Estimations simultanees du flot de scène et d'occultations par un système d'EDP

Frédéric Huguet, Frédéric Devernay

Estimations simultanées du flot de scène et d'occultations par un système d'EDP

Simultaneous scene flow and occlusions estimations by using a common PDE framework

Frédéric Huguet[1] Frédéric Devernay[2]

[1] INRIA Rhone Alpes, Laboratoire Jean Kuntzmann
[2] INRIA Rhone Alpes

INRIA Rhone Alpes
ZIRST 655 Avenue de l'Europe, 38330 Montbonnot Saint Ismiers
Frederic.Huguet@inrialpes.fr, Frederic.Devernay@inrialpes.fr

Résumé

Ce papier présente une méthode d'estimation du flot de scène à partir de séquences stéréo issues d'un couple de caméras calibrées. Le flot de scène représente le champ de déplacement 3D des points d'une scène, de telle sorte que le flot optique traditionnel peut être vu comme la projection de celui-ci dans les images. Nous proposons d'estimer le flot de scène en couplant l'évaluation du flot optique dans les séquences d'images associées à chaque caméra, à l'estimation de la correspondance stéréo dense entre les images. De plus, notre approche évalue, en même temps que le flot de scène, les occultations à la fois en flot optique et en stéréo. Nous obtenons au final un système d'EDP couplant le flot optique et la stéréo, que nous résolvons numériquement à l'aide d'un algorithme multirésolution original. Alors que les précédentes méthodes variationnelles estimaient la reconstrution 3D au temps t et le flot de scène séparément, notre méthode estime les deux simultanément. Nous présentons des résultats numériques sur des séquences synthétiques avec leur vérité terrain, et nous comparons également la précision du flot de scène projeté dans une caméra avec une méthode récente et performante d'estimation variationnelle du flot optique. Des résultats sont présentés sur une séquence stéréo réelle, se rapportant à un mouvement non rigide et à de larges discontinuités en flot optique et en stéréo.

Mots Clef

Flot de scène, occultations, stéréoscopie, EDP.

Abstract

This paper presents a method for scene flow estimation from a calibrated stereo image sequence. The scene flow contains the 3-D displacement field of scene points, so that the 2-D optical flow can be seen as a projection of the scene flow onto the images. We propose to recover the scene flow by coupling the optical flow estimation in both cameras with dense stereo matching between the images, thus reducing the number of unknowns per image point. Moreover our approach handles occlusions both for the optical flow and the stereo. We obtain a partial differential equations system coupling both the optical flow and the stereo, which is numerically solved using an original multi-resolution algorithm. Whereas previous variational methods were estimating the 3-D reconstruction at time t and the scene flow separately, our method jointly estimates both. We present numerical results on synthetic data with ground truth information, and we also compare the accuracy of the scene flow projected in one camera with a state-of-the-art single-camera optical flow computation method. Results are also presented on a real stereo sequence with large motion and stereo discontinuities.

Keywords

Scene flow, occlusions, stereo, PDE.

1 Introduction

Le flot de scène fut introduit par Vedula *et al.* [20, 21] comme étant le champ de vecteurs 3D défini sur les surfaces présentes dans une scène, décrivant le mouvement de chaque point 3D entre 2 instants consécutifs. Il peut être vu comme une extension 3D du flot optique, mais ce dernier peut lui-même être vu comme la projection du flot de scène dans les images, projection donnant un champ de vecteur 2D. Plusieurs méthodes proposent de reconstruire le flot de scène à partir du flot optique observé dans une ou plusieurs caméras [21, 22], mais l'étape de reconstruction est en général souscontrainte ou surcontrainte, et de plus les flots optiques obtenus avec plusieurs caméras peuvent ne pas être consistants entre eux.

Afin de surmonter ces problèmes, nous utilisons une paramétrisation minimale du flot de scène à partir du flot

optique et de la disparité d'une séquence d'images stéréo (cette description du flot de scène est parfois appelée *flot de disparité* [9]). Puisque cette paramétrisation est réalisée dans l'espace des images, le problème devient proche de celui de l'estimation du flot optique, avec plus d'inconnues et de mesures en chaque pixel.

Beaucoup de travaux ont été effectués dans le domaine de l'estimation du flot optique par méthodes variationnelles, depuis le travail pionnier de Horn et Schunck [3]. Certaines méthodes ont modifié le terme de régularisation afin de prendre en compte les discontinuités du flot optique [7]. De récents travaux se sont consacrés à la réduction du coût en temps de ces méthodes, aboutissant à des algorithmes temps réel [5] ou parallélisés [6].

Néanmoins, les meilleurs résultats en terme de précision furent obtenus par Brox *et al.* [4] :ils évitent toute linéarisation des différents termes d'énergie dans leur formulation variationnelle en warpant l'image à l'instant $t+1$ vers l'image à l'instant t, et l'énergie globale n'est linéarisée qu'au moment de la résolution numérique. Ainsi, ils évitent les imprécisions dues a l'approximation linéaire des termes de l'énergie à minimiser, en particulier des termes d'attache aux données qui avaient toujours été linéarisés depuis Horn et Schunck. Cette méthode est également robuste aux variations d'illumination, et par l'utilisation d'une fonction convexe, est également relativement robuste numériquement aux occultations et aux discontinuités (mais ne les traite pas explicitement).Slesareva *et al.* [17] ont adapté cette formulation variationnelle au problème d'estimation de cartes de disparité denses.

Concernant l'estimation du *flot de scène* dans le cadre variationnel, la seule méthode qui traitait à la fois de reconstruction et d'estimation du flot de scène était celle proposée par Pons *et al.* [15]. L'estimation du flot de scène est réalisée en optimisant alternativement la reconstruction et le champ de déplacement 3D. Ce dernier est évalué en optimisant une énergie qui prend en compte la différence entre les images consécutives reprojetées sur la reconstruction 3D évaluée. De récents travaux proposèrent une estimation jointe de la disparité et du flot optique : Dongbo Min *et al.* [13], qui néanmoins ne traitent pas les variations d'illumination et les occultations, ainsi que Isard et MacCormick [11], qui estiment des valeurs entières de disparité et de flot.

Nous proposons une méthode qui estime le flot de scène par évaluation couplée de la surface reconstruite et du champ de déplacement, à partir de séquences stéréoscopiques issues de caméras calibrées. Cette méthode prend en compte la contrainte épipolaire entre les images d'une paire stéréo prise à un instant t, aboutissant à une paramétrisation minimale du flot de scène. Seulement 4 variables sont cherchées comme minimum d'une certaine énergie, étant définies pour chaque pixel dans une image de référence de la manière suivante : la disparité au temps t, la disparité au temps $t+1$ et le flot optique (le système de caméras étant calibré, le flot de scène est calculable directement à partir de ces variables).

Ceci nous amène à résoudre numériquement, via un algorithme multirésolution, un système d'équations aux dérivées partielles (EDP) fortement non linéaires et couplées. Notre méthode évite la linéarisation de l'énergie à minimiser. En effet, Brox *et al.* ont prouvé que celà permettait d'améliorer la précision des résultats numériques. Ce principe est étendu à toutes les contraintes issues de la modélisation du problème du flot de scène. D'autre part, un terme de régularisation adapté nous permet de préserver les discontinuités à la fois dans la reconstruction et dans le champ de déplacement 3D, permettant ainsi à des fractures d'apparaître sur une surface régulière au cours du temps.

Le reste de cet article est organisé de la manière suivante : nous exposons d'abord la formulation mathématique couplant le flot optique et la stéréoscopie, et les différents termes de l'énergie à minimiser. Nous présentons ensuite les difficultés numériques inhérentes au problème, et l'algorithme global. Enfin, nous présentons des résultats numériques obtenus sur des séquences synthétiques avec la vérité terrain associée, et des résultats obtenus sur une séquence stéréo réelle représentant une scène non rigide, avec de larges discontinuités en mouvement et en stéréo.

2 Une formulation variationnelle unifiée pour le flot optique et la stéréo

Notre but est d'estimer un flot de scène dense, tout en préservant les discontinuités des surfaces et du mouvement. Zhang et Kambhamettu [23] réalisent celà en segmentant la scène observée, puis en appliquant une régularisation par morceaux, mais ce problème peut aussi être résolu en utilisant une fonctionnelle de régularisation appropriée. Étant donné que nous travaillons sur des séquences stéréo issues d'un système calibré, nous rectifions toutes les images obtenues afin de réduire la dimensionnalité du problème de mise en correspondance stéréoscopique : apres rectification, la disparité se trouve correspondre simplement à une différence d'abscisses entre deux points dans les images associés à un même point 3D.

Un lissage gaussien ($\sigma = 1.25$) est aussi appliqué aux images dans le but d'éviter ertaines instabilités numériques [2]. Notre méthode utilise les avantages numériques apportés par le travail de Brox *et al.* : robustesse aux variations d'illumination grâce à une contrainte sur l'invariance temporelle des gradients, et robustesse numérique aux occultations en stéréo et flot optique par l'utilisation d'une fonction Ψ appropriée.

Soient $I_l(x,y,t), I_r(x,y,t) : \Omega \subset \mathbb{R}^3$ les séquences d'images gauche et droite (Ω est le domaine de définition rectangulaire des images). Soit $(u,v) : \Omega \to \mathbb{R}^2$ le flot optique dans les images de gauche, et $(d,d') : \Omega \to \mathbb{R}^2$ les cartes de disparité aux temps t et $t+1$. $\mathbf{w} = (u,v,1)^\top$ est le vecteur déplacement entre l'image de gauche I_l au temps t et I_l au temps $t+1$, $\mathbf{d} = (d,0,0)$ est le dépla-

cement entre I_l et I_r au temps t, et $\mathbf{d}' = (d', 0, 0)$ est le déplacement entre I_l et I_r au temps $t+1$. Comme décrit dans Fig. 1, un point (x, y, t) dans I_l correspond aux points $(x + u(x, y), y + v(x, y), t + 1)$ dans I_l, $(x + d(x, y), y, t)$ dans I_r, et $(x + u(x, y) + d'(x, y), y + v(x, y), t + 1)$ dans I_r : *le domaine de définition pour les fonctions scalaires u, v, d et d' est toujours I_l au temps t.* Il est clair que la reconstruction 3D du point de la scène observé a la position (x, y) et au temps t dans I_l peut être obtenu à partir de d, et de manière similaire sa reconstruction à $t + 1$ est obtenue de u, v, et d'. Le flot de scène est ensuite aisément reconstruit en faisant la différence entre ces deux positions reconstruites.

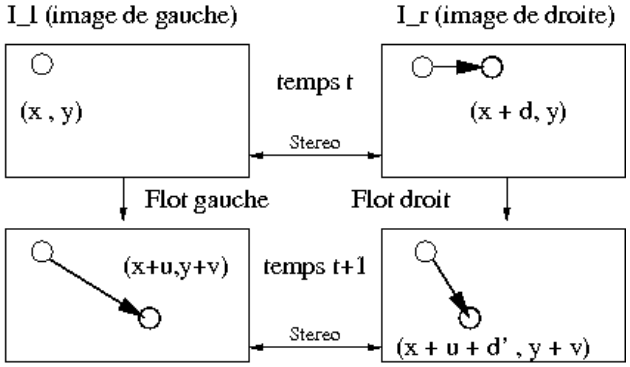

FIG. 1 – Le mouvement d'un point de la scène projeté dans les images stéréo, entre deux instants consécutifs

Nous écrivons l'énergie totale comme la somme d'un terme d'attache aux données et d'un terme de régularisation :

$$E(u, v, d, d') = E_{Data} + \alpha E_{Smooth}, \qquad (1)$$

α étant le paramètre de régularisation. E_{Data} est composé de 4 termes, correspondant aux 4 relations entre les images des séquences stéréo, montrées par Fig. 1 :

$$E_{Data} = \int_\Omega \left(\beta_{fl} E_{fl} + \beta_{fr} E_{fr} + \beta_{st} E_{st} + \beta_s E_s\right) \mathbf{dx}. \qquad (2)$$

$(x, y) : \Omega \rightarrow \beta_{fl}(x, y)$ vaut 1 pour les pixels non occultés pour le flot optique des images de gauche, et 0 en cas de présence d'une occultation. Les autres fonctions β jouent un rôle similaire pour les occultations associées à chaque composante de E_{data}. Introduisons la notation suivante pour la différence en intensité et illumination entre deux pixels :

$$\Delta(I, \mathbf{x}; I', \mathbf{y}) = |I'(\mathbf{y}) - I(\mathbf{x})|^2 + \gamma |\nabla I'(\mathbf{y}) - \nabla I(\mathbf{x})|^2, \qquad (3)$$

où $\nabla = (\partial_x, \partial_y)^\top$. Les 4 termes de E_{Data} peuvent être écrits de la manière suivante :

$$E_{fl}(u, v, d, d') = \Psi\left(\Delta(I_l, \mathbf{x}; I_l, \mathbf{x} + \mathbf{w})\right), \qquad (4)$$

$$E_{fr}(u, v, d, d') = \Psi\left(\Delta(I_r, \mathbf{x} + \mathbf{d}; I_r, \mathbf{x} + \mathbf{w} + \mathbf{d}')\right), \qquad (5)$$

$$E_{st}(u, v, d, d') = \Psi\left(\Delta(I_l, \mathbf{x} + \mathbf{w}; I_r, \mathbf{x} + \mathbf{w} + \mathbf{d}')\right), \qquad (6)$$

$$E_s(u, v, d, d') = \Psi\left(\Delta(I_l, \mathbf{x}; I_r, \mathbf{x} + \mathbf{d})\right). \qquad (7)$$

E_{fl} est le terme d'attache aux données relatif au flot optique gauche, E_{fr} correspondant au flot optique droit, qui a la même composante verticale v que le flot optique gauche, les images étant rectifiées. De manière similaire, E_s correspond à la correspondance stéréo entre les images gauche et droite au temps t, et E_{st} de même à $t + 1$. Les pixels dans l'image de gauche à t peuvent être occultés dans l'une ou plusieurs des 3 autres images, et l'utilisation d'un pénaliseur quadratique leur donnerait dans ce cas trop d'influence sur la solution numérique. Pour remédier à ce problème, nous utilisons la fonction Ψ [1, 4], définie par $\Psi(s^2) = \sqrt{s^2 + \epsilon^2}$ (avec $\epsilon = 0.001$), ce qui correspond à une minimisation L^1 modifiée pour avoir la dérivabilité en 0. La fonction est appliquée séparément à chaque composante de E_{data}, puisque les pixels peuvent par exemple être occultés en stéréo, mais pas en flot optique, et inversement. En outre, eq. (8) comprend un terme modélisant l'hypothèse de faible variation temporelle du gradient des images [4], introduisant la robustesse aux variations d'illumination (locales ou globales) et aux surfaces non lambertiennes (les termes de stéréo peuvent être affectés de manière importante par de telles surfaces, puisque utilisant des images issues de points de vue différents). Le paramètre γ doit être ajusté empiriquement, et dépend de l'amplitude de la variation d'illumination dans la scène.

Nous aurions pu considérer que la disparité à t puisse provenir de l'estimation du flot de scène entre les instants $t-1$ et t, mais si cette disparité obtenue contient une erreur non négligeable, celle-ci pourrait ensuite se propager à d', u, et v. En minimisant d en même temps que les autres inconnues, nous sommes capables de réévaluer toutes les composantes du flot de scène : la reonstruction 3D (de d) et le champ de déplacement 3D (de u, v, et $d' - d$).

Le terme de régularisation s'exprime ainsi :

$$E_{Smooth} = \int_\Omega \Psi(|\nabla u|^2 + |\nabla v|^2 + \lambda |\nabla (d' - d)|^2 + \mu |\nabla d|^2) \mathbf{dx}. \qquad (8)$$

En réduisant l'influence des hauts gradients du flot optique ou de la disparité sur l'énergie totale, la fonction Ψ a un rôle différent ici : elle permet de préserver les discontinuités de u, v, d, et d' [5]. Contrairement au cas de E_{data}, Ψ est appliquée à la somme des normes des gradients, car habituellement les discontinuités apparaissent simultanément dans la disparité d, le flot optique (u, v), et le flot de disparité $d' - d$ (sauf dans quelques cas spéciaux, comme dans l'exemple synthétique présenté dans la partie Résultats numériques).

L'effet de la régularisation sur le flot de scène ne devrait pas dépendre de l'orientation du champ de déplacement par rapport à la caméra, c'est pourquoi le paramètre λ devrait être ajusté pour rescaler l'influence de la régularisation sur le flot optique par rapport à la régularisation du flot de disparité, mais ne devrait pas être supérieur à μ pour éviter des oscillations durant l'optimisation : $\lambda < h/b$, où h est la distance moyenne des caméras à la scène, et b est la ba-

seline du dispositif stéréo. L'effet de ce paramètre sera un flot de disparité $(d' - d)$ plus régulier et des discontinuités plus petites lorsque la baseline est petite. Le paramètre μ règle l'influence relative de la disparité à t par rapport au flot optique. Puisque les discontinuités typiques dans les deux termes observées dans une scène devraient avoir le même effet sur E_{Smooth}, une bonne estimation devrait être $\mu = hs/bS$ où s est l'amplitude attendue (en unités du monde) du flot de scène, et S une taille caractéristique de la scène : si le mouvement entre t et $t+1$ est petit par rapport à la taille de la scène, alors μ devrait être petit lui aussi.

3 Optimisation

3.1 Équations d'Euler Lagrange

Un extremum de l'énergie E vérifie la condition nécessaire $\nabla E(u, v, d, d') = 0$, qui peut être réécrite sous la forme $(\partial_u E, \partial_v E, \partial_d E, \partial_{d'} E) = (0, 0, 0, 0)$. Ce sont les équations d'Euler Lagrange du problème. L'énergie E n'étant pas convexe, cette condition n'est pas suffisante et il se peut que l'on obtienne au final des minima locaux de E non souhaités. Nous verrons plus loin comment une approche multirésolution nous permet d'éviter ce cas de figure. Les 4 équations peuvent être trouvées de la même manière, en utilisant les outils du calcul des variations, et ont une structure similaire. Introduisons les abréviations suivantes :

$$I_{lx} := \partial_x I_l(\mathbf{x}+\mathbf{w}), \quad I_{lxz} := \partial_x I_l(\mathbf{x}+\mathbf{w}) - \partial_x I_l(\mathbf{x}), \quad (9)$$

$$I_{ly} := \partial_y I_l(\mathbf{x}+\mathbf{w}), \quad I_{lyz} := \partial_y I_l(\mathbf{x}+\mathbf{w}) - \partial_y I_l(\mathbf{x}), \quad (10)$$

$$I_{lz} := I_l(\mathbf{x}+\mathbf{w}) - I_l(\mathbf{x}), \quad I_{lyy} := \partial^2_{yy} I_l(\mathbf{x}+\mathbf{w}), \quad (11)$$

$$I_{lxx} := \partial^2_{xx} I_l(\mathbf{x}+\mathbf{w}), \quad I_{lxy} := \partial^2_{xy} I_l(\mathbf{x}+\mathbf{w}), \quad (12)$$

$$I_l^{t+1} := I_l(\mathbf{x}+\mathbf{w}) \quad (13)$$

et des abréviations similaires pour l'image de droite I_r. La dernière notation est utile pour voir l'index temporel dans les équations à venir.

$$\Psi'_{fl} = \partial_x \Psi(\Delta(I_l, \mathbf{x}; I_l, \mathbf{x}+\mathbf{w})) \quad (14)$$

$$\Psi'_{fr} = \partial_x \Psi(\Delta(I_r, \mathbf{x}+\mathbf{d}; I_r, \mathbf{x}+\mathbf{w}+\mathbf{d}')) \quad (15)$$

$$\Psi'_{st} = \partial_x \Psi(\Delta(I_l, \mathbf{x}+\mathbf{w}; I_r, \mathbf{x}+\mathbf{w}+\mathbf{d}')) \quad (16)$$

$$\Psi'_{\mathrm{div}} = \partial_x \Psi(|\nabla u|^2 + |\nabla v|^2 + \lambda|\nabla(d'-d)|^2 + \mu|\nabla d|^2). \quad (17)$$

En calculant $\partial_u E$ avec une dérivée de Gâteaux, nous obtenons :

$$\begin{aligned}&\beta_{fl}\Psi'_{fl}\cdot\big(I_{lx}I_{lz} + \gamma(I_{lxx}I_{lxz} + I_{lxy}I_{lyz})\big) + \\ &\beta_{fr}\Psi'_{fr}\cdot\big(I_{rx}I_{rz} + \gamma(I_{rxx}I_{rxz} + I_{rxy}I_{ryz})\big) + \\ &\beta_{st}\Psi'_{st}\cdot\big((I_r^{t+1} - I_l^{t+1})(I_{rx} - I_{lx}) + \gamma\big((I_{rx} - I_{lx})(I_{rxx} - I_{lxx}) + \\ &\quad (I_{ry} - I_{ly})(I_{rxy} - I_{lxy})\big)\big) - \alpha\,\mathrm{div}(\Psi'_{\mathrm{div}}\nabla u) = 0, \quad (18)\end{aligned}$$

Cette équation est composée d'un terme issu de E_{Data} et d'un terme diffusif en divergence issu du terme de régularisation de l'énergie. Les 3 autres équations issues de $\partial_v E = 0$, $\partial_{d'} E = 0$, $\partial_d E = 0$ sont tout à fait similaires à l'équation $\partial_u E = 0$, seuls les dérivées des images utilisées et le gradient embarqué dans le terme de divergence changeant d'une équation à l'autre ($\mathrm{div}(\Psi'_{\mathrm{div}}\nabla v)$ dans l'équation correspondant à $\partial_v E = 0$ par exemple.

Les conditions au bord du problème sont des conditions de Neumann : $\forall f \in \{u, v, d, d' - d\}, \nabla f \cdot \mathbf{n} = 0$, où $\mathbf{n}$ est la normale extérieure aux bords de l'image I_l. Dans ce système d'EDP, les 4 inconnues fonctionnelles sont fortement couplées notamment via des non linéarités, mais résoudre ces équations aboutit à la reconstruction du flot de scène.

3.2 Solution numérique

L'énergie E n'est malheureusement pas convexe, ceci étant dû à la non linéarisation des termes de l'énergie. Celà rend le problème mal posé, et nous ne pouvons pas utiliser une simple descente de gradient pour minimiser E comme dans [14]. Dans le but de contourner cette difficulté numérique, nous utilisons une approche multirésolution incrémentale, avec des itérations de point fixe imbriquées sur l'estimation courante de la solution (u, v, d, d') afin de l'affiner à chaque niveau de résolution. Une méthode similaire est utilisée par Brox *et al.* [4] pour résoudre le problème du flot optique. Les pyramides d'images stéréo sont estimées avec un facteur de sous-échantillonnage η, $0.5 < \eta < 1$ pour obtenir une transition régulière entre les niveaux des pyramides (nous utilisons en général $\eta = 0.9$). L'approche multirésolution assure la convergence vers un minimum global, comme celà est montré dans [12]. Cette méthode a prouvé son efficacité sur de nombreux problèmes, et a été récemment améliorée pour obtenir des performances temps réel [5].

Le terme de fidélité aux données dans les équations eq. (19) est composé de valeurs d'image et de gradients d'image évaluées par rapport à l'image de référence I_l au temps t. Ceci est équivalent à warper les 3 autres images (I_l à $t+1$, I_r à t et $t+1$) issues du même niveau dans les pyramides, vers l'image I_l au temps t, et en utilisant l'estimation courante de la solution (u, v, d, d'). Les termes de l'énergie s'évaluent ainsi à partir de ces images warpées et de leurs gradients.

À un niveau de résolution donné, nous traitons les non linéarités des équations en utilisant 2 itérations de point fixe imbriquées, obtenues en réalisant un développement de Taylor au premier ordre des équations d'Euler Lagrange. Ceci aboutit à la résolution d'un système linéaire pour chaque itération de l'itération de point fixe de bas niveau. Cette dernière évalue ainsi de petits incréments de la solution (du, dv, dd, dd'), et les images sont rewarpées en utilisant la transformation $(u+du, v+dv, d+dd, d'+dd')$ à chaque itération. L'itération de point fixe de haut niveau met à jour la solution totale (u, v, d, d') dès que l'itération de bas niveau a convergé, et permet aussi de rewraper les images en utilisant la nouvelle estimation (u, v, d, d'). Nous renvoyons à la section 3.2 de [4] pour davantage de détails sur la manière de transformer les équations d'Eu-

ler Lagrange en système linéaire (bien que cet article ne concerne que le problème du flot optique, ayant ainsi moins d'équations et de non linéarités). L'itération de point fixe de bas niveau utilise une méthode SOR modifiée pour résoudre le système linéaire final (décrit dans [10]). Dans la méthode SOR traditionnelle, la matrice du système est séparée en trois parties : diagonale, triangulaire supérieure et inférieure. Par conséquent, des ordonnancements différents des lignes et des colonnes du système (tout dépend de la manière dont on organise les pixels de l'image pour l'écriture du système) peuvent aboutir à des résultats différents à chaque itération. Notre implémentation utilise alternativement 4 ordonnancements différents, dans lesquels les pixels sont parcourus dans 4 différentes directions, dans le but de réduire l'asymétrie introduite par chaque itération de SOR. Ce problème d'asymétrie n'est pas visible dans le cas de l'estimation du flot optique seul, mais tendait à induire des ondes orientées dans la solution numérique du flot de scène.

Les conditions d'arrêt pour les itérations de point fixe sont estimées à partir des normes L^2 relatives entre des incréments consécutifs. Nous avons utilisé 0.05 comme condition d'arrêt pour les itérations de bas niveau et 0.01 pour celles de haut niveau. Une fois que la minimisation de E est effectuée à un niveau de résolution donné, la solution est multipliée par $1/\eta$, suréchantillonnée au niveau de résolution suivant, et le même processus est itéré jusqu'à ce que la résolution finale des images soit atteinte.

3.3 L'estimation des occultations

Les occultations sont gérées en estimant les fonctions β_{fl}, β_{fr}, β_{st}, β_s au début de chaque itération de point fixe de haut niveau. Nous prenons ainsi en compte dans cette estimation des occultations chaque mise à jour de la solution $(u, v, d, d^{'})$ au cours de l'optimisation. Les fonctions β valent 1 pour les pixels visibles et 0 pour les pixels occultés, de telle sorte que pour les pixels occultés partout, seul le terme de régularisation est pris en compte.

Nous décrivons les étapes nécessaires a l'estimation de la fonction β_s associée à la mise en correspondance stéréo à la date t, les autres fonctions étant évaluées similairement :

- Nous mappons d'abord la disparité d vers l'image de droite $I_r(., t)$ en utilisant le Z buffering.
- Nous remappons la carte de disparité obtenue vers l'image de gauche $I_l(., t)$, et nous ajoutons un seuil de tolérance (1.5 pixel) à la disparité remappée ainsi obtenue.
- Nous calculons la carte d'occultations en comparant d avec la disparité remappée associée à son seuil de tolérance.

La mise à jour de E_{data} pour chaque pixel de l'image de référence $I_l(., t)$ peut alors être effetuée, connaissant les occultations pour chaque composante de E_{data}.

3.4 Algorithme complet

Le problème étant fortement non linéaire et non convexe, l'algorithme de résolution doit, dans certains cas, être soigneusement initialisé afin d'éviter des minima locaux éventuellement présents au niveau de résolution le plus grossier. Dans le cas de l'estimation du flot optique [4], et spécifiquement lorsqu'on utilise un algorithme multirésolution, la résolution la plus grossière peut être aussi petite que possible, et le flot optique est habituellement initialisé à 0. La raison de ce choix est que le flot optique est généralement petit rapporté aux dimensions des images, et il est aisé de trouver une résolution d'image de départ telle que l'amplitude maximale du flot optique sous-échantillonné soit inférieure à 0.5 pixel, ce qui est dans la grande majorité des cas suffisant pour assurer la convergence vers un minimum global.

Dans le problème du flot de scène, nous avons un problème mixte : si nous considérons chaque caméra séparément, celà est semblable au problème du flot optique, mais nous essayons de résoudre simultanément un problème de mise en correspondance stéréo entre les images de gauche et de droite. Les caractéristiques d'un problème de stéréo sont très différentes de celles du flot optique : l'amplitude de la disparité est généralement comparable à la taille des images (et même plus grande que la taille des objets de la scène vus dans les images), et de plus il existe beaucoup de zones occultées. Pour cette raison, les approches multirésolution sont peu efficaces sur les problèmes stéréo si elles démarrent l'optimisation à un niveau de résolution trop grossier. Notre méthode souffrirait également de cet inconvénient dans cette situation.

Par conséquent, nous avons choisi de démarrer l'estimation du flot de scène à un niveau de résolution intermédiaire, et d'initialiser les inconnues (u, v, d, d') avec des valeurs judicieuses. Tout d'abord, dans les cas où il n'est pas possible d'initialiser la stéréo à 0, nous initialisons d avec un algorithme de l'état de l'art [8] qui calcule la disparité aux images de pleine résolution (niveau 1 des pyramides d'images). L'erreur commise sur la disparité d'un algorithme stéréo donné peut être aisément estimée en utilisant les benchmarks standards [16], et nous estimons ensuite le niveau de pyramide b tel que l'erreur nominale en disparité sous-échantillonnée se situe en dessous de 0.5 pixels. Nous estimons aussi un niveau de pyramide a, plus grand que b (correspondant donc à un niveau de résolution plus grossier), tel que l'amplitude maximale du flot optique à ce niveau soit en dessous de 0.5 pixels. Nous résolvons alors le problème du flot optique (en gardant les termes de eq. (1) en relation avec le flot optique) séparément sur les séquences gauche et droite, du niveau a au niveau b, et nous obtenons une estimation initiale pour les flots optiques gauche (u, v) et droite (u', v'). La disparité d à t est alors affinée entre un certain niveau de pyramide c et le niveau b (c supérieur ou égal à b)., utilisant la même méthode et gardant seulement les termes de E relatifs à la stéréo à la date t. $d^{'}$ est initialisée en additionnant à d la différence entre u' et u, et en warpant le résultat vers I_l au temps t (les détails sont donnés dans Algorithme 1). Enfin, l'algorithme d'estimation du flot de scène est appliqué aux 4 images, du

niveau de pyramide b au niveau 1 de la pyramide (résolution originale des images). L'algorithme d'estimation du flot de scène, incluant la phase d'initialisation, est détaillé dans Algorithme 1.

Algorithme 1 Estimation du flot de scène

SORTIES: Calculer le flot de scène (u, v, d, d') entre t et $t+1$ en utilisant des pyramides stéréo (chaque pyramide possède a niveaux)

ENTRÉES: $a, b, c \in \mathbb{N}, a > b >= 1, a > c >= b >= 1$

$u \leftarrow 0$, $v \leftarrow 0$, $u' \leftarrow 0$, $v' \leftarrow 0$

pour $l = a$ à b **faire**

$(u, v) \leftarrow$ flot optique gauche (u, v) au niveau l

$(u', v') \leftarrow$ flot optique droit (u', v') au niveau l

fin pour

$d \leftarrow$ stéréo avec [8]

pour $l = c$ à b **faire**

$d \leftarrow$ disparité d au temps t au niveau l

fin pour

$d'(\mathbf{x} + (u, v)) \leftarrow d + u'(\mathbf{x} + d) - u$

pour $l = b$ à 1 **faire**

$(u, v, d, d') \leftarrow$ flot de scène (u, v, d, d') au niveau l

fin pour

4 Résultats et évaluation

Alors qu'il existe de nombreuses séquences d'images de validation avec vérité terrain pour divers algorithmes en vision par ordinateur, le problème du flot de scène ne bénéficie pas encore de benchmarks reconnus. Néanmoins, de telles bases de données de validation existent pour les sous-problèmes du flot de scène : le flot optique et la stéréo.

Le benchmark standard pour le flot optique est la séquence Yosemite, une séquence de vol sur un paysage rendu par ray tracing, avec les vérités terrain en flot optique et en profondeur. Malheureusement, pour le moment la séquence n'est associée qu'à une seule caméra. Une séquence obtenue par une seconde caméra pourrait être générée en utilisant l'information de profondeur pour warper la première séquence d'images, mais la qualité ne serait pas très bonne et toutes les zones occultées seraient manquantes.

Pour le problème stéréo, plusieurs séquences de validation sont disponibles, chacune consistant en 8 vues d'une même scène, dans lesquelles tous les centres optiques sont alignés et espacés régulièrement, et les images sont rectifiées [16]. Il se trouve que ces images peuvent être utilisées pour évaluer un algorithme de flot de scène, en imaginant 2 caméras rectifiées observant une scène statique, et translatées dans la direction du segment joignant leurs centres optiques. Toutes ces images sont présentes dans les bases de données de validation stéréo. Néanmoins, elles représentent des configurations spéciales pour le flot de scène, puisque la composante flot optique est rigoureusement horizontale ($v = 0$), et la disparité reste constante ($d' = d$), mais puisque notre algorithme n'a aucune connaissance à priori de la nature de ce mouvement, celà reste encore un bon benchmark. Nous avons pris les images 2 et 6 des séquences Venus, Teddy et Cones comme paire stéréo à la date t, et les images 4 et 8 comme paire stéréo au temps $t+1$. La vérité terrain est donnée pour la disparité entre les images 2 et 6, et la vérité terrain du flot optique est celle de la disparité divisée par 2.

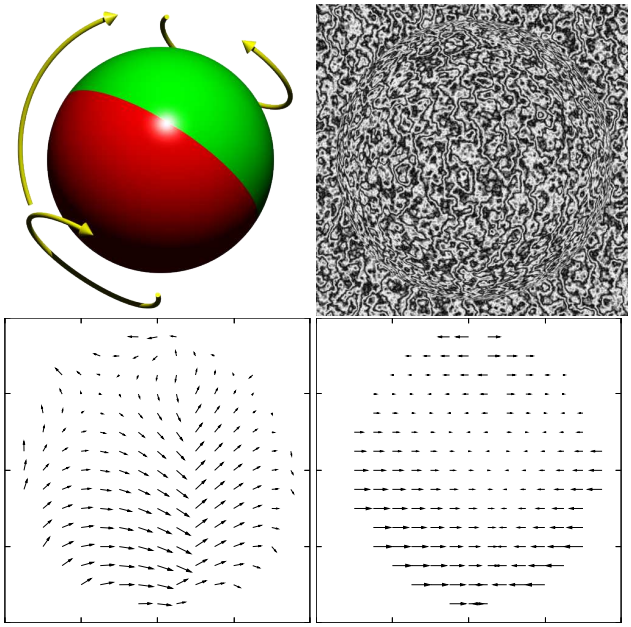

FIG. 2 – La scène synthétique présentée est une sphère texturée en rotation, pour laquelle les deux hémisphères tournent indépendamment dans des sens opposés (haut-bas, image en haut a droite). La reconstruction 3D est invariante par rotation. L'information du mouvement 3D est mesurable seulement à partir du flot de scène : (u, v) (en bas à gauche) and $d' - d$ (en bas à droite) montrent une discontinuité du flot de scène le long du méridien vertical.

Afin d'évaluer notre algorithme sur un flot de scène plus général, nous avons également généré des images synthétiques d'une sphère en rotation (Fig. 2). Cette scène représente un cas extrême où une reconstruction 3D ne donnera aucune information sur ce qu'il se passe dans la scène, et toute l'information est contenue dans le flot de scène : puisque la sphère est en rotation, la reconstruction reste identique au fil du temps. En outre, les hémisphères tournent dans des sens opposés, ce qui génère une forte discontinuité dans le flot de scène, et nous vérifions que l'algorithme est capable de retrouver cette discontinuité avec précision.

L'évaluation de la précision de l'algorithme est réalisée en estimant l'erreur RMS (root mean square) pour les 4 cartes de valeurs de u, v, d, d'. Les précisions des cartes de flot optique (u, v) sont évaluées ensemble, et les cartes de disparité sont évaluées séparément : bien qu'exprimées également en pixels, les disparités sont plus difficiles à mesurer du fait de leur plage de valeur et de la présence d'occul-

Dataset	(u, v)	d	d'
Venus	0.31	0.97	1.48
Teddy	1.25	2.27	6.93
Cones	1.11	2.11	5.24
Sphere	0.69	3.73	3.81

FIG. 3 – Erreur RMS en pixels sur les 4 composantes du flot de scène estimées par notre algorithme sur les différents jeux de données de validation.

Dataset	μ_{of}	σ_{of}	μ_{sf}	σ_{sf}
Venus	1.06	1.17	0.98	0.91
Teddy	0.43	0.49	0.51	0.66
Cones	0.66	1.21	0.69	0.77
Sphere	1.50	5.65	1.75	6.07

FIG. 4 – Espérance μ_{sf} et écart type σ_{sf} de l'erreur angulaire en degrés de la composante flot optique (u, v) du flot de scène, comparées a l'erreur angulaire (μ_{of}, σ_{of}) du flot optique estimé séparément.

tations. Les résultats de nos évaluations sont montrés en Fig. 3. Fig. 4 compare l'erreur angulaire de la composante flot optique du flot de scène avec le flot optique estimé en utilisant notre méthode ou [4]. Fig. 5 expose les valeurs finales de u, v, d et d' pour l'exemple de la sphère, montrant que la discontinuité a été correctement retrouvée par notre algorithme, et montrant les cartes d'occultation générées. Figures 6 et 7 présentent des résultats sur une séquence stéréo réelle représentant une scène non rigide. Dans cette séquence sont présentes des discontinuités en mouvement et en stéréo ainsi que des variations d'illumination.

5 Conclusion

Dans ce papier, nous avons présenté une méthode variationnelle pour estimer le flot de scène et les occultations associées simultanément à partir de séquences stéréoscopiques. Cette méthode couple l'estimation du flot optique avec la mise en correspondance stéréo dense, en minimisant une énergie construite à partir des relations entre les images des séquences stéréo. Notre algorithme gère également les discontinuités de la géométrie 3D et du champ de déplacement 3D, et est robuste aux variations d'illumination.

Notre algorithme généralise le travail de Brox *et al.* [4] sur l'estimation précise du flot optique, en ajoutant des contraintes issues de la géométrie épipolaire, et nous avons montré que le même type de schéma numérique peut être utilisé pour résoudre les deux problèmes. Néanmoins, la nature de la disparité est différente de celle du flot optique, dans le sens où les occultations sont plus larges, et que la plage de valeur en disparité est comparable à la taille des objets dans les images, ce qui cause beaucoup de difficultés à de nombreux algorithmes stéréo multirésolution. Nous avons ainsi proposé un algorithme qui, dans les cas difficiles pour la stéréo, évalue tout d'abord les composantes du flot de scène par une estimation séparée puis raffine cette estimation par une estimation conjointe de toutes les omposantes. Il s'agit du premier travail sur le flot de scène qui présente une évaluation quantitative de celui-ci, en comparant la composante de flot optique de notre flot de scène (en quelque sorte la projection dans les images de celui-ci) avec les résultats de la méthode variationnelle d'estimation du flot optique la plus précise à notre connaissance. De plus, nos expériences ont également démontré que l'algorithme présenté est capable de traiter des séquences stéréo réelles difficiles (mouvement non rigide, discontinuités en mouvement et en stéréo, variations d'illumination).

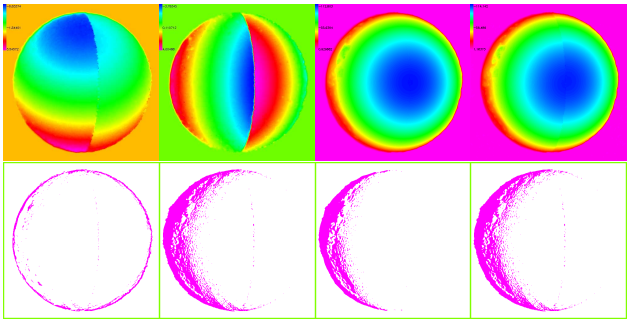

FIG. 5 – En haut : les valeurs u, v, d, d' estimées pour le cas de la sphère. ($-7 < u < 4$, $-4 < v < 4$, $-113 < d < 1$, $-115 < d' < 2$). Notez la discontinuité verticale dans d', due au fait que les coordonnées de référence sont celles de l'image de gauche au temps t. En bas : les cartes d'occultation pour les termes d'attache aux données correspondant au flot optique gauche, au flot optique droit, et aux disparités aux temps t et $t + 1$.

Dans un futur proche, nous espérons obtenir une preuve mathématique de la convergence de l'algorithme, et nous travaillons également sur des améliorations de la rapidité de l'algorithme, en adaptant des travaux récents sur les méthodes variationnelles de flot optique en temps réel [5]. En outre, nous souhaiterions modéliser les cartes d'occultations par des fonctions déterministes continues (coefficients β). Des travaux précédents utilisent une formulation probabiliste [19, 18], mais une formulation déterministe et continue pourrait être intégrée plus naturellement à notre formulation variationnelle.

FIG. 6 – Un exemple avec des données réelles (images de 854×854 pixels). L'intervalle de temps entre les paires stéréo du haut et du bas est de 1.5s. On observe des variations d'illumination, de grands mouvements (en translation et rotation), et une claire discontinuité du mouvement dans la région de la bouche. Les plages de valeurs en pixel pour les composantes du flot de scène sur cet exemple sont $u \in [-131, 1]$, $v \in [-49, 33]$, $d, d' \in [-122, -39]$.

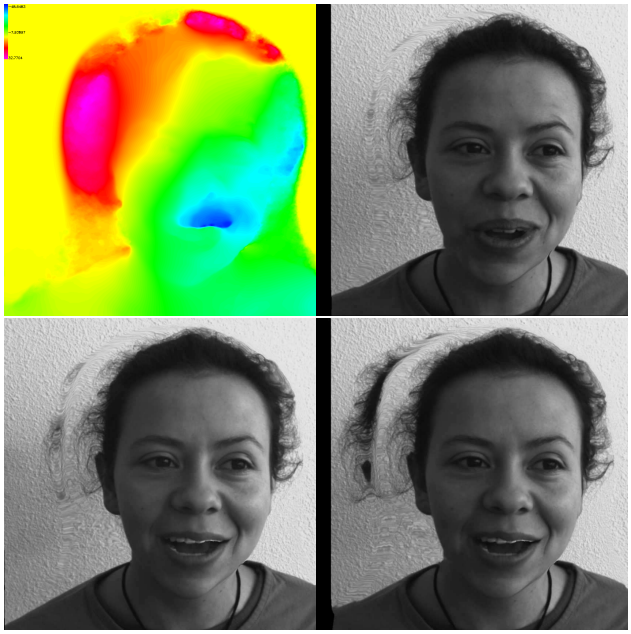

FIG. 7 – Résultats sur des données réelles : en haut à gauche, la composante flot optique verticale v du flot de scène montre clairement la reconstruction de la discontinuité créée par la bouche. L'image de droite au temps 0 (en haut a droite) et la paire stéréo à 1.5s ont été warpées vers l'image de gauche à l'instant 0, montrant où le flot de scène a été correctement estimé.

Références

[1] G. Aubert and P. Kornprobst. A mathematical study of the relaxed optical flow problem in the space BV(ω). *SIAM J. Math. Anal.*, 30(6) :1282–1308, 1999.

[2] J.L. Barron, D.J. Fleet, S.S. Beauchemin, and T.A. Burkitt. Performance of optical flow techniques. In *Proc. IEEE CVPR*, pages 236–242, 1992.

[3] B.Horn and B.Schunck. Determining optical flow. *Artifical Intelligence*, 17 :185–203, 1981.

[4] A. Brox, N. Bruhn, J. Papenberg, and T. Weickert. High accuracy optical flow estimation based on a theory for warping. In *Proc. 8th ECCV*, volume 3024 of *LNCS*, pages 25–36, Prague, Czech Republic, 2004. Springer-Verlag.

[5] A. Bruhn, J. Weickert, T. Kohlberger, and C. Schnorr. Discontinuity preserving computation of variational optic flow in real-time. In *ScaleSpace05*, pages 279–290, 2005.

[6] Andrés Bruhn, Joachim Weickert, Christian Feddern, Timo Kohlberger, and Christoph Schnörr. Variational optical flow computation in real time. *IEEE Trans. Image Processing*, 14(5) :608–615, 2005.

[7] Rachid Deriche, Pierre Kornprobst, and Gilles Aubert. Optical-flow estimation while preserving its discontinuities : A variational approach. In *Proc. ACCV*, pages 71–80, 1995.

[8] Pedro F. Felzenszwalb and Daniel P. Huttenlocher. Efficient belief propagation for early vision. *International Journal of Computer Vision*, 70(1), October 2006.

[9] Minglun Gong and Yee-Hong Yang. Disparity flow estimation using orthogonal reliability-based dynamic programming. In *Proc. 18th ICPR*, pages 70–73. IEEE, 2006.

[10] Frédéric Huguet and Frédéric Devernay. A variational method for scene flow estimation from stereo sequences. Research Report 6267, INRIA, August 2007.

[11] M. Isard and J.P. MacCormick. Dense motion and disparity estimation via loopy belief propagation. In *ACCV06*, pages II :32–41, 2006.

[12] Martin Lefébure and Laurent D. Cohen. Image registration, optical flow and local rigidity. *J. Math. Imaging Vis.*, 14(2) :131–147, 2001.

[13] D. Min and K. Sohn. Edge-preserving simultaneous joint motion-disparity estimation. In *ICPR06*, pages II : 74–77, 2006.

[14] J.-P. Pons, R. Keriven, O. Faugeras, and G. Hermosillo. Variational stereovision and 3D scene flow es-

timation with statistical similarity measures. In *Proc. IEEE ICCV*, page 597, 2003.

[15] Jean-Philippe Pons, Renaud Keriven, and Olivier Faugeras. Modelling dynamic scenes by registering multi-view image sequences. In *Proc. IEEE CVPR*, volume 2, pages 822–827, 2005.

[16] Daniel Scharstein and Richard Szeliski. A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. *International Journal of Computer Vision*, 47(1-3) :7–42, 2002.

[17] N. Slesareva, A. Bruhn, and J. Weickert. Optic flow goes stereo : A variational method for estimating discontinuity-preserving dense disparity maps. In *DAGM05*, page 33, 2005.

[18] C. Strecha, R. Fransens, and L.J. Van Gool. Wide-baseline stereo from multiple views : A probabilistic account. In *Proc. IEEE CVPR*, volume 1, pages 552–559, 2004.

[19] Christoph Strecha, Rik Fransens, and Luc J. Van Gool. A probabilistic approach to large displacement optical flow and occlusion detection. In *ECCV Workshop SMVP*, pages 71–82, 2004.

[20] S.Vedula, S.Baker, P.Rander, R.Collins, and T.Kanade. Three-dimensional scene flow. In *Proc. IEEE ICCV*, pages 722–729, 1999.

[21] Sundar Vedula and Simon Baker. Three-dimensional scene flow. *IEEE Trans. Pattern Anal. Mach. Intell.*, 27(3) :475–480, 2005.

[22] Y. Zhang and C. Kambhamettu. Integrated 3D scene flow and structure recovery from multiview image sequences. In *Proc. IEEE CVPR*, pages II : 674–681, 2000.

[23] Ye Zhang and Chandra Kambhamettu. On 3D scene flow and structure estimation. In *Proc. IEEE CVPR*, page 778, 2001.