Estimation de cartes de profondeurs basée sur un maillage triangulaire d'une paire stéréo

Brendan LE BOUILL[1,2], Yannick BERTHOUMIEU[1], Jean-François AUJOL[2], Charles-Alban DELEDALLE[2]

[1]Univ. Bordeaux, IMS, UMR 5218, F-33400 Talence, France, CNRS, UMR 5218, F-33400 Talence, France

[2]Univ. Bordeaux, IMB, UMR 5251, F-33400 Talence, France, CNRS, UMR 5251, F-33400 Talence, France

{brendan.lebouill, yannick.berthoumieu}@ims-bordeaux.fr
{jean-francois.aujol, charles-alban.deledalle}@math.u-bordeaux.fr

Résumé – La reconstruction d'environnement 3D par stéréovision est un problème complexe étudié depuis plusieurs décennies. Dans cet article, nous introduisons une méthode d'estimation de cartes de profondeurs basée sur un maillage triangulaire du domaine de l'image. Un modèle plan est associé à chaque triangle afin de maximiser la cohérence 3D de la scène reconstruite.

Abstract – 3D reconstruction by stereovision is a complex problem studied for decades. In this article, we introduce a depth map estimation method based on a triangulation of the image domain. A slanted plane is fitted to each triangle in order to maximize the 3D consistency of the reconstructed scene.

1 Introduction

La reconstruction d'environnement 3D par stéréovision est étudiée depuis plusieurs décennies par la communauté scientifique. Elle possède de nombreuses applications : de la réalité augmentée à la navigation autonome de véhicules ou de robots.

Plusieurs types d'approches permettent d'aborder le problème de la stéréovision. Elles sont divisées en deux classes par Scharstein et al. [10] : les méthodes locales et les méthodes globales. Pour les méthodes locales, le calcul de la disparité d'un pixel est effectué sur un support de taille fini (pixel, patch). Un terme de régularisation implicite permet d'améliorer les performances de ces algorithmes. L'utilisation de patchs ou de descripteurs définis sur un voisinage du pixel permettent un appariement plus robuste entre les images, bien que cela induise des problèmes au niveau des discontinuités. Contrairement aux méthodes locales, les méthodes globales modélisent de manière explicite les hypothèses de régularité du champ des disparités. Parmi les algorithmes les plus performants, on distingue les méthodes qui utilisent une régularisation du premier ou du second ordre [5][9], qui sont basées sur un support pixelique, de celles utilisant une segmentation de l'image [2][4]. Beaucoup d'approches récentes sont basées sur une (sur-)segmentation de l'image et associent un modèle 3D explicite à chaque segment de l'image, traditionnellement un plan. L'espace de recherche correspondant à une paramétrisation des plans 3D étant de grande dimension, de multiples auteurs se sont intéressés à différents types de discrétisation ou de subdivision de cet espace. Les méthodes de type *plane-sweeping* propagent des plans fronto-parallèles au plan image [6]. Cette simplification permet de calculer de manière rapide et efficace les reprojections entre images, ces plans ayant une disparité constante sur l'ensemble du segment. Ce type de modèle décrit cependant de manière imparfaite les scènes naturelles. La subdivision initialement proposée par Birchfield et al. [2] consiste à identifier les orientations principales présentes dans la scène observée. Seules ces orientations sont considérées lors de l'optimisation [8]. Pour les scènes urbaines, l'hypothèse de *Manhattan world* effectuée par Furukawa et al. [7] permet d'obtenir une reconstruction cohérente de l'environnement. D'autres modèles surfaciques associés aux segments de l'image ont été proposés, comme les *B-splines* par Bleyer et al. [3].

Certaines méthodes optimisent le plan associé à chaque segment de l'image sur l'espace continu des plans afin de maximiser la cohérence tridimensionnelle de la scène reconstruite et potentiellement augmenter la précision de la carte des disparités obtenue. Ces méthodes s'appuient généralement sur une décomposition en superpixels de l'image en utilisant l'algorithme SLIC [1], et associent un plan incliné à chaque superpixel [4][12]. Yamaguchi et al. [12] utilisent un champ de Markov hybride qui modélise explicitement les relations 3D entre superpixels à l'aide d'une variable discrète. Ce MRF considère une carte des disparités obtenue en utilisant l'algorithme SGM [9] ainsi qu'une décomposition en superpixels de l'image afin de maximiser la cohérence de la scène reconstruite.

L'approche présentée dans cet article propose de minimiser l'erreur de reprojection entre les deux images d'une paire stéréo plutôt qu'une régularisation basée uniquement sur la carte des disparités initialement estimée. Les estimations aberrantes peuvent ainsi être corrigées par notre méthode, contrairement à l'approche [12]. Une contrainte sur la forme des segments de l'image est appliquée. La forme libre des

superpixels augmente la complexité calculatoire des termes de régularisation car l'ensemble des pixels de la frontière doit être parcouru. De plus, l'absence de contrainte sur la forme des superpixels est par définition incompatible avec la modélisation de ces derniers par des plans. En effet, deux plans non parallèles de l'espace 3D s'intersectent en une ligne. Notre méthode génère un maillage triangulaire de l'image, qui est projeté en 3D pour minimiser une fonctionnelle. Les paramètres des plans sont optimisés sur un espace continu, contrairement à la méthode [6] qui ne considère que des plans fronto-parallèles à la caméra. La méthode la plus proche de la notre est celle introduite par Zhang et al. [13]. Notre modèle est cependant plus simple et une paramétrisation différente nous permet de régulariser de manière homogène la courbure du pseudo-maillage généré. En effet, la paramétrisation *normal-depth* adoptée par [13] privilégie les plans fronto-parallèles.

Une première contribution de cet article est l'introduction d'une méthode de génération d'un maillage triangulaire du domaine de l'image. Cela permet une modélisation moins complexe des discontinuités entre patchs triangulaires lors de la résolution du problème de stéréovision. La seconde contribution concerne la méthode d'estimation de carte des disparités. Les triangles du maillage obtenus sont modélisés par des plans 3D qui sont ensuite projetés sur l'axe optique afin d'obtenir une carte de disparités minimisant l'erreur de reprojection entre les deux images de la paire stéréo. La cohérence 3D de la scène observée est assurée grâce à deux termes de régularisation qui encouragent la continuité du pseudo-maillage ainsi que la coplanarité des triangles voisins.

Une première partie présente l'algorithme de génération du maillage triangulaire du domaine de l'image introduit. La seconde partie de cet article détaille l'algorithme d'estimation de carte des disparités basé sur le maillage triangulaire précédemment généré.

2 Génération d'un maillage triangulaire du domaine de l'image

La décomposition d'une image en triangles, ou triangulation, a fait l'objet de nombreuses études, notamment dans le cadre de la compression d'image. L'objectif est de construire un réseau triangulaire afin de minimiser une erreur de mesure entre l'image originale et l'image approximée. Celle-ci est interpolée à partir des valeurs d'intensité des sommets des triangles, parfois couplée à une équation définissant les variations d'intensité le long des arêtes des triangles. Deux types d'algorithmes de génération d'un maillage triangulaire du domaine de l'image ont été introduits par la communauté scientifique. Le premier type se fonde sur la détection de point d'intérêts. Une triangulation de Delaunay est alors générée à partir de ce jeu de points. Le maillage triangulaire est parfois densifié en ajoutant des sommets dans les zones

FIGURE 1 – Maillage triangulaire d'une image de la paire *Flowerpots*

homogènes. La seconde catégorie de méthodes est basée sur la décomposition en superpixels de l'image, qui sont polygonisés, puis divisés en triangles.

Dans cet article, l'image I est décomposée en un ensemble T de N_T triangles. Chacun des triangles T_m est défini par ses trois sommets $V_m = \{v_{m,i}\ ;\ i \in [\![1,3]\!]\}$. L'ensemble des sommets du maillage triangulaire est noté V et est de dimension N_V. A la manière de l'algorithme de sur-segmentation SLIC, on souhaite générer un maillage qui adhère aux contours de l'image et dont les segments sont homogènes en taille. Pour ce faire, nous proposons de minimiser itérativement l'énergie suivante.

$$\min_V \sum_{m=1}^{N_T} E_{\text{color}}(I, V_m) + \lambda \sum_{m=1}^{N_T} E_{\text{shape}}(V_m) \tag{1}$$

$$E_{\text{color}}(I, V_m) = \frac{1}{N_m} \sum_{p \in T_m} \|I(p) - c_m\|_2^2 \tag{2}$$

$$E_{\text{shape}}(V_m) \ = \ \left\| 1 - \frac{\text{area}_m}{\text{area}_{\text{ref}}} \right\|_1 \ , \tag{3}$$

où c_m et N_m sont respectivement le vecteur couleur moyen et le nombre de pixels du triangle T_m, et area_{ref} est l'aire d'un triangle d'un maillage triangulaire régulier de l'image.

La stratégie algorithmique utilisée par l'algorithme SLIC est incompatible avec une triangulation du domaine de l'image. En effet, la contrainte géométrique sur la forme des segments que l'on souhaite générer, ici des triangles, force l'algorithme d'optimisation à considérer l'ensemble du segment et non chaque pixel indépendamment.

A partir d'un maillage régulier initial de l'image, composé de triangles équilatéraux (excepté aux bords), la fonctionnelle (1) est itérativement minimisée. Pour chaque sommet du maillage, l'énergie est évaluée en considérant un mouvement de ce sommet dans son 4-voisinage. Le déplacement correspondant à l'énergie minimale est accepté. Puisque seuls les déplacements d'un pixel sont considérés, les pixels échangés entre triangles peuvent être obtenus de manière efficace en parcourant les arêtes induites par le sommet à sa position initiale et à sa position considérée. La Figure 1 donne un exemple de maillage triangulaire obtenu par la méthode présentée dans cette partie.

3 Projection tridimensionnelle du maillage triangulaire de l'image

3.1 Paramétrisation

La plupart des algorithmes de stéréovision paramétrisent le plan associé à chaque segment de l'image par une relation de la forme $\mathrm{disp}(\mathbf{p}) = ap_x + bp_y + c$ [3][12]. Les trois paramètres (a, b, c) ne sont cependant pas homogènes, ce qui peut poser des problèmes d'instabilité selon la stratégie d'optimisation utilisée. En partant de la paramétrisation proposée par Silveira et al. [11] et en tirant parti de la forme particulière des supports de l'image considérés pour la reprojection, les plans sont définis par les disparités des sommets des triangles. Le vecteur $\mathbf{d_m} = (d_{m,i_0}, d_{m,i_1}, d_{m,i_2})^T$ définit le plan associé au triangle T_m, lui même formé par les sommets V_m. Les trois sommets V_m formant un triangle non dégénéré, ces trois points sont non colinéaires. Ainsi, on peut calculer la disparité associée au pixel $\mathbf{p}$ appartenant au plan associé au triangle T_m par :

$$\mathrm{disp}(\mathbf{p}, m) = \mathbf{d_m}^T S_m^{-1} \tilde{\mathbf{p}} \ , \tag{4}$$

où $\tilde{\mathbf{p}} = (p_x, p_y, 1)^T$ est la position du pixel $\mathbf{p}$ en coordonnées homogènes, et $S_m = (\ \tilde{\mathbf{p}}_{\mathbf{i_0}}\ \tilde{\mathbf{p}}_{\mathbf{i_1}}\ \tilde{\mathbf{p}}_{\mathbf{i_2}})$ la matrice contenant les positions des pixels en coordonnées homogènes associées aux trois sommets du triangle T_m.

De plus, la relation (5) permet de calculer le vecteur normal normalisé $\bar{\mathbf{n}}_\mathbf{m}$ du plan associé au triangle T_m :

$$\bar{\mathbf{n}}_\mathbf{m} = \frac{\Omega_m}{\|\Omega_m\|} \ \text{avec} \ \Omega_m^T = \mathbf{d_m}^T S_m^{-1} K \ , \tag{5}$$

où K est la matrice de calibration de la caméra.

3.2 Fonctionnelle minimisée

On souhaite trouver les disparités en chaque sommet qui minimisent l'erreur de reprojection entre les deux images de la paire stéréo tout en maximisant la cohérence tridimensionnelle de la scène reconstruite. Pour cela, la fonctionnelle suivante est minimisée :

$$\begin{aligned} E_{\mathrm{stereo}}(I_L, I_R, T, D) \ &= E_{\mathrm{data}}(I_L, I_R, T, D) \\ &+ \alpha\, E_{\mathrm{split}}(T, D) \\ &+ \beta\, E_{\mathrm{normal}}(T, D) \ , \end{aligned} \tag{6}$$

où I_L et I_R sont les images rectifiées de la paire stéréo, T le maillage triangulaire (fixé) de l'image I_L et D l'ensemble des profondeurs des sommets des triangles.

Terme d'attache aux données. Ce terme pénalise les erreurs de reprojection entre les images de la paire stéréo I_L et I_R :

$$E_{\mathrm{data}}(I_L, I_R, T, D) = \sum_{m=1}^{N_T} \sum_{\mathbf{p} \in T_m} \left\| I_R \begin{pmatrix} p_x - \mathrm{disp}(\mathbf{p}, \mathbf{d_m}) \\ p_y \end{pmatrix} - I_L \begin{pmatrix} p_x \\ p_y \end{pmatrix} \right\|_2^2 \ . \tag{7}$$

Régularisation des discontinuités. On souhaite obtenir une reconstruction de la scène observée la plus continue possible, tout en préservant les discontinuités :

$$E_{\mathrm{split}}(T, D) = \sum_{i=1}^{N_V} \sum_{\substack{(m_1,m_2) \\ \in \mathrm{Vois}_i}} \omega_{m_1,m_2} \left\| d_{m_1,i} - d_{m_2,i} \right\|_2^2 \ , \tag{8}$$

où Vois_i est l'ensemble des triangles induits par le sommet i, et ω_{m_1,m_2} une pénalisation de la régularisation permettant de préserver les discontinuités

$$\omega_{m_1,m_2} = \exp\left(-\gamma f_{m_1,m_2} \left(\|\nabla I_L\|_2^2 \right) \right) \ \text{avec} \ \gamma > 0 \ . \tag{9}$$

La fonction $f_{m_1,m_2}\left(\|\nabla I_L\|_2^2\right)$ calcule la norme du gradient spatial moyen sur l'arête qui sépare les triangles m_1 et m_2.

Régularisation sur les normales. Enfin, on fait l'hypothèse que la courbure entre plans voisins est faible. Cette hypothèse permet de propager l'information entre plans voisins dans les zones homogènes, pour lesquelles le terme d'attache aux données n'est pas informatif :

$$E_{\mathrm{normal}}(T, D) = \sum_{j=1}^{N_E} \sum_{\substack{(m_1,m_2) \\ \in \mathrm{Vois}_j}} \omega_{m_1,m_2} \left\| \frac{1 - \bar{\mathbf{n}}_{\mathbf{m_1}}^T \bar{\mathbf{n}}_{\mathbf{m_2}}}{1 + \bar{\mathbf{n}}_{\mathbf{m_1}}^T \bar{\mathbf{n}}_{\mathbf{m_2}}} \right\|_2^2 , \tag{10}$$

où N_E est le nombre d'arêtes de T et Vois_j le couple de triangles qui partagent l'arête j.

3.3 Expérimentations et résultats

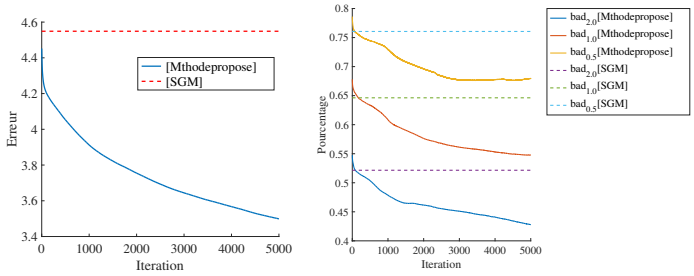
La fonctionnelle (7) est minimisée en utilisant une descente de gradient projetée. Une initialisation de la carte des disparités est obtenue en utilisant l'implémentation de l'algorithme SGM fournie par [5]. Le plan initial associé à chaque triangle est obtenu par moindres carrés sur cette initialisation. Les disparités étant nécessairement positives, elles sont projetées sur $\mathbb{R}^+$ à chaque pas de descente, ainsi la mise à jour de D se lit :

$$\underset{(\mathbb{R}^+)^{3N_T}}{\mathrm{proj}} \left[D - \delta \left(\frac{\partial E_{\mathrm{data}}}{\partial D} + \alpha \frac{\partial E_{\mathrm{split}}}{\partial D} + \beta \frac{\partial E_{\mathrm{normal}}}{\partial D} \right) \right] . \tag{11}$$

Le jeu de données Middlebury 2006 [10] a été utilisé pour valider l'approche proposée. Quatre critères d'évaluation ont été choisis pour valider les performances de notre méthode. L'erreur de disparité moyenne des pixels non occultés donne une indication sur la qualité globale de la reconstruction. Les indicateurs bad_ϵ donnent le pourcentage de pixels non occultés ayant une erreur de disparité inférieure à ϵ pixels. Le Tableau 1 montre le gain de performances obtenu par notre approche par rapport à l'initialisation avec SGM. On constate que l'erreur moyenne obtenue par notre approche est systématiquement inférieure à celle de l'algorithme SGM, ce qui indique que la reconstruction obtenue est plus cohérente. Cependant, dans le cas de la paire *Lampshade1*, les critères $\mathrm{bad}_{1.0}$ et $\mathrm{bad}_{0.5}$ sont faiblement dégradés par notre approche, ce qui peut être dû à la segmentation imparfaite de l'image ou au modèle plan associé aux segments de l'image, qui est parfois inadapté. La Figure 2

TABLE 1 – Comparaison des performances de l'algorithme SGM [9] et la méthode proposée (Meth. Propos.)

	μ_{erreur}	$\text{bad}_{2.0}$	$\text{bad}_{1.0}$	$\text{bad}_{0.5}$	Images
SGM	4.5488	0.5216	0.6463	0.7605	*Flowerpots*
Meth. Propos.	**3.4990**	**0.4279**	**0.5478**	**0.6797**	*Flowerpots*
SGM	7.7424	0.4910	**0.5591**	**0.6453**	*Lampshade1*
Meth. Propos.	**5.1231**	**0.4705**	0.5718	0.7007	*Lampshade1*
SGM	4.8894	0.1810	0.2478	0.4055	*Wood1*
Meth. Propos.	**4.0299**	**0.1766**	**0.2206**	**0.3895**	*Wood1*



(a) (b)

FIGURE 2 – Évolution de l'erreur moyenne (a) et des indicateurs bad_ϵ (b) en fonction du nombre d'itérations pour la paire *Flowerpots*

détaille l'évolution des critères estimés en fonction du nombre d'itérations de la descente de gradient. Pour ces exemples, environs 2000 triangles décomposaient l'image et les paramètres étaient fixés aux valeurs suivantes : $\alpha = 10^{-3}$, $\beta = 10^{-2}$, $\delta = 10^{1}$, $\gamma = 10^{-3}$ et $\lambda = 600$. La Figure 3 représente les reconstructions 3D obtenues par l'algorithme SGM et par notre méthode. Comme l'indiquaient les performances du Tableau 1, la cohérence 3D de la scène est améliorée par notre méthode.

4 Conclusion

La méthode proposée permet de générer une reconstruction cohérente de la scène observée par une paire stéréo. La modélisation explicite des relations entre plans tridimensionnels des supports du maillage triangulaire de l'image reprojeté permettent d'obtenir une meilleure reconstruction, notamment dans les zones homogènes en texture.

Bien que donnant des résultats satisfaisants, notre méthode, sujette à d'éventuels minima locaux, reste très liée à la qualité de l'initialisation. Ainsi, d'autres stratégies d'optimisation de notre modèle sont à envisager. En particulier, une stratégie pyramidale sera étudiée pour la génération du maillage triangulaire de l'image. Ceci permettra aussi de diminuer le temps de calcul nécessaire pour cette étape. Enfin, une gestion explicite des occultations devra être ajoutée au modèle.

Références

[1] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Süsstrunk. SLIC superpixels compared to state-of-the-art superpixel methods. *IEEE Transactions on PAMI*, 34(11) :2274–2282, 2012.

(a) (b)

FIGURE 3 – Reconstruction de la scène *Flowerpots* obtenue par l'algorithme SGM (a) et par notre méthode (b)

[2] S. Birchfield and C. Tomasi. Multiway cut for stereo and motion with slanted surfaces. In *ICCV*, volume 1, pages 489–495. IEEE, 1999.

[3] M. Bleyer, C. Rother, and P. Kohli. Surface stereo with soft segmentation. In *CVPR*, pages 1570–1577. IEEE, 2010.

[4] A. Bódis-Szomorú, H. Riemenschneider, and L. Van Gool. Fast, approximate piecewise-planar modeling based on sparse structure-from-motion and superpixels. In *CVPR*, pages 469–476, 2014.

[5] G. Facciolo, C. De Franchis, and E. Meinhardt. MGM : A significantly more global matching for stereovision. In *BMVC*, pages 90–1, 2015.

[6] G. P. Fickel, C. R. Jung, T. Malzbender, R. Samadani, and B. Culbertson. Stereo matching and view interpolation based on image domain triangulation. *IEEE Transactions on Image Processing*, 22(9) :3353–3365, 2013.

[7] Y. Furukawa, B. Curless, S. M. Seitz, and R. Szeliski. Manhattan-world stereo. In *CVPR*, pages 1422–1429. IEEE, 2009.

[8] D. Gallup, J.-M. Frahm, P. Mordohai, Q. Yang, and M. Pollefeys. Real-time plane-sweeping stereo with multiple sweeping directions. In *CVPR*, pages 1–8. IEEE, 2007.

[9] H. Hirschmuller. Stereo processing by semiglobal matching and mutual information. *IEEE Transactions on PAMI*, 30(2) :328–341, 2008.

[10] D. Scharstein, R. Szeliski, and R. Zabih. A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. In *SMBV*, pages 131–140. IEEE, 2001.

[11] G. Silveira, E. Malis, and P. Rives. An efficient direct approach to visual slam. *IEEE Transactions on robotics*, 24(5) :969–979, 2008.

[12] K. Yamaguchi, T. Hazan, D. McAllester, and R. Urtasun. Continuous markov random fields for robust stereo estimation. In *ECCV*, pages 45–58. Springer, 2012.

[13] C. Zhang, Z. Li, Y. Cheng, R. Cai, H. Chao, and Y. Rui. Meshstereo : A global stereo model with mesh alignment regularization for view interpolation. In *ICCV*, pages 2057–2065, 2015.