Reconstruction 3D régularisée à partir de séquences d'images aériennes

Marie-Lise Duplaquet, Guy Le Besnerais, Gilles Foulon

Office National d'Etudes et de Recherches Aérospatiales
DTIM - Unité Traitement d'Images, BP 72, F-92322 Châtillon Cedex, France
Marie_Lise.Duplaquet@onera.fr, Guy.Le_Besnerais@onera.fr, Gilles.Foulon@onera.fr

Résumé – Cet article concerne l'estimation dense des hauteurs à partir d'une séquence d'images aériennes en visée latérale. Dans un contexte calibré, nous utilisons un critère multi-image et séparable par pixel, régularisé par l'ajout d'un terme respectant les discontinuités (norme L1). Dans un deuxième temps, le mouvement est supposé mal connu et une étape préliminaire de recalage entre les images de la séquence est effectuée. Les bons résultats obtenus sur images de synthèse dégagent de nombreuses perspectives d'application aux images réelles.

Abstract – This paper deals with the estimation of an elevation map from side-looking aerial image sequences. In a calibrated context, a pixelwise multi-image criterion is used in association with an edge-preserving regularizing term (L1 norm). When motion is measured with a finite precision, we propose a preliminary motion estimation step. Good results are obtained on a realistic synthetic sequence and should lead to further developments on real sequences.

1 Introduction

L'extraction dense du relief d'une scène à partir d'une séquence d'images calibrées est une composante du thème très actif de la stéréo-mouvement (ou "Structure From Motion" dans la littérature anglo-saxonne). Ce thème recouvre en fait beaucoup de problématiques différentes, suivant le degré de calibration des images utilisées (mouvement et calibration interne connus ou pas) et l'approche retenue pour la mise en correspondance (de primitives, dense, etc.), au point qu'il est impossible de trouver des articles de synthèse approchant seulement l'exhaustivité.

Cette communication reprend le point de vue de [1], proche de celui de la stéréovision à plus de 2 vues, tel que décrit par Kanade et ses collaborateurs («multi-baseline stereo» [2], voir aussi [3]). Essentiellement, on considère que la calibration est parfaite et l'on définit un critère radiométrique global (sur les N images disponibles) de mise en correspondance.

Notre contribution concerne deux points : d'abord nous montrons comment régulariser la mise en correspondance de manière plus efficace, mais aussi plus coûteuse, que dans [1]. Ensuite nous proposons d'étendre l'approche au cas de paramètres de mouvement mal connus, en utilisant un algorithme proposé par Mandelbaum *et al.* [4]. Les résultats sont présentés sur une séquence de synthèse, dans une configuration à visée latérale particulièrement défavorable à la stéréovision.

2 Stéréovision multi-bases

Les travaux présentés dans cette première partie se situent à la suite de la thèse de B. Géraud [5] et concernent le traitement d'une séquence d'images aériennes en vue rasante dans un contexte de stéréovision calibrée à N vues. La configuration choisie (observation avec un angle de dépression de 20 degrés) est très défavorable à la stéréovision, avec un rapport B/H inférieur à 0.1 : même avec des images à résolution métrique, la reconstruction de la hauteur par stéréovision est donc plus que decamétrique. On montre qu'il est théoriquement possible d'améliorer la précision par le traitement d'un grand nombre d'images, ce qu'a démontré empiriquement l'étude [1].

Afin de maîtriser exactement les conditions de prise de vue, nous travaillons sur une séquence de synthèse. L'image présentée en figure 1 montre la vue centrale de la séquence. On dispose pour la synthèse du modèle numérique de terrain, sur lequel est plaquée la texture d'une image aérienne réelle, et du modèle de bâtiments, dont les faces sont texturées aléatoirement. Un modèle capteur est appliqué aux images synthétiques suréchantillonnées, comprenant une fonction de transfert (on se limite à une fonction de transfert détecteur) et un bruit pixel.

La reconstruction est relative à un plan moyen de la scène, supposé connu (mais qui pourrait être estimé par recalage homographique) : on cherche la hauteur des éléments de la scène relativement à ce plan. En simulation, la carte des hauteurs de référence est disponible (voir figure 2) ce qui permet des comparaisons quantitatives avec les cartes reconstruites.

L'estimation dense de la carte des hauteurs utilise un critère global qui mesure la variabilité radiométrique des images d'un même point 3D le long de la séquence :

$$C(x,y,h) = \sum_{k=1}^{n} |I_r(x,y) - I_k(H_{r,k}(x,y,h))|^2 \qquad (1)$$

où (x,y) désigne les coordonnées du pixel considéré de

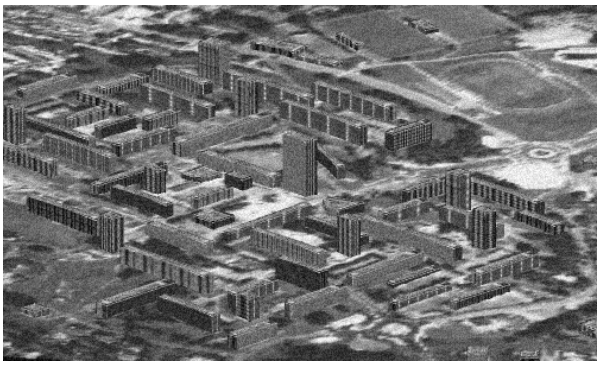

FIG. 1: Image centrale de la séquence

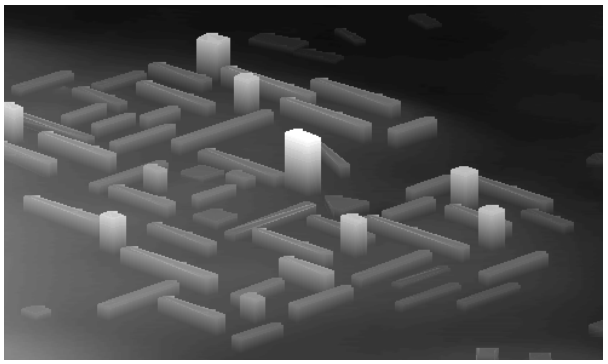

FIG. 2: Carte des hauteurs de référence

l'image de référence, h est l'élévation du point et $H_{r,k}(x,y,h)$ est la transformation qui définit l'image du point 3D (x,y,h) dans la vue k. Il s'agit d'une homographie plane, parfaitement connue dans un contexte calibré.

Le critère précédent a l'inconvénient d'être biaisé, à cause de la présence de l'image de référence dans tous les résidus. D'autres critères portant sur la variance du vecteur des radiométries ont été proposés [3], mais leur comportement en cas d'occlusion partielle d'un point au cours de la séquence peut être moins satisfaisant. La caractéristique essentielle commune à ces critères est la séparation pixel-par-pixel, à la différence des critères de corrélation de fenêtre qui induisent un lissage des hauteurs estimées.

FIG. 3: Carte des hauteurs estimée

La solution simple retenue dans [1] pour minimiser $\mathcal{C}$ consiste à procéder par échantillonnage de l'espace des hauteurs h de la scène. La figure 3 montre la carte obtenue pour une séquence de 61 images avec un bruit pixel de 11 dB. Bien que relativement bruité, le résultat permet de reconnaitre les bâtiments constitutifs de la scène. Par ailleurs il n'est pas lissé, puisque le critère est pixellique.

Une analyse quantitative du résultat, présentée en figure 4, permet de vérifier empiriquement que pour un déplacement total du porteur donné (i.e. à B/H constant) le traitement d'un nombre croissant d'images de la séquence permet d'améliorer la précision relativement à l'utilisation des 2 vues extrêmes. On note aussi que la précision maximale est déjà quasiment obtenue pour 20 images traitées.

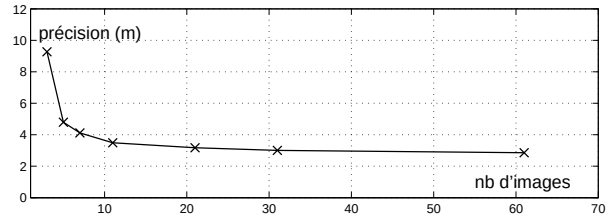

FIG. 4: Précision en fonction du nombre d'images traitées à B/H fixé

3 Régularisation

Dans [1], le résultat était lissé *a posteriori* par un filtre médian. Nous proposons ici une approche globale dont l'enjeu est de respecter les discontinuités de hauteur de la carte estimée.

La méthode par échantillonnage conduit à calculer une valeur de critère en chaque point et pour chaque hypothèse de hauteur. Au lieu de choisir le minimum indépendamment en chaque point, on ajoute un terme de régularisation au critère liant les points voisins. La carte des hauteurs doit ensuite être obtenue par optimisation globale.

Le critère global, régularisé par un terme en valeur absolu, pour conserver les ruptures, est le suivant :

$$\mathcal{C}(\boldsymbol{h}) = \sum_{(x,y)} C(x,y,h(x,y)) + \lambda \sum_{(x',y')\in V(x,y)} |h(x,y) - h(x',y')|$$

où $V(x,y)$ désigne le voisinage aux 4 plus proches voisins du site (x,y). Le paramètre de régularisation est choisi à $\lambda = 1$.

Nous réalisons l'optimisation globale en utilisant un recuit simulé [6]. Bien que très coûteux, l'algorithme du recuit permet de sortir des minima locaux dans lesquels les optimisations simples ont échoué. La décroissance de la température est choisie en $1/\log$. Le résultat de la figure 5 est obtenu pour 1000 itérations, soit environ 3 heures de calcul. La carte obtenue en figure 5 est très satisfaisante : la régularisation est efficace, même dans les zones bruitées du haut de l'image dans lesquelles on distingue des bâtiments invisibles dans la carte non régularisée. Cette intégration spatiale s'obtient sans nuire à la précision de position des ruptures de hauteurs, comme le montrent certains bords

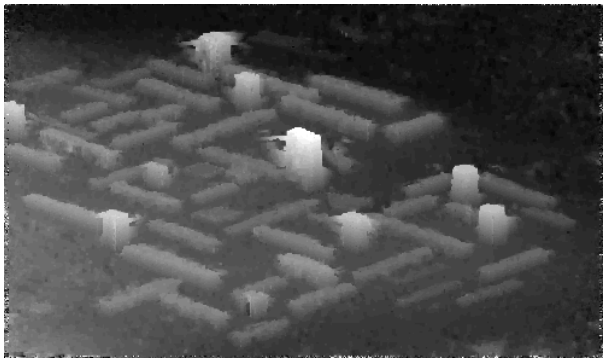

FIG. 5: Carte des hauteurs régularisée

de bâtiments. Les principaux artefacts restants sont dus aux occlusions qui ne sont pas modélisées.

4 Estimation de mouvement

Même en équipant le porteur (robot, avion) de capteurs odométriques de précision, les paramètres de rotation et translation entre les images de la séquence ne seront jamais parfaitement connus. Des tests avec l'algorithmie précédente montrent que dès que l'erreur sur le mouvement entraîne des décalages supérieurs au pixel, la carte des hauteurs estimées se dégrade rapidement et devient inexploitable.

La figure 6 présente un exemple de reconstruction à partir de paramètres de mouvement erronés, obtenus en ajoutant des perturbations gaussiennes indépendantes sur chaque mouvement. L'écart-type est de 1m sur les paramètres de translation et de 1/10 d'arc-seconde pour les paramètres angulaires. Ces perturbations faibles entraînent néanmoins, à cause de la configuration de prise de vue, des erreurs de plusieurs pixels sur le recalage des points et les résultats de reconstruction se dégradent très rapidement.

D'où la nécessité de développer un algorithme permettant l'estimation de la structure et du mouvement. Nous faisons cependant l'hypothèse que l'erreur de mouvement reste faible, c'est-à-dire se traduit par des décalages de quelques pixels au maximum.

FIG. 6: Résultat avec 21 images et un mouvement imparfaitement connu

Nous utilisons une approche d'estimation conjointe de la structure et du mouvement développée par Mandelbaum *et coll.* [4]. Il s'agit d'une approche dense (au sens où tous les pixels sont considérés) qui utilise des corrélations de fenêtres. Les surfaces de corrélation, calculées pour tous les pixels, entre une image de référence et toutes les autres images, sont résumées par une quadrique et servent à la fois à l'estimation du mouvement et du relief. Si la quadrique n'est pas définie (dans une zone homogène par exemple), le pixel est écarté des traitements ultérieurs.

Sans détailler la méthode, notons qu'il s'agit d'une méthode itérative de descente locale alternant des étapes d'estimation des mouvements inter-caméras et des étapes de reconstruction de la carte de hauteurs. La carte de hauteurs obtenue, présentée en figure 7, n'est pas complètement dense puisque les points sont sélectionnés en fonction de la forme de leur surface de corrélation. Par ailleurs elle est nettement lissée, d'une part à cause de l'utilisation des fenêtres de corrélation, d'autre part car il s'agit du résultat d'une seule itération. Cependant, le résultat est satisfaisant et indique que la méthode a dû estimer correctement les paramètres de mouvement.

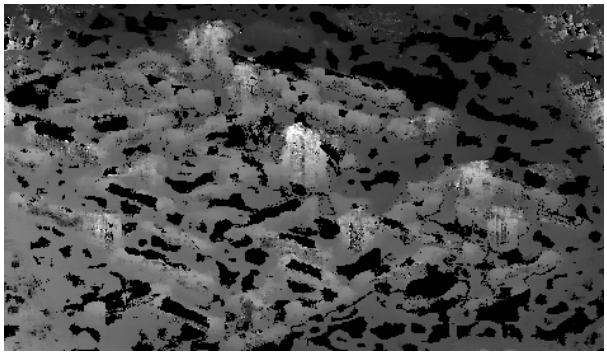

FIG. 7: Résultat de l'algorithme de Mandelbaum *et coll.* (en noir, zones non estimées : 30 %)

De fait, nous sommes intéressés par le mouvement estimé, que nous utilisons pour lancer la méthode de reconstruction des hauteurs présentée dans les sections précédentes. La figure 8 montre la carte obtenue, de qualité comparable à celle obtenue avec mouvement connu. Après régularisation, on vérifie que le relief est correctement estimé, voir figure 9.

FIG. 8: Résultat du critère non régularisé après estimation du mouvement

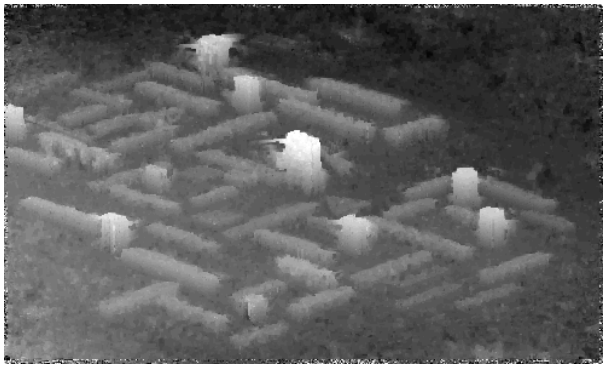

FIG. 9: Carte obtenue après régularisation

Pour quantifier les résultats obtenus, on compare à la carte des hauteurs de référence et on calcule d'une part un pourcentage de points aberrants, c'est-à-dire présentant une erreur supérieure à 10 m, d'autre part un écart quadratique moyen calculé sur 75% des points, considérés comme non aberrants. La table 1 résume les résultats obtenus.

TAB. 1: Résultats de reconstruction 3D. En col. 1 : nombre d'images utilisées, col 2. : type de perturbation sur les données (bruit pixel et/ou perturbation du mouvement), col. 3 : méthode : non régularisée, régularisée et/ou estimation du mouvement, col. 4 : RMS en mètres.

nbi	bruit	méthode	RMS	aberrants
21	non	non rég.	1.28	8.13 %
21	pixel	non rég.	3.24	20.12 %
21	pixel	régul.	1.13	4.31 %
61	pixel	non rég.	2.86	18.17 %
61	pixel	régul.	0.91	4 %
21	pix+mvt	non rég.	6.25	37.32 %
21	pix+mvt	mvt non rég.	3.73	22.96 %
21	pix+mvt	mvt régul	1.49	4.65 %

Notons d'abord que tous ces résultats sont nettement meilleurs que ceux qui pourraient être obtenus avec des critères comparables utilisant seulement deux images. L'intégration temporelle s'avère efficace dès qu'une vingtaine d'images sont traitées. Elle permet de compenser le bruit radiométrique que l'on a choisi élevé (RSB : 11 dB) pour correspondre aux effets atmosphériques sur les prises de vues aériennes rasantes. La régularisation spatiale introduite en section 3 permet un gain remarquable à la fois en précision et pourcentage de données aberrantes. La perturbation du mouvement dégrade considérablement l'estimation de hauteur, mais sa correction par l'algorithme de Mandelbaum *et coll.* est efficace : les résultats après correction du mouvement et régularisation sont comparables à ceux obtenus avec une calibration parfaite.

5 Conclusion

Le travail présenté étend les résultats de [1] à la fois en ce qui concerne la régularisation spatiale de l'estimation et la prise en compte des erreurs de mesure du mouvement de l'observateur. Les bons résultats obtenus en synthèse avec un mouvement mal connu nous permettent d'envisager le traitement de séquence réelles, qui ne sont jamais parfaitement renseignées.

Les perspectives de ce travail sont nombreuses. Le calcul de la carte régularisée est très lourd : d'autres algorithmes d'optimisation discrète sont en cours de test.

L'étape d'estimation du mouvement peut être développée. D'une part il faudra dans certains cas réels compenser des erreurs de recalage plus fortes, par exemple par une approche multi-échelle. D'autre part on dispose de modèles des erreurs de mesures du mouvement et d'une modélisation dynamique du porteur qui peuvent être intégrés dans l'estimation.

Les principaux défauts de l'estimation sont liés aux parties cachées. On peut essayer d'introduire une modélisation des occlusions dans une première phase d'estimation dense, au risque d'alourdir encore cette étape. On peut aussi considérer ce problème dans une phase ultérieure d'interprétation en termes de superstructures de la carte de hauteur.

Références

[1] B. Géraud, G. Le Besnerais et G. Foulon, « Determination of dense depth map from an image sequence: application to aerial imagery », in *European Symposium on Remote Sensing, Image and Signal processing for Remote sensing IV*, septembre 1998.

[2] T. Kanade, M. Okutomi et T. Nakahara, « A multibaseline stereo method », in *proceedings of DARPA image understanding workshop*, janvier 1992, pp. 409–426.

[3] S. Roy, « Stereo without epipolar lines: a maximum flow formulation », *International Journal of Computer Vision*, vol. 34, n° 2/3, pp. 147–161, 1999.

[4] R. Mandelbaum, G. Salgian et H. Sawhney, « Correlation-based estimation of ego-motion and structure from motion and stereo », in *ICCV'99*. IEEE, 1999.

[5] B. Géraud, *Reconstruction tridimensionnelle à partir d'une séquence d'images : application à l'imagerie aérienne*, Thèse de Doctorat, Université Paris V, Paris, France, avril 1999.

[6] S. Geman et D. Geman, « Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images », *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. PAMI-6, n° 6, pp. 721–741, novembre 1984.