Reconstruction 3D de scènes complexes par maillage de cartes de disparité

Lionel Oisel[1], Luce Morin[1], Etienne Mémin[1], Claude Labit[1]

[1]IRISA/INRIA
Campus universitaire de Beaulieu
35042, Rennes, France
{loisel, lmorin, memin, labit}@irisa.fr

Résumé – Nous présentons dans cet article une méthode de reconstruction 3D de scènes complexes à partir de séquences d'images non calibrées. Dans un premier temps un champ de disparité contraint par la géométrie épipolaire est estimé. Ce champ est alors segmenté suivant un modèle de mouvement homographique associé à une triangulation de Delaunay itérative. L'utilisation d'une technique d'auto-étalonnage permet de passer du modèle 2D calculé au modèle VRML de la scène. Ce dernier permet alors de nombreuses manipulations telles que la création de vues extrapolées.

Abstract – We present a new algorithm for fully automatic 3D reconstruction from uncalibrated images in order to deal with modeling complex and rigid scenes. A 2D triangular model of the scene is calculated using a two steps algorithm mixing discrete and continuous approaches. Each triangular patch corresponds to the projection of a 3D plane. First a robust and regularized dense disparity field is estimated under the epipolar geometry constraint. Then, the resulting field is segmented according to an homographic model using iterative Delaunay triangulation. In association with a classical self-calibration algorithm, this 2D planar model is then used to obtain a 3D model of the scene which is compatible with a VRML representation. This representation allows to modify original view-points in order to generate interpolated or extrapolated views.

1 Introduction

Notre travail s'inscrit dans le domaine de l'analyse/synthèse de séquences d'images vidéo numériques. L'objectif est de parvenir à construire un modèle 3D d'une scène complexe ne contenant pas d'objets en mouvement. Les scènes ou objets ainsi modélisés trouvent leurs applications dans de nombreux domaines (commerce électronique, post production, simulateurs pour les entraînements en milieu hostile ...). Les approches généralement rencontrées peuvent être séparées en deux classes : les approches basées modèles et les approches basées images. La génération de modèle [6]offre généralement de bons résultats en terme de qualité visuelle mais repose sur des hypothèses de scènes polyédriques. Les approches basées image [3, 5] permettent quant à elles de s'affranchir des hypothèses sur le contenu de la scène mais offre une qualité visuelle très aléatoire (problèmes des zones de discontinuités, d'occultations, sensibilité à la précision de la mise en correspondance). Nous proposons dans cet article une méthode basée modèle permettant de s'affranchir des hypothèses sur le contenu de la scène.

Nous considérons ici deux images extraites de la séquence originale. La phase de modélisation est décomposée en trois étapes bien distinctes que nous détaillerons par la suite :

- la première étape consiste à estimer un champ de disparité s'appuyant sur des techniques d'estimation de flot optique avec prise en compte de la contrainte épipolaire;
- la carte de disparité ainsi obtenue est alors maillée en utilisation une triangulation de Delaunay associée à un modèle de mouvement homographique;
- une information de profondeur est alors associée à chaque noeud du maillage pour obtenir un modèle 3D VRML (le remplissage des triangles étant effectué par projection homographique de texture).

Nous présentons maintenant ces trois étapes avant de les illustrer par quelques résultats.

2 Estimation du champ de disparité

L'algorithme développé s'inspire des travaux réalisés par E. Mémin et P. Pérez [7]. Afin de faciliter l'estimation de la disparité et d'obtenir une meilleure reconstruction 3D, il s'avère nécessaire de prendre en compte la contrainte épipolaire [4]. Cette contrainte permet d'associer à un point de la première image I_1 une droite de correspondants potentiels dans la deuxième image I_2 (appelée droite épipolaire). Cette contrainte peut s'exprimer sous forme matricielle par la matrice dite fondamentale. La caméra n'étant pas étalonnée, la géométrie épipolaire est directement extraite de la paire d'images par extraction puis appariement de points d'intérêts couplés à une estimation robuste de la matrice fondamentale utilisant la parallaxe virtuelle [2, 8]. La disparité d_s à estimer en un point s peut alors être décomposée sous la forme d'un vecteur normal et tangentiel à la droite épipolaire (voir figure 1). La DFD traduisant

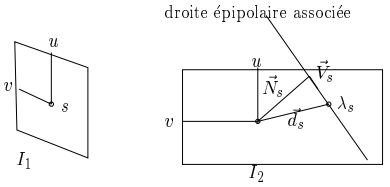

FIG. 1: La contrainte épipolaire

la conservation de la luminance entre deux points correspondants s'écrit alors :

$$DFD(s) = I_1(s) - I_2(s + \vec{N}_s + \lambda_s \vec{V}_s)$$

La composante normale étant connue (géométrie épipolaire), il reste donc à estimer les λ_s en chaque pixel. Pour ce faire, nous nous ramenons à un problème de minimisation d'une fonction d'énergie. Cette fonction se décompose en une somme de deux termes :

$$\begin{aligned} \widehat{\lambda} &= \arg\min_{\lambda \in \mathbb{R}^{|s|}} H(\lambda) \\ &= \arg\min_{\lambda \in \mathbb{R}^{|s|}} (H_1(\lambda) + \alpha[H_2(\lambda)]) \end{aligned} \quad (1)$$

où α est une constante fixée arbitrairement. Le premier terme H_1 est lié à la linéarisation de la DFD. Il représente le terme d'attache aux données :

$$H_1(\lambda) = \sum_{s \in S} \rho \left[\lambda_s \vec{V}_s \nabla \tilde{I}_2(s) + \tilde{I}_2(s) - I_1(s) \right]$$

$$\text{avec } \tilde{I}_2(s) = I_2(s + \vec{N}_s)$$

H_2 représente quant à lui le terme de lissage qui tend à favoriser des déplacements similaires d_s et d_r pour chaque paire de points voisins $<s,r> \in \mathcal{C}$:

$$H_2(\lambda) = \sum_{<s,r> \in \mathcal{C}} \rho(||d_s - d_r||)$$

Ces deux termes utilisent également des estimateurs robustes ρ qui permettent d'autoriser des déviations soit par rapport au modèle (zones d'occultation) soit par rapport au lissage (discontinuités de profondeur). Sous certaines hypothèses, la fonction robuste se ramène à une somme pondérée de termes quadratiques : une forte déviation par rapport au modèle ou au lissage entraîne une faible pondération de l'énergie associée.

La résolution du problème est effectuée par un schéma déterministe de Gauss-Seidel multi-résolution associé à une formulation incrémentale [8]. La matrice fondamentale associée à un niveau de résolution k est calculé à partir de la matrice fondamentale obtenue au niveau de résolution le plus fin par changement de base :

$$\mathbf{F}^k = \mathbf{M}^{kT} \mathbf{F} \mathbf{M}^k$$

où $\mathbf{M} = diag(2,2,1)$ est la matrice associée à ce changement de base. La matrice $\mathbf{F}^k$ permet alors de calculer les vecteurs $\vec{N}_s^k$ et $\vec{V}_s^k$ pour chaque position s. Le champ de vecteurs de disparité obtenu au niveau de résolution k est alors projeté au niveau de résolution $k-1$ afin d'en initialiser le processus de minimisation énergétique.

Devant l'importance des déplacements à estimer, un schéma multirésolution seul ne suffit pas. Afin d'assurer la convergence de l'algorithme, un champ initial est calculé en interpolant les vecteurs disparité résultant de l'extraction et de l'appariement des points d'intérêt. Ce champ est alors projeté vers la résolution la plus faible (image la plus petite) afin de fournir une estimée initiale proche de la solution optimale.

3 Maillage du champ de disparité

A l'issue de l'étape précédente nous disposons donc d'un champ de disparité dense. L'étape de segmentation consiste maintenant à mailler ce champ de telle sorte que chaque triangle corresponde à la projection d'une partie d'un plan de l'espace. En écrivant les équations du plan 3D associé au modèle de projection perspective et au déplacement de la caméra dans l'espace, on peut montrer que la disparité sur la projection de ce plan doit répondre à un modèle de mouvement homographique [1]. En récrivant ce modèle pour deux points $\mathbf{p}_1$ et $\mathbf{p}_2$ correspondant aux projections dans les deux images d'un point 3D appartenant au plan, nous obtenons une équation linéaire en coordonnées homogènes :

$$\mathbf{H}\tilde{\mathbf{p}}_1 = \lambda \tilde{\mathbf{p}}_2$$

où $\mathbf{H}$ est une matrice 3×3 homogène appelée matrice d'homographie associée aux projections du plan dans les deux images et $\tilde{\mathbf{p}}_i$ est un point de la ième image exprimé en coordonnées homogènes. Un maillage initial est tout d'abord construit en prenant arbitrairement 4 points proches des angles de la première image. Le découpage est alors effectué de manière itérative. Pour chaque triangle, l'homographie associée est calculée avec l'ensemble des points en correspondance appartenant à la facette hormis ceux représentant une zone d'occultation (information donnée par l'estimateur robuste associée au terme d'attache aux données). L'estimation de $\mathbf{H}$ est également contrainte par la géométrie épipolaire [9].

Un critère de découpage caractérisant la distance entre le champ de disparité dense et le modèle homographique estimé est alors calculé. Ce critère prend également en compte l'information sur les discontinuités de profondeur donnée par l'estimateur robuste associé au terme de lissage. Le triangle est découpé si ce critère est inférieur à un seuil donné. Cette opération est effectuée de manière itérative jusqu'à ce que le modèle homographique associé à chaque triangle respecte le champ dense associé.

4 Reconstruction 3D

Le modèle 2D correspondant à des facettes planes doit maintenant être rétro-projeté dans l'espace 3D. Pour ce faire, il est nécessaire d'affecter une information de profondeur à chaque sommet du maillage. La mise en correspondance étant connue, il reste à connaître les matrices de projection pour parvenir à retrouver cette information. Dans notre cas, une information 3D précise n'est pas nécessaire (seule la géométrie globale de la scène doit être respectée). Les paramètres de la caméra (appelés paramètres intrinsèques) sont donc fixés arbitrairement. Une fois ces paramètres fixés, il est possible de remonter aux rotation (matrice $\mathbf{R}$) et translation (vecteur $\mathbf{t}$) entre le repère associé à la première prise de vue et celui associé à la deuxième.

Comme le montrent Tsai et al. dans [10], connaissant la matrice de paramètres intrinsèques $\mathbf{A}$, il est possible de retrouver le mouvement de la caméra entre les deux vues à partir de la matrice fondamentale. Ce calcul se fait par l'intermédiaire de la matrice essentielle $\mathbf{E}$ obtenue par l'équation :

$$\mathbf{E} = \mathbf{A}^T\mathbf{F}\mathbf{A} = \hat{\mathbf{t}}\mathbf{R}$$

où $\hat{\mathbf{t}}$ représente la matrice de produit vectoriel associé au vecteur de translation $\mathbf{t}$.

Les paramètres de rotation et de translation sont obtenus en décomposant la matrice essentielle par SVD (singular value decomposition). La rotation n'est ici obtenue qu'à un angle de Π près alors que la translation n'est déterminée qu'à un facteur d'échelle algébrique près. Cette ambiguïté peut cependant être levée à partir d'une paire de points en correspondance dans les deux images. Il suffit pour cela de reconstruire le point 3D correspondant et de modifier l'angle de rotation et le signe de la translation jusqu'à ce que le point reconstruit se situe devant les deux caméras. La translation n'est alors connue qu'à un facteur d'échelle positif près.

Le modèle 3D ainsi reconstruit est mis au format VRML afin de pouvoir être transmis et manipulé en temps réel. L'information de texture utilisée lors de l'étape de remplissage des triangles est extraite de la première image.

5 Résultats

Nous avons testé l'ensemble de la chaîne algorithmique précédemment décrite sur des images d'une scène non polyédrique (séquence Yosemite). Nous présentons les deux images utilisées en figure 2. La disparité obtenue prend des valeurs comprises entre 7 et 26 pixels. De tels déplacements s'avèrent trop importants pour permettre à des algorithmes de calcul de flot optique non contraints et non initialisés de converger vers une solution satisfaisante. La carte de disparité obtenue (module du vecteur disparité en chaque point, cf figure 3) met bien en évidence la montagne au premier rang (zone claire). La figure 4 présente quant à elle, les poids associés aux deux estimateurs robustes utilisés. La première carte montre les discontinuités de profondeur: quand le poids associé à une paire de pixel est faible, l'apport énergétique du terme de lissage est borné ce qui traduit la présence de discontinuités (zones noires). Il en va de même pour la carte d'observation où les zones d'occultation sont mises en évidence par les zones plus sombres (notamment le pourtour non commun aux deux images). Les figures 5, 6 et 7 sont constituées quant à elles de captures d'écran réalisées à partir d'un outil de visualisation VRML. Les images résultent d'un mouvement de translation le long de l'axe z ainsi que de mouvements plus complexes. Si la géométrie globale de la scène parait respectée, une translation trop importante fait apparaître un flou.

L'algorithme a également été testé sur des séquences polyédriques. Un modèle VRML a été construit à partir de deux images extraites de la séquence «armel» en utilisant l'algorithme que nous venons de présenter. (cf figure 8). Parallèlement, un deuxième modèle a été construit à

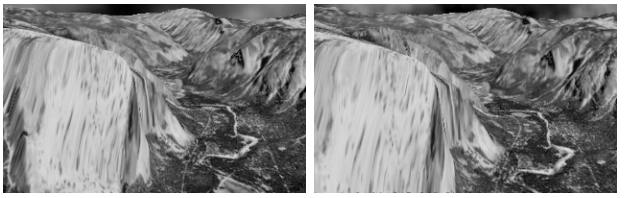

FIG. 2: Images originales 3 et 12 et carte de disparité associée

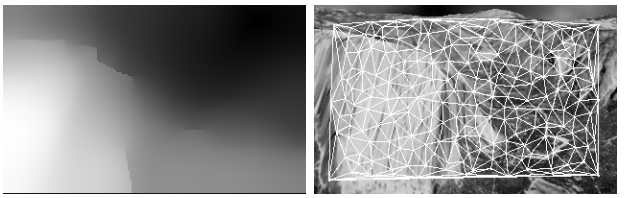

FIG. 3: Images de disparité et triangulation finale obtenue

FIG. 4: Cartes de discontinuité et d'observation : plus le pixel est noir plus le poids associé est faible.

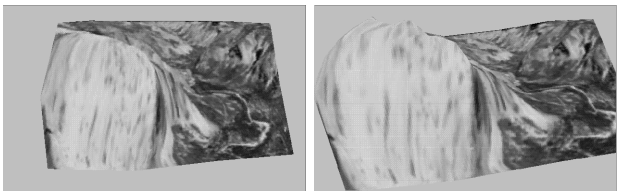

FIG. 5: Simulation de translation en z

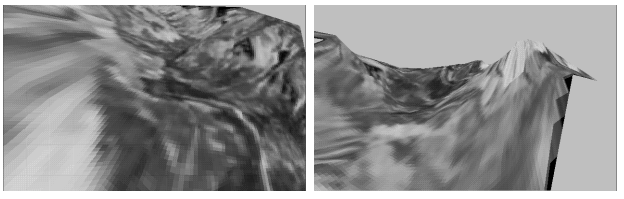

FIG. 6: Simulation de translation en z et vue de derrière de la vallée

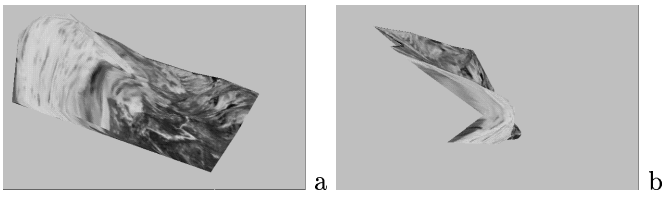

FIG. 7: Simulation de mouvements complexes

partir de la triangulation de points extraits et appariés. Ces points ont été fournit par un logiciel de calcul de la géométrie épipolaire «image-matching» développé par Zhang [11]. Ces deux modèles ont été utilisés afin de générer un certain nombre de vues plus ou moins proches des images originales (cf figure 9). Les différents résultats obtenus mettent en évidence l'apport de notre méthode en terme de qualité de rendu visuel par rapport à une technique reposant sur une simple extraction de points d'intérêt.

6 Conclusion

Nous proposons dans cet article la description d'une chaîne algorithmique complète et automatique de modélisation 3D de scènes à partir de séquence d'images non étalonnées. L'intérêt principal de notre approche réside dans le caractère automatique de la méthode associé à l'absence d'hypothèses sur les scènes considérées. Nous avons ainsi pu présenter des résultats de reconstruction sur une scène non polyédrique. Afin d'améliorer le modèle VRML obtenu, il semble nécessaire d'effectuer une phase d'auto-étalonnage permettant une meilleure estimation de la géométrie de la prise de vues associée à une étape de remise à jour du modèle permettant d'intégrer l'information provenant d'images supplémentaires.

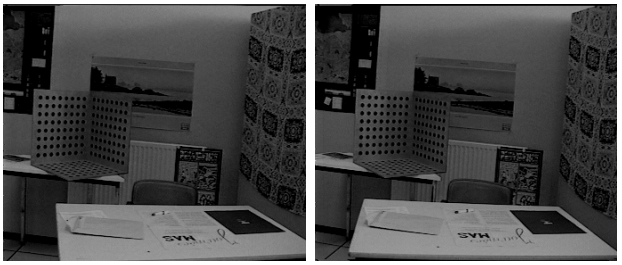

FIG. 8: Images originales d'une scène d'intérieur.

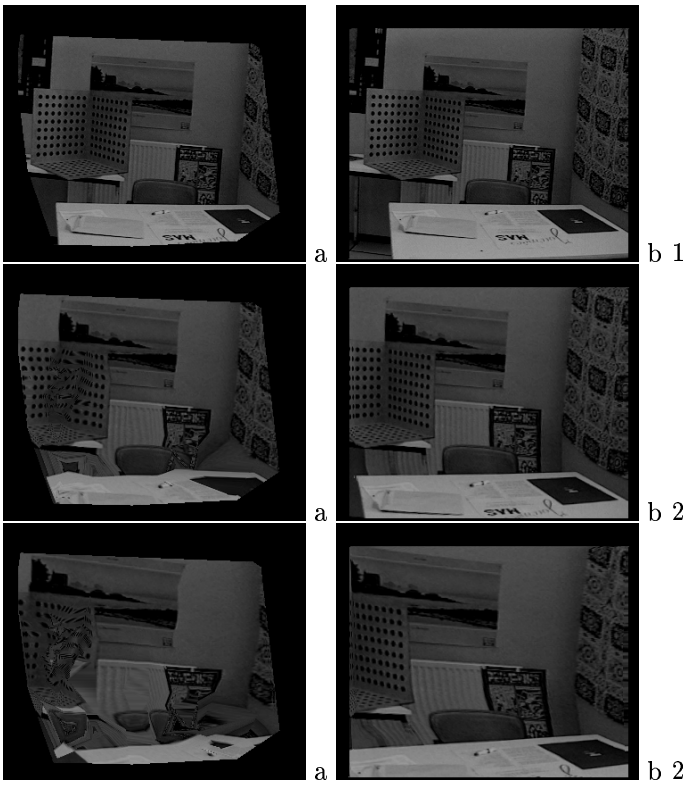

FIG. 9: Images gauches resynthétisées (1) et vues extrapolées (2) : modèle calculé à partir de points d'intérêt extraits et appariés automatiquement (a), modèle obtenu par notre méthode

Références

[1] Adiv (G.). – Determining three-dimensional motion and structure from optical flow generated by several moving objects. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 7, n° 4, juillet 1985, pp. 384–401.

[2] Couapel (B.) et Bainian (K.). – Stereo vision with the use of a virtual plane in the space. *Chinese Journal of Electronics*, vol. 4, n° 2, 1995, pp. 32–39.

[3] Faugeras (O.) et Laveau (S.). – Representing three-dimensional data as a collection of images and fundamental matrices for image synthesis. *In : Proceedings of International Conf. on Pattern Recognition*, pp. 689–691. – Jerusalem, Israel, 1994.

[4] Hartley (R.I.) et Sturm (P.). – Triangulation. *Computer Vision and Image Understanding*, vol. 68, n° 2, novembre 1997, pp. 146–157.

[5] Kanade (T.), Narayanan (P.J.) et Rander (P.W.). – Virtualized reality: Concept and early results. *In : Workshop on Representation of Visual Scenes.* – Cambridge, USA, 1995.

[6] Koch (R.), Pollefeys (M.) et Gool (L. Van). – Automatic 3d model acquisition from uncalibrated image sequences. *In : Computer Graphics International*, pp. 597–604. – Hannover, 1998.

[7] Mémin (E.) et Perez (P.). – Dense estimation and object-based segmentation of the optical-flow with robust techniques. *IEEE Image Processing*, vol. 7, n° 5, mai 1998, pp. 703–719.

[8] Oisel (L.). – *Reconstruction 3D de scènes complexes à partir de séquences vidéo non calibrées : estimation et maillage d'un champ de disparité.* – Thèse de doctorat, IRISA, Université de RENNES I, novembre 1998.

[9] Robert (L.) et Faugeras (O.). – Relative 3-D positioning and 3-D convex hull computation from a weakly calibrated stereo pair. *Image and Vision Computing*, vol. 13, n° 3, 1995, pp. 189–197.

[10] Tsai (R.) et Huang (T.). – Uniqueness and estimation of three dimensional motion parameters of rigid objects with curved surface. *Transaction on Pattern Analysis ans Machine intelligence*, vol. 6, 1984, pp. 13–26.

[11] Zhang (Z.). – Estimating motion and structure from correspondences of line segments between two perspective images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 17, n° 12, 1995, pp. 1129–1139.