computational complexity and improve estimation accuracy, a joint estimation technique is proposed and addressed in a variational framework, by using an iterative algorithm based on recently developed convex analysis tools. Experimental results involving real sequences indicate the feasibility and robustness of our approach both in terms of reconstruction and consistency of estimated displacement fields.

1 Introduction

Après la haute définition (HD), la prochaine révolution de la télévision sera d'intégrer la perception de la vidéo en trois dimensions. Cette nouvelle génération de téléviseurs permettra d'apprécier des films en relief sans le port de lunettes adaptées ou de naviguer librement dans un évènement sportif. D'autres domaines d'application sont aussi envisageables, tels que le cinéma numérique, la médecine, le trafic aérien, les technologies militaires, les jeux vidéos etc. Le développement de la télévision en relief (TV3D) et des supports autostéréoscopiques a suscité de nombreuses études portant sur le traitement et la compression de séquences vidéos multi-capteurs. Il s'agit, en particulier, d'extraire de l'information 3D à partir des vues disponibles, sous forme de cartes de profondeurs ou de disparités, afin de générer des vues virtuelles intermédiaires ou de compresser les séquences d'une manière plus efficace, permettant une transmission compacte et compatible avec les équipements existants.

Une séquence vidéo multi-vues, constituée d'images captées par plusieurs caméras synchrones, montre l'évolution temporelle d'une scène 3D à partir de points de vue distincts. Le codage d'une telle séquence entraîne une quantité importante de données à traiter proportionnelle aux nombre de vues utilisées. Un moyen efficace d'améliorer les performances d'un codage vidéo multi-vues est de prendre en compte, en plus de la corrélation temporelle, la corrélation inter-vues entre les caméras, tout en s'appuyant sur les liens spatio-temporels qui existent entre les différentes images de la séquence. Ces liens se résument à un champ de déplacement inter-vues (disparité) et un champ de déplacement temporel (mouvement apparent). Une extension du codeur H.264/AVC dans le cas du codage vidéo multi-vues (MVC, Multiview Video Coding), prenant en compte les redondances entre les vues, a été développée dans [1]. En comparaison avec un codage Simulcast, où chaque vue est codée indépendamment avec l'encodeur H.264/AVC, cette extension apporte un gain de compression significatif [1], [2].

Contrairement aux autres standards vidéo, H.264/AVC considère dans son estimation temporelle plusieurs trames de référence dans le processus d'encodage d'une trame. Pour choisir la meilleure trame de référence pour chaque macrobloc (MB), une optimisation débit-distortion basée sur une fonction de coût Lagrangienne est utilisée. En MVC cette spécificité est mise à profit, en insérant parmi les trames de référence, les trames provenant des vues voisines. Ainsi, une mise en compétition de codage par blocs sera utilisée, consistant à choisir de manière adaptative pour chaque MB un codage *intra* ou un codage *inter* par compensation de mouvement ou *inter* par compensation de disparité. Par ailleurs, les méthodes proposées estiment les champs de déplacements (mouvements apparents et disparité) séparément et ne prennent pas en considération les contraintes qui existent entre le mouvement et la disparité dans les séquences multi-vues.

Dans cet article, nous proposons une technique d'estimation conjointe qui permet d'exploiter les différentes relations existantes entre les champs de déplacements présents dans une séquence d'images multi-vues. La stratégie adoptée permet de réduire le nombre de variables à estimer et augmenter ainsi la cohérence, la fiabilité et la précision des estimations. Le problème d'estimation conjointe est formulé comme un problème de programmation convexe, consistant à minimiser une fonction objectif convexe sur l'intersection d'ensembles convexes construits à partir des connaissances *a priori* et d'observations.

Cet article est structuré comme suit. La section 2 décrit la relation entre les vecteurs de disparité et de mouvement, dans le cas de séquences vidéo stéréoscopiques. Nous détaillerons dans la section 3 la formulation du problème d'estimation conjointe et nous introduisons les ensembles de contraintes convexes que nous proposons. Des résultats expérimentaux sont présentés dans la section 4.

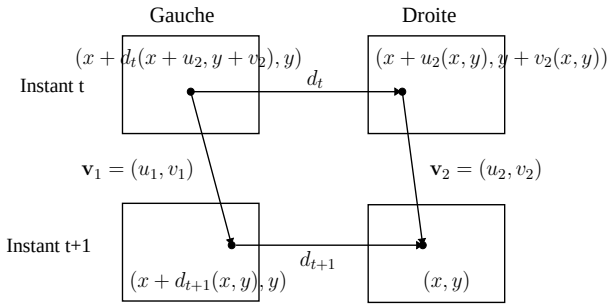


FIG. 1 – La contrainte de cohérence stéréo-cinétique.

2 Relation entre disparité et mouvement

Un système de caméras stéréoscopiques permet d'observer l'évolution d'une scène à partir de deux points de vues distincts. Deux paires stéréoscopiques consécutives sont donc constituées de deux images au temps t, une image gauche I_1^t et une image droite I_2^t, et deux images au temps $t+1$, I_1^{t+1} et I_2^{t+1}. Soient $\mathbf{d}_t$ le vecteur de disparité au temps t, liant un point de l'image gauche à un point de l'image droite, et $\mathbf{d}_{t+1}$ la disparité au temps $t+1$. Soient $\mathbf{v}_1 = (u_1, v_1)$ et $\mathbf{v}_2 = (u_2, v_2)$ les vecteurs donnant le déplacement temporel entre les instants t et $t+1$. Si tous ces vecteurs se rapportent aux projections du même point physique dans la scène, la contrainte suivante doit être vérifiée [3], [4] :

$$\mathbf{d}_t + \mathbf{v}_2 - \mathbf{d}_{t+1} - \mathbf{v}_1 = 0. \tag{1}$$

Cette contrainte, appelée contrainte de *cohérence* (cf. figure 1), établie une relation linéaire entre les vecteurs de disparité et les vecteurs de mouvement. Supposons que les caméras sont parallèles (la disparité se réduit à une seule composante horizontale) et que le point physique est projeté au pixel $s = (x, y)$ dans l'image I_2^{t+1}, l'équation (1) peut être réécrite comme suit :

$$\begin{cases} u_1(x + d_{t+1}(s), y) \simeq d_t(s + \mathbf{v}_2(s)) + u_2(s) - d_{t+1}(s), \\ v_1(x + d_{t+1}(s), y) \simeq v_2(s). \end{cases}$$

Cette équation implique qu'il est possible d'obtenir u_1 et v_1 à partir de u_2, v_2 et d_{t+1}, à condition de connaître la disparité d_t. L'estimation des paramètres $\{u_1, v_1, u_2, v_2, d_{t+1}\}$ peut donc se faire par l'estimation conjointe des variables $\{u_2, v_2, d_{t+1}\}$.

3 Estimation conjointe des champs de déplacements

L'existence d'une relation entre la disparité et le mouvement apparent permet d'envisager une technique d'estimation conjointe. La stratégie adoptée, dans ce travail, consiste à estimer simultanément le mouvement apparent droit et la disparité à l'instant $t+1$ en minimisant une fonctionnelle objectif convexe sous des contraintes convexes. La disparité au temps t, supposée connue, peut avoir été obtenue par la méthode proposée dans [5] ou par une analyse conjointe effectuée à l'instant précédent. Le mouvement apparent gauche $\mathbf{v}_1$ se déduit immédiatement des trois vecteurs de déplacements estimés grâce à la contrainte de cohérence.

3.1 La fonction objectif

Supposons que les quatre points homologues, qui sont les projections du même élément de la scène, ont la même valeur de luminance, le mouvement apparent droit et la disparité à l'instant $t+1$ peuvent être estimés simultanément en minimisant la fonction objectif suivante :

$$\begin{aligned} \tilde{J}(\mathbf{v}_2, d_{t+1}) = & \sum_{(x,y)\in\mathcal{D}} [I_2^{t+1}(x,y) - I_2^t(x+u_2, y+v_2)]^2 \\ & + \sum_{(x,y)\in\mathcal{D}} [I_2^{t+1}(x,y) - I_1^{t+1}(x+d_{t+1}, y)]^2 \\ & + \sum_{(x,y)\in\mathcal{D}} [I_2^t(x+u_2, y+v_2) - I_1^{t+1}(x+d_{t+1}, y)]^2 , \end{aligned} \tag{2}$$

où $\mathcal{D} \subset \mathbb{N}^2$ est le support de l'image. Cette fonction combine trois mesures de corrélation : les deux premières sont relatives au mouvement droit et à la disparité et la troisième corrélation se rapporte aux deux champs simultanément. Cependant, la minimisation de cette fonction, non-convexe par rapport à la disparité d_{t+1} et au champ de mouvement $\mathbf{v}_2$, ne permet que de converger vers des minima locaux. Pour contourner cette difficulté, nous supposons que des estimées initiales $\bar{d}_{t+1}$ et $\bar{\mathbf{v}}_2 = (\bar{u}_2, \bar{v}_2)$ sont accessibles et nous développons, autour de ces estimées, les termes non-linéaires en série de Taylor au premier ordre comme suit :

$$\begin{aligned} I_2^t(x+u_2, y+v_2) \simeq & \; I_2^t(x+\bar{u}_2, y+\bar{v}_2) \\ & + (u_2-\bar{u}_2) I_2^{t,x} + (v_2-\bar{v}_2) I_2^{t,y} , \\ I_1^{t+1}(x+d_{t+1}, y) \simeq & \; I_1^{t+1}(x+\bar{d}_{t+1}, y) + (d_{t+1}-\bar{d}_{t+1}) I_1^{t+1,x} , \end{aligned}$$

où $I_2^{t,x}$ et $I_2^{t,y}$ sont, respectivement, le gradient horizontal et vertical de l'image de droite compensée et $I_1^{t+1,x}$ est le gradient horizontal de l'image de gauche compensée. En posant $w = (u_2, v_2, d_{t+1})^\top$, le vecteur de paramètres à estimer, et en introduisant les linéarisations ci-dessus, l'expression du critère (2) se réécrit de la manière suivante :

$$J_{\mathcal{D}}(w) = \sum_{i=1}^{3} \sum_{s\in\mathcal{D}} [L_i(s)\, w(s) - r_i(s)]^2 , \tag{3}$$

où

$$\begin{cases} L_1 = [I_2^{t,x},\ I_2^{t,y},\ 0] \\ L_2 = [0,\ 0,\ I_1^{t+1,x}] \\ L_3 = [I_2^{t,x},\ I_2^{t,y},\ -I_1^{t+1,x}], \end{cases}$$

et

$$\begin{cases} r_1 = -I_2^t + \bar{u}_2 I_2^{t,x} + \bar{v}_2 I_2^{t,y} + I_2^{t+1} \\ r_2 = -I_1^{t+1} + \bar{d}_{t+1} I_1^{t+1,x} + I_2^{t+1} \\ r_3 = -I_2^t + I_1^{t+1} + \bar{u}_2 I_2^{t,x} + \bar{v}_2 I_2^{t,y} - \bar{d}_{t+1} I^{t+1,x}. \end{cases}$$

Minimiser le critère $J_{\mathcal{D}}$ est un problème inverse *mal posé* car d'une part, on ne dispose, en chaque point, que d'une unique équation pour trouver trois inconnus et d'autre part les composantes de $\{L_i\}_i$ peuvent s'annuler, simultanément, en certains points de l'image. Pour obtenir des solutions uniques et fiables, il faut donc incorporer autant d'information *a priori* que possible sur les champs à estimer. Pour résoudre le problème d'optimisation sous contraintes ainsi résultant, nous avons mis en œuvre une approche ensembliste où chaque contrainte est représentée par un ensemble convexe et l'intersection de tous

ces ensembles convexes constitue l'ensemble des solutions admissibles. Le problème d'estimation conjointe revient ainsi à trouver une solution admissible qui minimise la fonction objectif convexe (3). Il se formule dans un espace hilbertien réel $\mathcal{H}$ comme suit

$$\text{Trouver } w \in S = \bigcap_{i=1}^{m} S_i \text{ tel que } J(w) = \inf J(S) \ , \qquad (4)$$

où la fonction objectif $J : \mathcal{H} \rightarrow]-\infty, +\infty]$ est une fonction convexe et les ensembles de contraintes $(S_i)_{1\leq i\leq m}$ sont des convexes fermés de $\mathcal{H}$. Les contraintes étant modélisées comme des ensembles de niveau de fonctions convexes, les ensembles $(S_i)_{1\leq i\leq m}$ sont de la forme

$$\forall i \in \{1, ..., m\}, \qquad S_i = \{w \in H \mid f_i(w) \leq \delta_i\} \ , \qquad (5)$$

où $(f_i)_{1\leq i\leq m}$ est une famille de fonctions convexes continues et $(\delta_i)_{1\leq i\leq m} \in \mathbb{R}$.

3.2 Les contraintes convexes

En supposant que les vecteurs de mouvement sont bornés par les limites horizontales et verticales du déplacement, la première contrainte que l'on peut imposer est la contrainte de plage de valeurs. De même, la fonction de disparité est toujours bornée et comprise entre une valeur minimale $d_{\min} \geq 0$ et une valeur maximale $d_{\max}$. Les ensembles de contraintes associés à ces informations sont :

$$S_1 = \{w \in \mathcal{H} \mid u_{\min} \leq u_2 \leq u_{\max}\} \ , \qquad (6)$$

$$S_2 = \{w \in \mathcal{H} \mid v_{\min} \leq v_2 \leq v_{\max}\} \ , \qquad (7)$$

$$S_3 = \{w \in \mathcal{H} \mid d_{\min} \leq d_{t+1} \leq d_{\max}\} \ . \qquad (8)$$

Le second *a priori* que nous avons considéré concerne la régularité des champs de mouvement et de disparité, lisses dans les homogènes et discontinus aux frontières des objets. Pour traduire cette propriété, nous avons utilisé une contrainte de régularisation basée sur la variation totale (VT), qui est une mesure de la somme des amplitudes des discontinuités présentes dans l'image. La régularisation par variation totale a été appliquée pour résoudre divers problèmes inverses mal posés et s'est particulièrement imposée comme une approche performante en traitement d'images [5], [6]. L'idée de cette régularisation est motivée par l'observation que, dans de nombreux types de problèmes, la variation totale de l'image originale ne dépasse pas une certaine borne connue. Cette information restreint la solution à appartenir aux ensembles convexes suivants :

$$S_4 = \{w \in \mathcal{H} \mid \text{tv}(u_2) \leq \tau_{u_2}\} \ , \qquad (9)$$

$$S_5 = \{w \in \mathcal{H} \mid \text{tv}(v_2) \leq \tau_{v_2}\} \ , \qquad (10)$$

$$S_6 = \{w \in \mathcal{H} \mid \text{tv}(d_{t+1}) \leq \tau_d\} \ , \qquad (11)$$

où τ_{u_2}, τ_{v_2} et τ_d sont des constantes positives qui peuvent être estimées à partir d'expérimentation ou en exploitant des bases de données d'images du même type.

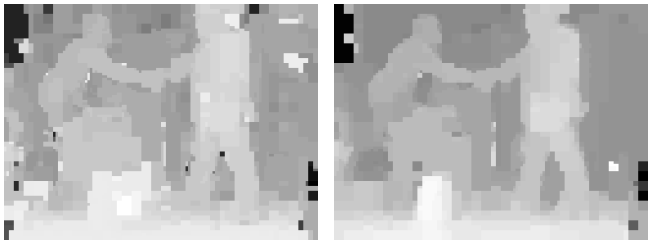
Le problème d'estimation conjointe est finalement formulé comme celui de la minimisation de la fonctionnelle (3) sous les contraintes $(S_i)_{1\leq i\leq 6}$. Pour résoudre numériquement ce problème d'optimisation sous contraintes, nous avons adapté l'algorithme parallèle et itératif par blocs proposé par Combettes [7] pour résoudre des problèmes de restauration d'images. Cet algorithme offre une méthode de résolution puissante et efficace pour l'optimisation d'une fonction convexe et différentiable.

4 Résultats expérimentaux

Nous évaluons la méthode proposée sur les séquences multi-vues réelles "Book Arrival" et "Doors Flowers" [8]. Les images originales, de taille 512×384 pixels, sont montrées à la figure 3(a). Pour les deux séquences, nous considérons seulement 4 vues des seize vues disponibles. Un schéma d'interdépendance simple est utilisé entre les vues, i.e, chaque vue utilise la vue de gauche comme vue de référence. Nous utilisons le logiciel JMVM 8.0 [9] pour effectuer nos tests.

Pour procéder à l'estimation conjointe des champs de disparité et de mouvement impliquées dans ces séquences vidéo multi-vues, il faut disposer d'un champ de disparité entre les deux premières vues de la première trame. Ce champ est obtenu par la méthode proposée dans [5]. Le mouvement apparent droit et la disparité à la trame courante sont ensuite estimés en utilisant la technique d'estimation conjointe présentée à la section 3. Le mouvement apparent gauche se déduit, enfin, des trois vecteurs de déplacements estimés. Cependant, pour pallier au phénomène d'occlusion, le champ résultant est affiné en utilisant l'approche d'optimisation convexe proposée dans [5]. La figure 3 présente les résultats obtenus. Ces résultats montrent que notre approche permet d'obtenir des champs de mouvement et de disparité cohérents et lisses tout en respectant les discontinuités autour des objets.

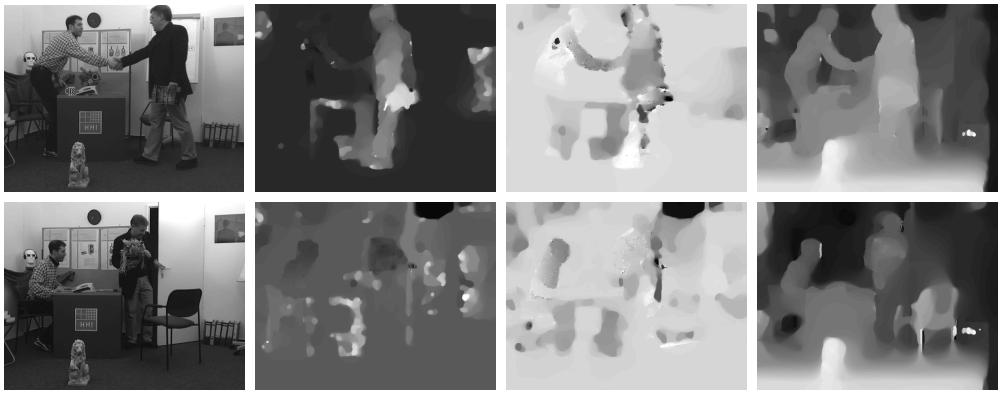
Une fois les champs de mouvement et de disparité calculés conjointement, ils seront insérés dans l'encodeur H.264/AVC. Pour assurer leur compatibilité avec la structure par blocs de l'encodeur, une segmentation de ces vecteurs de déplacement permet d'aboutir à des champs de vecteurs épars (par blocs). Cette segmentation est établie suivant un critère débit-distorsion et de manière compatible avec les modes de partition de H264. Le résultat de cette segmentation sur un champ de disparité et pour un QP égal à 22 est présenté à la figure 2. Comparée à la référence H264/AVC, notre approche (estimation dense suivie d'une segmentation) produit des champs de vecteurs plus réguliers, favorisant la sélection du mode SKIP et offrant ainsi des gains significatifs en terme de réduction de débit. En effet, lorsque le mode SKIP est sélectionné parmi l'ensemble des possibilités de codage, aucune information additionnelle (vecteur de déplacement, résiduel de bloc) n'est transmise au décodeur. Pour les deux séquences, aussi bien en bas débit qu'en haut débit, la méthode proposée obtient un pourcentage d'augmentation du nombre de MB encodés avec le mode SKIP supérieur au pourcentage obtenu avec H264.



(a) H.264/AVC (b) VT

FIG. 2 – Exemple de champs de vecteurs de disparité par blocs à QP 22 (de la séquence "Book arrival", trame 48).

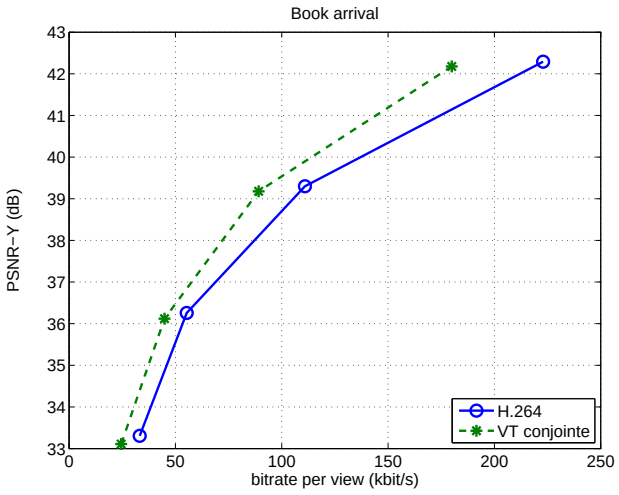
La figure 4 affiche les résultats en termes de performance débit-distortion. La comparaison de l'estimation dense conjointe



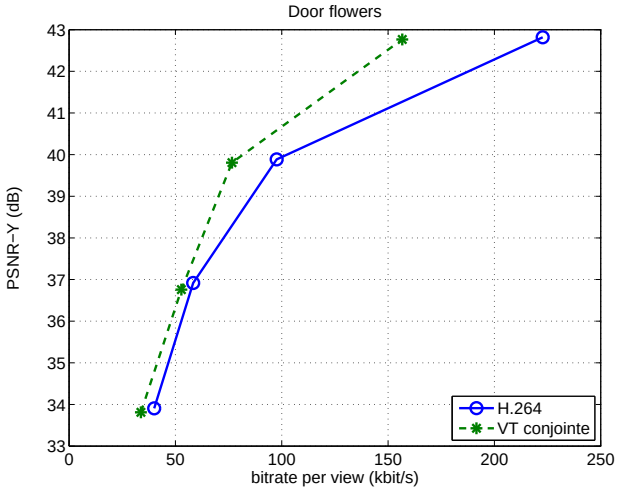
(a) (b) (c) (d)

FIG. 3 – Résultat de l'analyse conjointe sur les séquences "Book arrival" (en haut) et "Door flowers" (en bas) : (a) image de gauche (b) mouvement horizontal gauche (c) mouvement vertical gauche, (d) disparité horizontale à la trame courante.

indique clairement le bénéfice d'une estimation dense. Les courbes correspondent à 4 valeurs de QP qui sont 22, 27, 32, 37.



(a) Book arrival



(b) Door flowers

FIG. 4 – Rate-distortion coding results.

5 Conclusion

Dans cet article, nous avons proposé une méthode d'estimation conjointe des champs de déplacements dans une séquence d'images stéréoscopiques. Les résultats présentés soulignent l'intérêt de l'estimation conjointe aussi bien en termes de reconstruction qu'en termes de cohérence des champs estimés.

Références

[1] E. Martinian, A. Behrens, J. Xin, A. Vetro and H. Sun, " Extensions of H.264/AVC for Multiview Video Compression," *IEEE Int. Conf. on Image Process.*, ISSN : 1522-4880, pp. 2981-2984, Oct. 2006.

[2] Y. Chen, Y. K. Wang, K. Ugur, M. M. Hannuksela, J. Lainema, and M. Gabbouj, "The Emerging MVC Standard for 3D Video Services," *EURASIP Journal on Advances in Signal Processing*, Vol. 2009, No. 1, 2009, pp. 1-13.

[3] A. Tamtaoui and C. Labit, "Coherent disparity and motion compensation in 3DTV image sequencecoding schemes," in *Proc. Int. Conf. Acoustics, Speech, and Signal Process.*, Toronto, Canada, Apr. 14-17, 1991, vol. 4, pp. 2845–2848.

[4] W. Yang, K. Ngan, J. Lim and K. Sohn, "Joint Motion and Disparity Fields Estimation for Stereoscopic Video Sequences," *Signal Processing : Image Communication*, vol. 20, no. 3, pp. 265-276, Mar. 2005.

[5] W. Miled, J. C. Pesquet and M. Parent, "Disparity map estimation using a total variation bound," in *Proc. 3rd Canadian Conf. Comput. Robot Vis.*, Quebec, Canada, Jun. 7-9, 2006, pp. 48–55.

[6] L. I. Rudin, S. Osher, and E. Fatemi, "Nonlinear total variation based noise removal algorithms," *Physica D*, vol. 60, pp. 259–268, 1992.

[7] P. L. Combettes, "A block iterative surrogate constraint splitting method for quadratic signal recovery," *IEEE Trans. Signal Process.*, vol. 51, pp. 1771–1782, Jul. 2003.

[8] I. Feldmann, M. Muller, F. Zilly, R. Tanger, K. Muller, A. Smolic, P. Kauff and T. Wiegand, "HHI Test Material for 3D Video," *IEEE Trans. Signal Process.*, Archamps, France, May 2008.

[9] A. Vetro, P. Pandit, H. Kimata, A. Smolic and Y. Wang, "Joint Multiview Video Model (JMVM) 8.0," *IEEE Trans. Signal Process.*, JVT-AA207, Geneva, Apr. 2008.